Measuring Text Readability by Lexical Relations Retrieved from WordNet

Shu-yen Lin, Cheng-chao Su, Yu-da Lai, Li-chin Yang, Shu-kai Hsieh
English Department
National Taiwan Normal University
sylin@ntnu.edu.tw

Abstract

Current readability formulae have often been criticized for being unstable or not valid. They are mostly computed in regression analysis based on intuitively-chosen variables and graded readings. This study explores the relation between text readability and the conceptual categories proposed in Prototype Theory. These categories form a hierarchy: Basic level words like *guitar* represent the objects humans interact with most readily. They are acquired by children earlier than their superordinate words (or hypernyms) like *stringed instrument* and their subordinate words (or hyponyms) like *acoustic guitar*. Therefore, the readability of a text is presumably associated with the ratio of basic level words it contains. WordNet, a network of meaningfully related words, provides the best online open source database for studying such lexical relations. Our preliminary studies show that a basic level word can be identified by its frequency to form compounds (e.g. *chair* → *armchair*) and the length difference from its hyponyms in average. We compared selected high school English textbook readings in terms of their basic level word ratios and their values calculated in several readability formulae. Basic level word ratios turned out to be the only one positively correlated with the text levels.

Keywords: Readability, Ontology, Prototype Theory, WordNet, Basic Level Word

1. Introduction

Reading process is the core of language education. Teachers now have access to a vast amount of texts extractable from the Internet inter alia, but the materials thus found are rarely classified according to comprehension difficulty. It is not uncommon to see foreign language teachers using texts not compatible with the students' reading abilities.

Traditional methods of measuring text readability typically rely on the counting of sentences, words, syllables, or characters. However, these formulae have been criticized for being unstable and incapable of providing deeper information about the text. Recently, the focus of readability formula formation has shifted to the search for meaningful predictors and stronger association between the variables and the comprehension difficulty.

We start our research by assuming in line with Rosch et al.'s Prototype Theory [1] that words form conceptual hierarchies in that words at different hierarchical levels pose different processing difficulties. This processing difficulty is presumably correlated with the reading difficulty of the text containing the words. Putting the logic into templates, the measurement

of text readability can be done by calculating the average hierarchical levels at which the words of a text fall.

Our study comprises two stages. In the preliminary experiments, we utilized WordNet [2], an online lexical database of English, to identify basic level words. In the subsequent experiment, we compared selected readings in terms of their basic level word ratios and their values calculated in several readability formulae. Basic level word ratios turned out to be the only one positively correlated with the text levels.

The remainder of this paper is organized as follows: Section 2 reviews the common indices the traditional readability formulae are based on and the criticism they have received. In Section 3, we first review an approach that centers on ontology structure, and then propose our own ontology-based approach. Section 4 is about methodology – how to identify basic level words, and how to assess the validity of our method against other readability formulae. Section 5 reports the results of the assessment and discusses the strength and weaknesses of our approach. In this section, we also suggest what can be done in further research.

2. Literature Review

In this section we first summarize the indices of the traditional readability formulae and then give an account of the criticism these formulae face.

2.1 Indices of Readability – Vocabulary, Syntactic, and Semantic Complexity

The earliest work on readability measurement goes back to Thorndike [3] where word frequency in corpus is considered an important index. This is based on the assumption that the more frequent a word is used, the easier it should be. Followers of this logic have compiled word lists that include either often-used or seldom-used words whose presence or absence is assumed to be able to determine vocabulary complexity, thus text complexity. Vocabulary complexity is otherwise measured in terms of word length, e.g., the Flesch formula [4] and FOG formula [5]. This is based on another assumption that the longer a word is, the more difficult it is to comprehend [6].

Many readability formulae presume the correlation between comprehension difficulty and syntactic complexity. For Dale and Chall [7], Flesch formula [4], and FOG index [5], syntactic complexity boils down to the average length of sentences in a text. Heilman, Collins-Thompson, Callan, and Eskenazi [8] also take morphological features as a readability index for morphosyntactically rich languages. Das & Roychoudhury's readability index [9] for Bangla has two variables: average sentence length and number of syllables per word.

Flesch [4] and Cohen [10] take semantic factors into account by counting the abstract words of a text. Kintsch [11] focuses on propositional density and inferences. Wiener, M., Rubano, M., and Shilkret, R. [12] propose a scale based on ten categories of semantic

relations including, e.g., temporal ordering and causality. They show that the utterances of fourth-, sixth-, and eighth-grade children can be differentiated on their semantic density scale.

Since 1920, more than fifty readability formulae have been proposed in the hope of providing tools to measure readability more accurately and efficaciously [13]. Nonetheless, it is not surprising to see criticism over these formulae given that reading is a complex process.

2.2 Criticism of the Traditional Readability Formulae

One type of criticism questions the link between readability and word lists. Bailin and Grafstein [14] argue that the validity of such a link is based on the prerequisite that words in a language remain relatively stable. However, different socio-cultural groups have different core vocabularies and rapid cultural change makes many words out of fashion. The authors also question the validity of measuring vocabulary complexity by word length, showing that many mono- or bi-syllabic words are actually more unfamiliar than longer polysyllabic terms.

These authors also point out the flaw of a simple equation between syntactic complexity and sentence length by giving the sample sentences as follows:

(1) I couldn't answer your e-mail. There was a power outage.

(2) I couldn't answer your e-mail because there was a power outage.

(2) is longer than (1), thus computed as more difficult, but the subordinator "because" which explicitly links the author's inability to e-mail to the power outage actually aids the comprehension. The longer passage is accordingly easier than the shorter one.

Hua and Wang [15] point out that researchers typically select, as the criterion passages, standard graded texts whose readability has been agreed upon. They then try to sort out the factors that may affect the readability of these texts. Regression analyses are used to determine the independent variables and the parameters of the variables. However, the researchers have no proof of the cause-effect relation between the selected independent variables and the dependent variable, i.e., readability.

Challenge to the formula formation is also directed at the selection of criterion passages. Schriver [16] argue that readability formulae are inherently unreliable because they depend on criterion passages too short to reflect cohesiveness, too varied to support between-formula comparisons, and too text-oriented to account for the effects of lists, enumerated sequences and tables on text comprehension.

The problems of the traditional readability formulae beg for re-examination of the correlation between the indices and the readability they are supposed to reflect.

3. Ontology-based Approach to Readability Measurement

3.1 An ontology-based method of retrieving information

Yan, X., Li, X., and Song, D. [17] propose a domain-ontology method to rank documents on

the generality (or specificity) scale. A document is more specific if it has broader/deeper Document Scope (DS) and/or tighter Document Cohesion (DC). DS refers to a collection of terms that are matched with the query in a specific domain. If the concepts thus matched are associated with one another more closely, then DC is tighter. The authors in their subsequent study [18] apply DS and DC to compute text readability in domain specific documents and are able to perform better prediction than the traditional readability formulae.

In what follows we describe the approach we take in this study, which is similar in spirit to Yan et al.'s [18] method.

3.2 An Ontology-based Approach to the Study of Lexical Relations

In this small-scaled study, we focus on lexical complexity (or simplicity) of the words in a text and adopt Rosch et al.'s Prototype Theory [1].

3.2.1 Prototype Theory

According to Prototype Theory, our conceptual categorization exhibits a three-leveled hierarchy: basic levels, superordinate levels, and subordinate levels. Imagine an everyday conversation setting where a person says "Who owns this piano?"; the naming of an object with 'piano' will not strike us as noteworthy until the alternative "Who owns this string instrument?" is brought to our attention. Both terms are truth-conditionally adequate, but only the former is normally used. The word 'piano' conveys a basic level category, while 'string instrument' is a superordinate category. Suppose the piano in our example is of the large, expensive type, i.e., a grand piano, we expect a subordinate category word to be used in e.g. "Who owns this grand piano?" only when the differentiation between different types of pianos is necessary.

Basic level is the privileged level in the hierarchy of categorical conceptualization. Developmentally, they are acquired earlier by children than their superordinate and subordinate words. Conceptually, basic level category represents the concepts humans interact with most readily. A picture of an apple is easy to draw, while drawing a fruit would be difficult, and drawing a crab apple requires expertise knowledge. Informatively, basic level category contains a bundle of co-occurring features – an apple has reddish or greenish skin, white pulp, and a round shape, while it is hard to pinpoint the features of 'fruit', and for a layman, hardly any significant features can be added to 'crab apple'.

Applying the hierarchical structure of conceptual categorization to lexical relations, we assume that a basic level word is easier for the reader than its superordinate and subordinate words, and one text should be easier than another if it contains more basic level words.

3.2.2 WordNet – An Ontology-Based Lexical Database of English

WordNet [2] is a large online lexical database of English. The words are interlinked by means of conceptual-semantic and lexical relations. It can be used as a lexical ontology in computational linguistics. Its underlying design principle has much in common with the hierarchical structure proposed in Prototype Theory illustrated in 3.2.1. In the vertical dimension, the hypernym/hyponym relationships among the nouns can be interpreted as hierarchical relations between conceptual categories. The direct hypernym of 'apple' is 'edible fruit'. One of the direct hyponyms of 'apple' is 'crab apple'. Note, however, hypernyms and hyponyms are relativized notions in WordNet. The word 'crab apple', for instance, is also a hypernym in relation to 'Siberian crab apple'. An ontological tree may well exceed three levels. No tags in WordNet tell us which nouns fall into the basic level category defined in Prototype Theory. In the next section we try to retrieve these nouns.

4. Methodology

4.1 Experiment 1

We examined twenty basic level words identified by Rosch et al. [1], checking the word length and lexical complexity of these basic level words and their direct hypernyms as well as direct hyponyms in WordNet [2]. A basic level word is assumed to have these features: (1) It is relatively short (containing less letters than their hypernyms/hyponyms in average); (2) Its direct hyponyms have more synsets[1] than its direct hypernyms; (3) It is morphologically simple. Notice that some entries in WordNet [2] contain more than one word. We assume that an item composed of two or more words is NOT a basic level word. A lexical entry composed of two or more words is defined as a COMPOUND in this study. The first word of a compound may or may not be a noun, and there may or may not be spaces or hyphens between the component words of a compound.

Table 1: Twenty basic level words in comparison with their direct hypernyms and hyponyms on (average) word length, number of synsets, and morphological complexity*

Item	Basic Level		Direct Hypernym			Direct Hyponym		
	W. Length	M. Complexity	W. Length	Synsets	M. Complexity	W. Length	Synsets	M. Complexity
guitar	6	A	18	1	B	10	6	A, B
piano	5	A	18	3	B	10	3	A, B
drum	4	A	20	1	B	7.4	8	A, B
apple	5	A	7.5	2	A, B	10.67	3	A, B
peach	5	A	9	1	B	9	0	N/A

[1] A synset is a set of synonyms. The direct hypernym of 'piano', for instance, is grouped into three synsets: (1) keyboard instrument, (2) stringed instrument, and (3) percussion instrument, percussive instrument.

grape	5	A	11	1	B	11.8	3	A, B
hammer	6	A	8	1	B	9.7	9	A, B
saw	2	A	8	1	B	8.7	7	A, B
screwdriver	11	A	8	1	B	19.8	3	B
pants	5	A	7	1	A	8.9	18	A, B
socks	4	A	7	1	A	7.2	5	A, B
shirt	5	A	7	1	A	7.56	9	A, B
table	5	A	5	1	A	13.8	6	A, B
lamp	4	A	20	1	B	9.88	17	A, B
chair	5	A	4	1	A	11.2	15	A, B
car	3	A	12	1	A, B	7	31	B
bus	3	A	15	1	A, B	8	3	B
truck	5	A	12	1	A, B	8	11	B
dog	3	A	10	2	A, B	7	18	B
cat	3	A	6	1	A, B	9	2	B

*A refers to “single word” and B refers to “compound”.

The results confirm our assumption. First, the average word length (number of letters) of both the hypernyms and the hyponyms is much longer than that of the basic level words. Second, the hyponyms have a lot more synsets than the hypernyms. Third, in contrast to the basic level words which are morphologically simple, their direct hypernyms and hyponyms are more complex. Many of the hypernyms are compounds. The hyponyms are even more complex. Every basic level word (except ‘peach’) has at least one compounded hyponym.

4.2 Experiment 2

In this experiment, we examined the distribution of the compounds formed by the basic level words and their hypernyms and hyponyms. We also randomly came up with five more words that seem to fall into the basic level category defined by Rosch et al. [1]. These basic level words (e.g. ‘guitar’) are boldfaced in each item set in Table 2 below. Above each basic level word is its (or one of its) direct hypernym(s) (e.g. ‘stringed instrument’), under the basic level word is the first-occurring direct hyponym (e.g. ‘acoustic guitar’). When the basic level word has more than one level of hyponym, the first word at the second hyponymous level was also examined (e.g. ‘movable barrier’, ‘door’, ‘car door’, ‘hatchback’). For words that have more than one sense, we focused only on the sense defined in Rosch et al. [1]. For example, the noun ‘table’ has six senses in WordNet; we only focused on the sense ‘a piece of furniture’.

For each target item, we clicked on its FULL HYPONYM in WordNet 3.0 [2] to find the compounds formed by the target item. The next step was to count the compounds formed by the target words. For example, among the twelve hyponyms of ‘guitar’, five are compounds formed by ‘guitar’ – ‘acoustic guitar’, ‘bass guitar’, ‘electric guitar’, ‘Hawaiian guitar’, and ‘steel guitar’. In contrast, only one hyponym of ‘stringed instrument’ is a compound containing ‘stringed instrument’. As for ‘acoustic guitar’, it has no hyponyms. We assume

that basic level words are more apt to form compounds than their hypernyms as well as hyponyms, so their compound ratios are calculated: Number of compounds is divided by number of hyponyms. We also keep record of the level where a compound occurs.

Table 2: Compound ratios and distribution of compounds in hyponymous levels

Hypernym Basic Level Word Hyponym	Cpd # / Hyponym #	Cpd Ratio (%)	Number of Compounds at Hyponymous Levels					
			1st Level	*2nd Level*	*3rd Level*	*4th Level*	*5th Level*	*6th Level*
stringed instrument	1 / 86	1	1	0	0	0		
guitar	5 / 12	42	5					
acoustic guitar	0 / 0	0						
keyboard	0 / 35	0	0	0	0			
piano	8 / 16	50	4	4				
grand piano	3 / 8	38	3					
baby grand piano	0 / 0	0						
percussion	0 / 68	0	0	0	0			
drum	5 / 14	36	5					
bass drum	0 / 0	0						
edible fruit	0 / ...	0	0	0	0	0		
apple	5 / 29	17	5	0	0			
crab apple	2 / 8	25	2					
Siberian crab	0 / 0	0						
N/A	N/A	N/A						
peach	0 / 0	0						
N/A	N/A	N/A						
edible fruit	0 / ...	0	0	0	0	0		
grape	6 / 17	35	3	2	1			
muscadine	0 / 0	0						
hand tool	0 / ...	0	0	0	0	0		
hammer	7 / 16	44	7	0				
ball-peen hammer	0 / 0	0						
hand tool	0 / ...	0	0	0	0	0	0	
saw	25 / 30	83	13	12	0			
bill	0 / 0	0						
hand tool	0 / ...	0	0	0	0	0	0	
screwdriver	4 / 4	100	4					
flat tip screwdriver	0 / 0	0						
garment	4 / 448	0	3	1	0	0	0	
pants	9 / 49	18	8	1				
bellbottom trousers	0 / 0	0						
hosiery	0 / 29	0	0	0				

socks	5 / 13	38	5					
anklet	0 / 0	0						
garment	4 / 448	0	3	1	0	0	0	
shirt	8 / 17	47	8	0				
camise	0 / 0	0						
furniture	4 / ...	0	4	0	0	0	0	
table	39 / 79	49	32	7	0	0		
alter	0 / 0	0						
source of	0 / 108	0	0	0	0	0	0	
lamp	27 / 68	40	14	12	1	0		
Aladdin's lamp	0 / 0	0						
seat	6 /102	6	2	3	1	0		
chair	31 / 48	65	17	14	0			
armchair	0 / 10	0	0	0				
captain's chair	0 / 0	0						
motor vehicle	0 / 153	0	0	0	0	0		
car	21 / 76	28	19	2				
amphibian	0 / 2	0	0					
public transport	0 / 38	0	0	0	0			
bus	3 / 5	60	3					
minibus	0 / 0	0						
motor vehicle	0 / 153	0	0	0	0	0		
truck	15 / 48	31	10	5	0			
dump truck	0 / 0	0						
canine	0 / ...	0	0	0	0	0	0	0
dog	51 / 279	18	13	20	16	2	0	
puppy	0 / 0	0						
feline	0 / ...	0	0	0	0			
cat	35 / 87	40	4	31				
domestic cat	0 / 33	0	0					
kitty	0 / 0	0						
publication	1 / 211	0	0	1	0	0	0	
book	39 / 145	27	21	14	4	0	0	
authority	0 / 7	0	0					
power of	0 / 0	0						
language unit	0 / ...	0	0	0	0	0	0	0
word	35 / 220	16	28	7	0	0	0	
anagram	0 / 1	0	0					
antigram	0 / 0	0						
material	16 / ...	0	14	2	0	0		
paper	59 / 210	28	40	18	1			

card	14 / 57	25	6	8				
playing card	0 / 48	0						
movable barrier	0 / 46	0	0	0	0			
door	18 / 23	78	13	5				
car door	0 / 1	0	0					
hatchback	0 / 0	0						
leaf	2 / 23	9	2	0	0			
page	5 / 20	25	5	0				
full page	0 / 0	0						

Note: The symbol "#" stands for "number". Cpd refers to "compound". The three dots indicate that the number of hyponyms is too many to count manually. The number is estimated to exceed one thousand.

The most significant finding is that basic level words have the highest compound ratios. In comparison with their hypernyms and hyponyms, they are much more frequently used to form compound words. Although some hyponyms like 'grand piano' and 'crab apple' also have high compound ratios, they should not be taken as basic level items because such compounds often contain the basic level words (e.g. 'Southern crab apple'), indicating that the ability to form compounds is actually inherited from the basic level words.

Our data pose a challenge to Prototype Theory in that a subordinate word of a basic level word may act as a basic level word itself. The word 'card', a hyponym of 'paper', is of this type. With its high compound ratio of 25%, 'card' may also be deemed to be a basic level word. This fact raises another question as to whether a superordinate word may also act as a basic level word itself.

Many of the basic level words in our list have three or more levels of hyponym. It seems that what is cognitively basic may not be low in the ontological tree. A closer look at the distribution of the compounds across the hyponymous levels reveals another interesting pattern. Basic level words have the ability to permeate through two to three levels of hyponyms in forming compounds. By contrast, words at the superordinate levels do not have such ability, and their compounds mostly occur at the direct hyponymous level.

4.3 Experiment 3

The goal of this experiment is to show that whether a word belongs to the basic level affects its readability. This in turn affects the readability of a text and should be considered a criterion in measuring text readability. An easy text presumably contains more basic level words than a difficult one. Put in fractional terms, the proportion of basic level words in a text is supposed to be higher than that of a more difficult text.

To achieve this goal, we need independent readability samples to be compared with our prediction. As readability is subjective judgment that may vary from one person to another, such independent samples are extremely difficult, if ever possible, to obtain. In this study, we

resorted to a pragmatic practice by selecting the readings of English textbooks for senior high school students in Taiwan. Three textbooks from Sanmin Publishing Co., each used in the first semester of a different school year, were selected. We tried to choose the same type of text, so that text type will not act as a noise. Furthermore, since we do not have facility to run large-scale experiment yet, we limited the scope to two-hundred-word text at each level. Accordingly, the first two hundred words of the first reading subjectively judged as narrative were extracted from the textbooks (Appendix 1). All the nouns occurring in these texts, except proper names and pronouns, were searched for in WordNet [2]. Considering the fact that for a word with more than one sense, the distribution of hyponyms differs from one sense to another, we searched for the hyponyms of the word in the particular sense occurring in the selected readings. We know that this practice, if used in a large-scale study, is applicable only if sense tagging is available, and we hope that it will be available in the near future.

Based on the results of the two preliminary experiments, we assume that basic level words have at least the following two characteristics: (1) They have great ability to form compounded hyponyms; (2) Their word length is shorter than the average word length of their direct hyponyms. These characteristics can be further simplified as the **Filter Condition** to pick out basic level words:

(1) Compound ratio of full hyponym $\geqq$ 25%;

(2) Average word length of direct hyponym minus target word length $\geqq$ 4.

Note in passing that the second criterion differs fundamentally from the commonly used criterion of word length. Ours compares the target word with its full hyponyms. Word length is measured in relative terms: What is counted is the word length difference, not the word length itself. Based on the two assumed characteristics of our filter condition, the information for each noun we need includes: (1) Length of the target word, i.e. how many letters the word contains; (2) Compound ratio of the target word, i.e. how many hyponyms of the word are compounds formed by the word. Note that here the hyponyms refer to the full hyponyms, so all the words in every hyponymous synset were counted; (3) Average word length of the direct hyponyms. The next section reports the computed information via WordNet [2].

5. Results and Discussion

The three selected readings contain sixty nouns in total, of which twenty-one conform to the proposed Filter Condition of basic level words. They are given in Table 3 below. A comprehensive list of all the sixty nouns are given in Appendix 2 at the end of this paper. Note in passing that the level numbers refer to the presumed difficulty levels of the selected readings. Level 1 is presumably the easiest; Level 3, the hardest. These numbers should not be taken as ratio measurement. Level 3, for example, is not assumed to be three times harder than Level 1. We intend these numbers to stand for ordinal relations.

Table 3: Basic Level Words from the 200-word Texts at Three Levels

Target Word	Level	Compound Ratio (%)	Length of Target Word	Average Length of Direct Hyponyms
food	1	53	4	8
apple	1	56.6	5	10
vinegar	1	60	7	11
potato	1	62.5	6	11
cold	1	66.6	4	8
test	1	72.7	4	9
orange	1	88.8	6	11
soap	1	93	4	9
language	2	35.12	8	12
job	2	37.5	3	8
heart	2	40	5	15
technology	2	47.22	10	19
factor	2	63.64	6	12
culture	2	85.19	7	19
physics	3	32.84	7	12.6
question	3	35.71	7	15
barometer	3	40	9	13.25
system	3	60.95	6	12.93
time	3	62.22	4	10
office	3	72.22	6	11.5
call	3	93.33	4	11

In order to measure the text difficulty, basic level word ratios of the selected texts were computed. Table 4 shows the statistics. Diagrammatically, it is clear in Figure 1 that the basic level word ratios are decreasing as the difficulty levels of the selected readings increase. The text from Level-1 has the highest basic level word ratio; the text from Level-3 has the lowest basic level word ratio. This finding conforms to the levels of these textbooks, and proves the usefulness of the basic level word concept in the measurement of readability.

Table 4: Basic level word ratio at different levels

	Number of nouns	Number of Basic Level Words	Ratio of Basic Level Words
Level-1	17	8	47.1
Level-2	15	6	40.0
Level-3	28	7	25.0

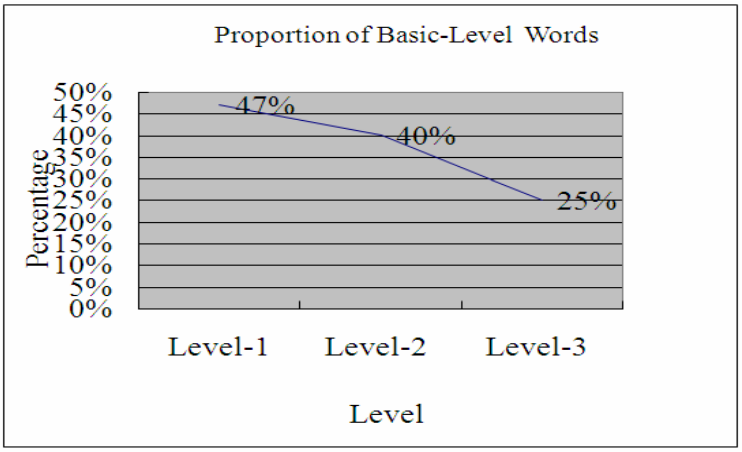


Figure 1: Basic Level Word Ratio of Selected Texts

Table 5 shows the readability scores of the selected readings measured by several readability formulae. Figure 2 displays the overall tendency computed by these formulae: Level-1 is the easiest, while Level-2 and Level-3 are at about the same difficulty level. The readability formulae seem not to be able to decipher the difference between the texts of Level-2 and Level-3 while our basic level word ratio can easily show their different difficulty levels.

Table 5: Readability of the 200-word Texts Computed by Several Readability Formulae

	Dale-Chall	Flesch Grade Level	FOG	Powers	SMOG	FORCAST	Spache
Level-1	4.6	2.1	7.8	4	6.4	7.7	2.4
Level-2	7.4	8.3	18.9	6.2	10.2	11.8	3.9
Level-3	6.3	9.5	16.4	5.9	10.5	9.1	4.8

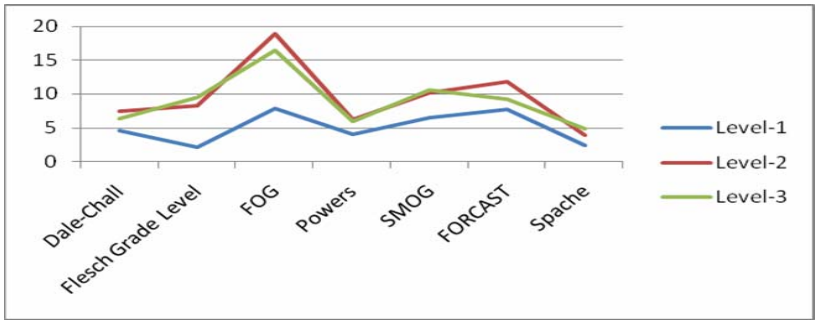


Figure 2: Readability of the 200-word Texts Computed by Several Formulae

This paper is just the first step to measure readability by lexical relations retrieved from WordNet [2]. Twenty-five percent of the twenty basic level words defined by Rosch et al. [1]

are NOT identified by our Filter Condition (e.g. ‘truck’, ‘shirt’, socks’). Among the identified basic level words in the three selected texts, some look rather dubious to us (e.g. ‘barometer’, ‘technology’). The filter condition proposed in this study certainly leaves room to be fine-tuned and improved in at least two respects. First, the two criteria of compound ratios and word length difference have been used as sufficient conditions. We will postulate the possibility of weighting these criteria in our subsequent research. Second, in addition to the lexical relations proposed in this study, there are presumably other lexical relations between basic level words and their hypernyms/hyponyms that are retrievable via WordNet [2]. Doubts can also be raised as to whether all basic level words are equally readable or easy. Can it be that some basic level words are in fact more difficult than others and some hypernyms/ hyponyms of certain basic level words are actually easier than certain basic level words?

We thank our reviewers for raising the following questions, and will put them in the agenda of our subsequent study: (1) The examined words in this study are all nouns. Can we find relationships between verbs, adjectives, and even adverbs like the hypernym/hyponym relationships with the basic level “nouns”? The tentative answer is yes and no. Take the example of the verb ‘run’. It has hypernyms in WordNet (‘speed’, ‘travel rapidly’, etc.). It also has subordinate lexical relation called ‘troponym’, which is similar to hyponym of nouns. Admittedly, English verbs do not constitute compounds so often as English nouns, but other lexical relations may exist between the verbs, and the relations are likely to be retrievable. (2) Although the small scale size of our experiments makes the validity of the results challengeable, the exciting findings of this study have provided the outlook of a large-scale project in the future. (3) Are basic level words frequent words in general? Can we use frequency to substitute for ‘basichood’ if the two criteria have approximately the same indexing power? We like to extend this question and ask whether the ontological relations between the lexical units in WordNet are correlated with word frequency. We hope we will be able to answer this question in a study of larger scale.

Laying out the groundwork for further research, we aim to tackle the following issues too. All traditional readability formulae implicitly suppose an isomorphic relation between form and meaning as if each word has the same meaning no mater where it occurs. We acknowledge that one of the biggest challenges and the most badly needed techniques of measuring readability is to disambiguate the various senses of a word in text since the same word may have highly divergent readability in different senses. Another tacit assumption made by the traditional readability formulae is that the units of all lexical items are single words. This assumption overlooks many compounds and fixed expressions and affects the validity of these formulae.

Although our research has provided the study of readability a brand new perspective and has offered exciting prospects, our challenges are still many and the road is still long.

References

[1] Rosch, Eleanor, Mervis, Carolyn, Gray, Wayne, Johnson, David, & Boyes-Braem, Penny, "Basic objects in natural categories," Cognitive Psychology 8: 382-439, 1976.

[2] WordNet, version 3.0. Princeton, N.J.: Princeton University. Retrieved from World Wide Web: http://wordnet.princeton.edu/perl/webwn?s=word-you-want, 2006.

[3] Thorndike, E.L., The Teacher's Word Book. New York: Teacher's College, Columbia University, 1921.

[4] Flesch, R., "A new readability yardstick", Journal of Applied Psychology 32: 221-233, 1948.

[5] McCallum, D. R., & Peterson, J. L., "Computer-based readability indices," *Proceedings of the ACM '82 Conference*, 1982.

[6] Chall, J., & Dale, E., *Readability revisited: The new Dale-Chall readability formula*. Cambridge, Massachusetts: Brookline Books, 1995.

[7] Dale, E., Chall, J., "Formula for predicting readability," *Educational Research Bulletin* 27 (1–20), 37–54, 1948.

[8] Heilman, Collins-Thompson, Callan & Eskenazi, "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts," *Proceedings of the HLT/NAACL Annual Conference*, 2007.

[9] Das & Roychoudhury, "Readability Modelling and Comparison of One and Two Parametric Fit: A Case Study in Bangla," *Journal of Quantitative Linguistics, 13*, 17-34, 2006.

[10] Cohen, J.H., "The effects of content are material on cloze test performance," *Journal of Reading*, 19/3: 247-50, 1975.

[11] Kintsch, W., *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum, 1974.

[12] Wiener, M., Rubano, M., and Shilkret, R., "A measure of semantic complexity among predications," *Journal of Psycholinguistic Research*, Vol. 19, No. 2: 103-123, 1990.

[13] Crossley, S.A, Dufty, D.F., McCarthy, P.M., & McNamara, D.S., "Toward a new readability: A mixed model approach," *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society, 2007.

[14] Bailin, A. & Grafstein, Ann, "The Linguistic Assumptions Underlying Readability Formulae: A Critique," *Language and Communication* 21(3): 285-301, 2001.

[15] Hua, N. & Wang, G., "Lun chuantong keduxing gongshi de bu-kexuexing," [On the non-scientific aspects of traditional readability formulae]. *KaoShiZhouKan* 18: 119-120, 2007.

[16] Schriver, K. A., "Readability formulas in the new millennium: What's the use?" *ACM Journal of Computer Documentation, 24.3*: 138-140, 2000.

[17] Yan, X., Li, X., and Song. D., "Document generality: its computation for ranking," In: G. Dobbie and J. Bailey *Seventeenth Australiasian Database Conference (ADC2006)*, Hobart, Australia, 16-19 January, 2006a.

[18] Yan, X., Li, X., and Song. D., "Concept-based Document Readability in Domain Specific Information Retrieval," CIKM 2006: 540-549, 2006b.

Appendix 1: Three Pieces of 200-word-text from a Senior High School Textbook

Level 1: Book 1 Lesson 2

Scientist say that your tongue can recognize only four tastes. It can tell if something is sour (like vinegar) or bitter (like soap). But that's all. To tell different foods apart, we also have to use our noses.

Can you remember a time when you had a bad cold? Your food tasted very plain then. It seemed to have little taste at all. That wasn't because your tongue wasn't working. It was because your nose was stopped up. You couldn't smell the food, and that made it seem tasteless. You can prove this to yourself. Try eating something while you pinch your nose shut. It won't seem to have much taste.

Here's another test. It shows how important the nose is in tasting. First you blindfold a person. Then you put a piece of potato in his mouth. You tell him to chew it. At the same time, you hold a piece of apple under his nose. Then ask what food is in his mouth. Most people will say, " An apple." The smell of the apple fools them. The test works best when two foods feel the same in the mouth. It won't work well with apple and orange slices.

Level 2: Book 3 Lesson 2

When people from different cultures live and work together much more than before, change takes place. The languages of the world's dominant cultures are replacing the languages of the smaller cultures. You're learning English right now. Could this be the beginning of the end for the Chinese language? Of course not. *Mandarin* remains the healthy, growing language at the heart of Chinese culture. Mandarin steadily continues to spread among Chinese people worldwide. Elsewhere, *Swahili* grows in Africa. Spanish continues to thrive in *South America*. *Hindi* rules India. And of course almost everyone these days wants to learn English. However, many less common regional languages haven't been so lucky, because most young people have stopped learning them.

When less common languages disappear, two factors are to blame: trade and technology. Most international trade takes place in major world languages such as English or Mandarin. Cultures that isolate themselves from international business and major world languages have difficulty prospering.

Most children respect their own culture and traditions. But when it comes to getting a job, knowing a major world language if often essential. It may mean the difference between success and failure. For many, using a less common reginal language simply isn't

Level 3: Book 5 Lesson 2

Some time ago, I received a call from a colleague who asked if I would be the referee on the grading of an examination question. He was about to give a student a zero for his answer to a physics question, while the student claimed he should receive a perfect score and would if the system were not set up against the student. The instructor and the student agreed to submit this to an impartial judge, and I was selected.

I went to my colleagues' office and read the examination question: "Show how it is possible to determine the height of a tall building with the aid of a barometer." The student had answered: "Take the barometer to the top of the building, attach a long rope to it and lower the barometer to the street. Then bring it up and measure the length of the rope. The length of the rope is the height of the building."

I pointed out that the student really had a strong case for full credit, since he had answered the question completely and correctly. On the other hand, if full credit were given, it could well contribute to a high grade for the

Appendix 2: Nouns Extracted from the Three Pieces of 200-word-text.

Item	Target Item				Direct Hyponyms	
	Level	Cpd # / Hyponym #	Cpd Ratio (%)	Length	Avg. Length	Number
scientist	1	37/174	21	9	13	20
tongue	1	0/4	0	6	0	0
taste	1	4/34	11.7	5	6	9
vinegar	1	3/5	60	7	11	3
soap	1	14/15	93	4	9	8
food	1	1234/2310	53	4	8	15
nose	1	4/22	18	4	6	8
time	1	1/0	0	4	0	0
cold	1	2/3	66.6	4	8	1
test	1	8/11	72.7	4	9	5
person	1	3152/13235	23.8	6	8	401
potato	1	10/16	62.5	6	11	5
mouth	1	3/10	30	5	4	6
apple	1	17/30	56.6	5	10	3
smell	1	2/23	8.6	5	6	4
orange	1	8/9	88.8	6	11	3
slice	1	2/10	20	5	6	2
culture	2	23/27	85.19	7	19	7
language	2	425/1210	35.12	8	12	16
world	2	2/9	22.22	5	11	3
end	2	23/54	42.59	3	6	14
heart	2	2/5	40	5	15	2
factor	2	14/22	63.64	6	12	6

trade	2	16/66	24.24	5	10	3
technology	2	17/36	47.22	10	19	7
business	2	54/163	33.13	8	8	12
child	2	34/55	61.82	5	7	21
tradition	2	0/7	0	9	6	4
job	2	3/8	37.5	3	8	15
difference	2	0/11	0	10	10	9
success	2	24/58	41.38	7	7	5
failure	2	7/50	14	7	7	8
time	3	28/45	62.22	4	10	16
call	3	14/15	93.33	4	11	8
colleague	3	0/0	0	9	N/A	N/A
referee	3	0/0	0	7	N/A	N/A
grading	3	0/0	0	7	N/A	N/A
examination	3	20/32	62.5	11	9	24
question	3	10/28	35.71	7	15	3
student	3	16/48	33.33	7	9.25	20
zero	3	0/0	0	4	N/A	N/A
answer	3	0/2	0	6	8	2
physics	3	22/67	32.84	7	12.6	18
score	3	1/5	20	5	8.5	4
system	3	103/169	60.95	6	12.93	28
instructor	3	30/55	54.55	10	10.86	21
judge	3	7/33	21.21	5	7.33	3
office	3	13/18	72.22	6	11.5	8
height	3	0/7	0	6	7.5	2
building	3	212/485	43.71	8	9.76	54
aid	3	0/1	0	3	8	1
barometer	3	2/5	40	9	13.25	4
top	3	0/9	0	3	5.8	5
rope	3	15/37	40.54	4	7.21	19
street	3	22/32	68.75	6	8.95	21
length	3	1/19	5.26	6	8.8	5
case	3	0/2	0	4	8	1
credit	3	1/9	11.11	6	7	3
hand	3	0/1	0	4	4	1
grade	3	1/5	20	5	8.5	4

Note 1: Level ranges from 1 to 3, which respectively represents the English textbooks of Book I for the first-year senior high school students, Book III for the second-year, and Book V for the third-year senior high school students in Taiwan.

Note 2: Cpd ratio refers to the ratio of compounds formed by the target item to the total number of the target item's full hyponyms.

Note 3: Direct hyponyms refer to the lexical items at the level immediate below the target item.

A Semantic Composition Method for Deriving Sense Representations of Determinative-Measure Compounds in E-HowNet

Chia-hung Tai, Shu-Ling Huang, Keh-Jiann Chen
Institute of Information Science, Academia Sinica
glaxy, josieh, kchen @iis.sinica.edu.tw

摘要

本篇論文利用定量複合詞為例，示範如何利用廣義知網的語意合成機制來推導複合詞的語意及其表達式。首先我們定義了所有但有限數量的定詞跟量詞的廣義知網表達式，接著我們利用語意合成的規則針對任何新的定量詞去產生候選的語意表達式。然後我們在從調整語料集合去設計語意解歧規則，利用啟發式語意解歧規則跟參考上下文的詞來解決定量詞的廣義知網表達式的歧異，實驗顯示在語意推導跟解歧之後有 88%的正確率。

Abstract

In this paper, we take Determinative-Measure Compounds as an example to demonstrate how the E-HowNet semantic composition mechanism works in deriving the sense representations for all determinative-measure (DM) compounds which is an open set. We define the sense of a closed set of each individual determinative and measure word in E-HowNet representation exhaustively. We then make semantic composition rules to produce candidate sense representations for any newly coined DM. Then we review development set to design sense disambiguation rules. We use these heuristic disambiguation rules to determine the correct context-dependent sense of a DM and its E-HowNet representation. The experiment shows that the current model reaches 88% accuracy in DM identification and sense derivation.

關鍵詞：語意合成，定量複合詞，語意表達，廣義知網，知網

Keywords: Semantic Composition, Determinative-Measure Compounds, Sense Representations, Extended How Net, How Net

1. Introduction

Building knowledge base is a time consuming work. The CKIP Chinese Lexical Knowledge Base has about 80 thousand lexical entries and their senses are defined in terms of the E-HowNet format. E-HowNet is a lexical knowledge and common sense knowledge representation system. It was extended from HowNet [1] to encode concepts. Based on the

framework of E-HowNet, we intend to establish an automatic semantic composition mechanism to derive sense of compounds and phrases from lexical senses [2][3]. Determinative-Measure compounds (abbreviated as DM) are most common compounds in Chinese. Because a determinative and a measure normally coin a compound with unlimited versatility, the CKIP group does not define the E-HowNet representations for all DM compounds. Although the demonstrative, numerals, and measures may be listed exhaustively, their combination is inexhaustible. However their constructions are regular [4]. Therefore, an automatic identification schema in regular expression [4] and a semantic composition method under the framework of E-HowNet for DM compounds were developed.

In this paper, we take DMs as an example to demonstrate how the E-HowNet semantic composition mechanism works in deriving the sense representations for all DM compounds. The remainder of this paper is organized as follows. The section 2 presents the background knowledge of DM compounds and sense representation in E-HowNet. We'll describe our method in the section 3 and discuss the experiment result in the section 4 before we make conclusion in the section 5.

2. Background

There are numerous studies on determinatives as well as measures, especially on the types of measures.[1] Tai [5] asserts that in the literature on general grammar as well as Chinese grammar, classifiers and measures words are often treated together under one single framework of analysis. Chao [6] treats classifiers as one kind of measures. In his definition, a measure is a bound morpheme which forms a DM compound with the determinatives enumerated below. He also divides determinatives word into four subclasses:

i.Demonstrative determinatives, e.g. 這" this", that"那"…
ii.Specifying determinatives, e.g. 每"every", 各" each"…
iii.Numeral determinatives, e.g. 二"two", 百分之三"three percentage", 四百五十" four hundred and fifty"…
iv.Quantitative determinatives, e.g. 一" one", 滿" full", 許多" many"…

Measures are divided into nine classes by Chao [6]. Classifiers are defined as 'individual measures', which is one of the nine kinds of measures.

i.classifiers, e.g. 本"a (book)",

[1] Chao [6] and Li and Thompson [7] detect measures and classifiers. He [8] traces the diachronic names of measures and mentions related literature on measures. The dictionary of measures pressed by Mandarin Daily News Association and CKIP [9] lists all the possible measures in Mandarin Chinese.

ii.classifier associated with V-O constructions, e.g. 手 “hand”,
iii.group measures, e.g. 對”pair”,
iv.partitive measures, e.g. 些”some”,
v.container measures, e.g. 盒“box”,
vi.temporary measures, e.g. 身”body”,
vii.Standard measures, e.g. 公尺”meter”,
viii.quasi-measure, e.g. 國”country”,
ix.Measures with verb, e.g. 次”number of times”.

As we mentioned in the section of introduction, Chao considers that determinatives are listable and measures are largely listable, so D and M can be defined by enumeration, and that DM compounds have unlimited versatility. However, Li and Thompson [7] blend classifiers with measures. They conclude not only does a measure word generally not take a classifier, but any measure word can be a classifier. In Tai’s opinion [5], in order to better understand the nature of categorization in a classifier system, it is not only desirable but also necessary to differentiate classifiers from measure words. These studies on the distinction between classifiers and measures are not very clear-cut. In this paper, we adopt the CKIP DM rule patterns and Part-of-Speeches for morpho-syntactic analysis, and therefore inherit the definition of determinative-measure compounds (DMs) in [10]. Mo et al. define a DM as the composition of one or more determinatives together with an optional measure. It is used to determine the reference or the quantity of the noun phrase that co-occurs with it. We use the definition of Mo et al. to apply to NLP and somewhat different from traditional linguistics definitions.

2.1 Regular Expression Approach for Identifying DMs

Due to the infinite of the number of possible DMs, Mo et al. [10] and Li et al. [4] propose to identify DMs by regular expression before parsing as part of their morphological module in NLP. For example, when the DM compound is the composition of one determinative, e.g. for numerals in (1), roughly rules (2a), (2b) or (2c) will be first applied, and then rules (2d), (2e) or (2f) will be applied to compose complex numeral structures, and finally rules (2g) will generate the pos Neu of numeral structures. From the processes of regular expression, the numerals 534 and 319 in (1) is identified and tagged as Neu.[2]

(1) 鼓勵***534***人完成***319***鄉之旅

guli wubaisanshisi ren wancheng sanbaiyishijiu xiang zhi lu

encourage 534 persons to accomplish the travel around 319 villages

[2] The symbol “Neu” stands for Numeral Determinatives. Generation rules for numerals are partially listed in (2).

(2) a. NO1 = {○,一,二,兩,三,四,五,六,七,八,九,十,廿,卅,百,千,萬,億,兆,零,幾};

b. NO2 = {壹,貳,參,肆,伍,陸,柒,捌,玖,拾,佰,仟,萬,億,兆,零,幾};

c. NO3 = {１,２,３,４,５,６,７,８,９,０,百,千,萬,億,兆};

d. IN1 -> { NO1*, NO3*} ;

e. IN2 -> NO2* ;

f. IN3 -> {IN1,IN2} {多,餘,來,幾} ({萬,億,兆}) ;

g. Neu -> {IN1,IN2,IN3 } ;

Regular expression approach is also applied to deal with ordinal numbers, decimals, fractional numbers and DM compounds for times, locations etc.. The detailed regular expressions can be found in [4]. Rule patterns in regular expression only provide a way to represent and to identify morphological structures of DM compounds, but do not derive the senses of complex DM compounds.

2.2 Lexical Sense Representation in E-HowNet

Core senses of natural language are compositions of relations and entities. Lexical senses are processing units for sense composition. Conventional linguistic theories classify words into content words and function words. Content words denote entities and function words without too much content sense mainly serve grammatical function which links relations between entities/events. In E-HowNet, the senses of function words are represented by semantic roles/relations [11]. For example, 'because' is a function word. Its E-HowNet definition is shown in (1).

(1) because|因為 def: reason={};

which means reason(x)={y} where x is the dependent head and y is the dependent daughter of '因為'.

In following sentence (2), we'll show how the lexical concepts are combined into the sense representation of the sentence.

(2) Because of raining, clothes are all wet. 因為下雨，衣服都濕了

In the above sentence, '濕 wet', '衣服 clothes' and '下雨 rain' are content words while '都 all', '了 Le' and '因為 because' are function words. The difference of their representation is

that function words start with a relation but content words have under-specified relations. If a content word plays a dependent daughter of a head concept, the relation between the head concept and this content word will be established after parsing process. Suppose that the following dependent structure and semantic relations are derived after parsing the sentence (2).

(3) S(reason:VP(Head:Cb:因為|dummy:VA:下雨)|theme:NP(Head:Na:衣服)｜quantity:Da:都｜Head:Vh:濕|particle:Ta:了)。

After feature-unification process, the following semantic composition result (4) is derived. The sense representations of dependent daughters became the feature attributes of the sentential head 'wet|濕'.

(4) def:{wet|濕:
theme={clothing|衣物},
aspect={Vachieve|達成},
manner={complete|整},
reason={rain|下雨}}

In (3), function word '因為 because' links the relation of 'reason' between head concept '濕 wet' and '下雨 rain'. The result of composition is expressed as reason(wet|濕)={rain|下雨}, since for simplicity the dependent head of a relation is normally omitted. Therefore reason(wet|濕)={rain|下雨} is expressed as reason={rain|下雨}; theme(wet|濕)={clothing|衣物} is expressed as theme={clothing|衣物} and so on.

2.3 The sense representation for determinatives and measures in E-HowNet

The sense of a DM compound is determined by its morphemes and the set of component morphemes are determinatives and measures which are exhaustively listable. Therefore in order to apply semantic composition mechanism to derive the senses of DM compounds, we need to establish the sense representations for all morphemes of determinatives and measures first. Determinatives and measure words are both modifiers of nouns/verbs and their semantic relation with head nouns/verbs are well established. We thus defined them by a semantic relation and its value like (5) and (6) bellowed.

(5) The definition of determinatives in E-HowNet

this 這	def: quantifier={definite|定指}
first 首	def: ordinal={1}
one 一	def: quantity={1}

We find some measure words contain content sense which need to be expressed, but for some measure words, such as classifiers, their content senses are not important and could be neglect. So we divided measure words into two types: with or without content sense, their

sense representations are exemplified below:

(6) The definition of measure words in E-HowNet

a) Measure words with content sense

bowl 碗 def: container={bowl|碗}

meter 米 def: length={meter|公尺}

month 月 def: time={month|月}

b) Measure words without content sense

本 copy def:{null}

間 room def:{null}

樣 kind def:{null}

3. Semantic Composition for DM Compounds

To derive sense representations for all DM compounds, we study how to combine the E-HowNet representations of determinative and measure words into a DM compound representation, and make rules for automatic composition accordingly. Basically, a DM compound is a composition of some optional determinatives and an optional measure. It is used as a modifier to describe the quantity, frequency, container, length…etc. of an entity. The major semantic roles played by determinatives and measures are listed in the Table 1.

The basic feature unification processes [12]:

If a morpheme *B* is a dependency daughter of morpheme *A*, i.e. *B* is a modifier or an argument of *A*, then unify the semantic representation of *A* and *B* by the following steps.

Step 1: Identify semantic relation between *A* and *B* to derive relation(*A*)={*B*}. Note: the possible semantic relations are shown in Table 1.

Step 2: Unify the semantic representation of *A* and *B* by insert relation(*A*)={*B*} as a sub-feature of *A*.

It seems that a feature unification process can derive the sense representation of a DM compound, as exemplified in (7) and (8), once its morpheme sense representations and semantic head are known.

(7) one 一 def:quantity={1} + bowl 碗 def: container={bowl|碗} →

one bowl 一碗 def: container={bowl|碗:quantity={1}}

(8) this 這 def: quantifier={definite|定指} + 本 copy def:{null} →

this copy 這本 def: quantifier={definite|定指}

Table 1. Major semantic roles played by determinants and measures

Semantic Role	D/M

quantifier	e.g. 這、那、此、該、本、貴、敝、其、某、諸
ordinal	e.g. 第、首
qualification	e.g. 上、下、前、後、頭、末、次、首、其他、其餘、別、旁、他、另、另外、各
quantity	e.g. 一、二、萬、雙、每、任何、一、全、滿、整、一切、若干、有的、一些、部份、有些、許多、很多、好多、好幾、好些、少許、多、許許多多、幾許、多數、少數、大多數、泰半、不少、個把、半數、諸多
Formal={.Ques.}	e.g. 何、啥、什麼
Quantity={over, approximate, exact}	e.g. 餘、許、足、之多、出頭、好幾、開外、整、正
position	e.g. 桌子、院子、地、屋子、池、腔、家子
container	e.g. 盒(子)、匣(子)、箱(子)、櫃子、櫥(子)、籃(子)、簍(子)、爐子、包(兒)、袋(兒)、池子、瓶(子)、桶(子)、聽、罐(子)、盆(子)、鍋(子)、籠(子)、盤(子)、碗、杯(子)、勺(子)、匙(湯匙)、筒(子)、擔(子)、籮筐、杓(子)、茶匙、壺、盅、筐、瓢、鍬、缸
length	e.g. 公厘、公分、公寸、公尺、公丈、公引、公里、市尺、營造　尺、台尺、吋(inch)、呎(feet)、碼(yard)、哩(mile)、　(海)浬、庹、噚、尺、里、釐、寸、丈、米、厘、厘米、海　哩、英尺、英里、英呎、英寸、米突、米尺、微米、毫米、　英吋、英哩、光年
size	e.g. 公畝、公頃、市畝、營造畝、坪、畝、分、甲、頃、平方公里、平方公尺、平方公分、平方尺、平方英哩、英畝
weight	e.g. 公克、公斤、公噸、市斤、台兩、台斤(日斤)、盎司(斯)、磅、公擔、公衡、公兩、克拉、斤、兩、錢、噸、克、英磅、英兩、公錢、毫克、毫分、仟克、公毫
volume	e.g. 公撮、公升(市升)、營造升、台升(日升)、盎司、品脫(pint)、加侖(gallon)、蒲式耳(bushel)、公斗、公石、公秉、公合、公勺、斗、毫升、夸、夸特、夸爾、立方米、立方厘米、立方公分、立方公寸、立方公尺、立分公里、立方英尺、石、斛、西西
time	e.g. 微秒、釐秒、秒、秒鐘、分、分鐘、刻、刻鐘、點、點鐘、時、小時、更、夜、旬、紀(輪, 12 年) 、世紀、天(日)、星期(禮拜、週、周) 、月、月份、季、年(載、歲) 、週年、周歲、年份、晚、宿、世、輩、輩子、代、學期、學年、年代

address	e.g. 國、省、州、縣、鄉、村、鎮、鄰、里、郡、區、站、巷、弄、段、號、樓、衖、市、洲、地、街
place	e.g 部、司、課、院、科、系、級、股、室、廳
duration	e.g 陣(子)、會、會兒、下子

However there are some complications need to be resolved. First of all we have to clarify the dependent relation between the determinative and the measure of a DM in order to make a right feature unification process. In principle, a dependent head will take semantic representation of its dependent daughters as its features. Usually determinatives are modifiers of measures, such as 這碗, 一碗, 這一碗. For instance, the example (9) has the dependent relations of
NP(quantifier:DM(quantifier:Neu:一|container:Nfa:碗)|Head:Nab:麵)

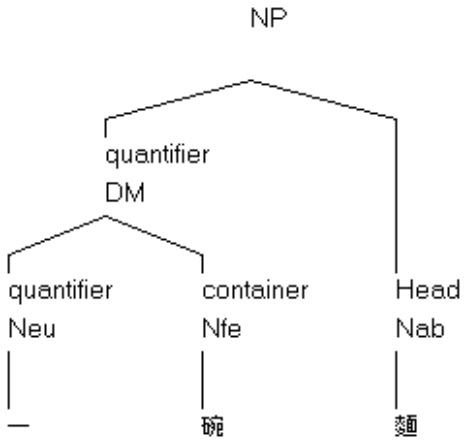


Figure 1. The dependent relations of 一碗麵"a bowl of noddle".

After feature unification process, the semantic representation of "一 def: quantity={1}" becomes the feature of its dependent head "碗 def: container={bowl|碗} and derives the feature representation of "one bowl 一碗 def: container={bowl| 碗 :quantity={1}}". Similarly, "one bowl 一碗" is the dependent daughter of "noodle|麵 def:{noodle|麵}". After unification process, we derive the result of (9).

(9)one bowl of noodle|一碗麵 def:{noodle|麵:container={bowl|碗:quantity={1}}}

The above feature unification process written in term of rule is expressed as (10).

(10) Determinative + Measure (D+M) → def: semantic-role(M) = {Sense-representation(M): Representation(D)}

The rule (10) says that the sense representation of a DM compound with a determinative D

and a measure M is a unification of the feature representation of D as a feature of the sense representation of M as exemplified in (9).
However a DM compound with a null sense measure word, such as 'this copy|這本' , ' a copy|一本', or without measure word, such as 'this three|這三', will be exceptions, since the measure word cannot be the semantic head of DM compound. The dependent head of determinatives become the head noun of the NP containing the DM and the sense representation of a DM is a coordinate conjunction of the feature representations of its morphemes of determinatives only.

For instance, in (8), 'copy' has weak content sense; we thus regard it as a null-sense measure word and only retain the feature representation of the determinative as the definition of "this copy|這本". The unification rule for DM with null-sense measure is expressed as (11).

(11) Determinative + {Null-sense Measure} (D+M) → def: Representation(D);

If a DM has more than one determinative, we can consider the consecutive determinatives as one D and the feature representation of D is a coordinate conjunction of the features of all its determinatives. For instance, "this one|這一" and "this one|這一本" both are expressed as "quantifier={definite|定指}; quantity={1}".
Omissions of numeral determinative are occurred very often while the numeral quantity is "1". For instance, "這本" in fact means "this one|這一本". Therefore the definition of (8) should be modified as:
這本 def: quantifier={definite|定指}; quantity={1};
The following derivation rules cover the cases of omissions of numeral determinative.

(12) If both numeral and quantitative determinatives do not occur in a DM, then the feature quantity={1} is the default value of the DM.

Another major complication is that senses of morphemes are ambiguous. The feature unification process may produce many sense representations for a DM compound. Therefore sense disambiguation is needed and the detail discussions will be in the section 3.1.

Members of every type of determinatives and measures are exhaustively listable except numeral determinatives. Also the formats of numerals are various. For example, "5020" is equal to "五零二零" and "五千零二十" and "五千二十". So we have to unify the numeral representation into a standard form. All numerals are composition of basic numeral as shown in the regular expressions (2). However their senses are not possible to define one by one. We take a simple approach. For all numeral, their E-HowNet sense representations are expressed

as themselves. For example, 5020 is expresses as quantity={5020} and will not further define what is the sense of 5020. Furthermore all non-Arabic forms will be convert into Arabic expression, e.g. “五千零二十” is defined as quantity={5020}.

The other problem is that the morphological structures of some DMs are not regular patterns. Take “兩個半 two and half” as an example. “半 half” is not a measure word. So we collect those word like “多many, 半half, 幾many, 上up, 大big, 來more” for modify the quantity definition. So we first remove the word “半” and define the “兩個” as quantity={2}. Because the word “半” means quantity={0.5}, we define the E-HowNet definition for “兩個半” as quantity={2.5}. For other modifiers such as “多 many, 幾 many, 餘 more, 來 more”, we use a function over() to represent the sense of “more”, such as “十多個 more than 10” is represented as quantity={over(10)}

The appendix A shows the determinatives and measures used and their E-HowNet definition in our method. Now we have the basic principles for compositing semantics of DM under the framework of E-HowNet.

Below steps is how we process DMs and derive their E-HowNet definitions from an input sentence.

I. Input: a Chinese sentence.
II. Apply regular expression rules for DM to identify all possible DM candidates in the input sentence.
III. Segment DM into a sequence of determinatives and measure words.
IV. Normalize numerals into Arabic form if necessary
V. Apply feature unification rules (10-12) to derive candidates of E-HowNet representations for every DM.
VI. Disambiguate candidates for each DM if necessary.
VII. Output: DM Compounds in E-HowNet representation.

For an input Chinese sentence, we use the regular expression rules created by Li et al. [2006] to identify all possible DMs in the input sentence. Then, for every DM compound, we segment it into a sequence of determinatives and measures. If any numeral exists in the DM, every numeral is converted into decimal number in Arabic form. For every DM, we follow the feature unification principles to composite semantics of DM in E-HowNet representations and produce possible ambiguous candidates. The final step of sense disambiguation is described in the following section.

3.1 Sense Disambiguation

Multiple senses will be derived for a DM compound due to ambiguous senses of its morpheme components. For instance, the measure word ”頭 head” has either the sense of

{頭|head}, such as “滿頭白髮 full head of white hairs” or the null sense in ”一頭牛 a cow”. Some DMs are inherent sense ambiguous and some are pseudo ambiguous. For instances, the above example “一頭” is inherent ambiguous, since it could mean “full head” as in the example of “一頭白髮 full head of white hairs” or could mean “one + classifier” as in the example of ”一頭牛 a cow”. For inherent ambiguous DMs, the sense derivation step will produce ambiguous sense representations and leave the final sense disambiguation until seeing collocation context, in particular seeing dependent heads. Some ambiguous representations are improbable sense combination. The improbable sense combinations should be eliminated during or after feature unification of D and M. For instance, although the determiner “一” has ambiguous senses of “one”, “first”, and “whole”, but “一公尺” has only one sense of “one meter”, so the other sense combinations should be eliminated.

The way we tackle the problem is that first we find all the ambiguous Ds and Ms by looking their definitions shown in the appendix A. We then manually design content and context dependent rules to eliminate the improbable combinations for each ambiguous D or M types. For instance, according to the appendix A, “頭” has 3 different E-HowNet representations while functions as determinant or measure, i.e. “def:{null}”, “def:{head|頭}”, and “def:ordinal={1}”. We write 3 content or context dependent rules below to disambiguate its senses.

(13) 頭”head”, Nfa, E-howNet: “def:{null}” : while E-HowNet of head word is “動物({animate|生物}” and it’s subclass.

(14) 頭“head“, Nff, E-howNet: “def:{頭}” : while pre-determinant is 一(Neqa)”one” or 滿”full” or 全”all” or 整”total”.

(15) 頭”first”, Nes, E-howNet: “def:ordinal={1}” : while this word is being a demonstrative determinatives which is a leading morpheme of the compound.

The disambiguation rules are shown in appendix B. In each rule, the first part is the word and its part-of-speech. Then the E-HowNet definition of this sense is shown, and followed by the condition constraints for this sense. If there is still ambiguities remained after using the disambiguation rule, we choice the most frequent sense as the result.

4. Experiment and Discussion

We want to know how good is our candidate production, and how good is our disambiguation rule. We randomly select 40628 sentences (7536 DM words) from Sinica Treebank as our development set and 16070 sentences (3753 DM words) as our testing set. We use development set for designing disambiguation rules and semantic composition rules. Finally, we derive 36 contextual dependent rules as our disambiguation rules. We randomly select 1000 DM words from testing set. We evaluate the composition quality of DMs with E-HowNet representation before disambiguation. For 1000 DM words, the semantic

composition rules produce 1226 candidates of E-HowNet representation from 939 words. The program fails to produce E-HowNet representations for the rest of 61 words because of undefined morphemes. There are 162 words out of the 939 words having ambiguous senses. The result shows that the quality of candidates is pretty good. Table 2 gives some examples of the result. For testing the correctness of our candidates, we manually check the format of 1226 candidates. Only 5 candidates out of 1226 are wrong or meaningless representations. After disambiguation processes, the resulting 1000 DM words in E-HowNet representation are judged manually. There are 880 correct E-HowNet representations for 1000 DM words in both sense and format. It is an acceptable result. Among 120 wrong answers, 57 errors are due to undefined morpheme, 28 errors are unique sense but wrong answer and the number of sense disambiguation errors is 36. Therefore accuracy of sense disambiguation is (162-36)/162=0.778.

Table 2. The result of semantic composition for DM compounds.

DM Compounds	E-HowNet Representation
二十萬元	def:role={money|貨幣:quantity={200000}}
另一個	def:qualification={other|另},quantity={1}
二百三十六分	def:role={分數:quantity={236}}
前五天	def:time={day| 日 :qualification={preceding| 上 次 }, quantity={5}}
一百一十六點七億美元	def:role={美元:quantity={11670000000}}

After data analysis, we conclude the following three kinds of error types.

A. Unknown domain error:

七棒”7th batter”, 七局”7th inning”

Because there is no text related to baseball domain in development set, we get poor performance in dealing with the text about baseball. The way to resolve this problem is to increase the coverage of disambiguation rules for the baseball domain.

B. Undefined senses and morphemes:

每三個“each three”

We do not define the sense of 每 ”each” and we only define 每 ”all”, so we have to add the sense of “each” in E-HowNet representation about 每.

有三位 ”there are three persons”, 同一個 ”the same”

Because 有 “have” and 同 ”the same” do not appear in our determinative list, it is not possible to composite their E-HowNet definitions.

C. Sense ambiguities:

In parsed sentence: NP(property:DM:上半場”first half ”|Head:DM:二十分”twenty

minutes or twenty points") . The E-HowNet representation of 二十分"twenty minutes or twenty points" can be defined as "def:role={分數:quantity={20}}" or "def:time={分鐘:quantity={20}}". More context information is needed to resolve this kind of sense ambiguity.

For unknown domain error and undefined rule, the solution is to expand the disambiguation rule set and sense definitions for morphemes. For sense ambiguities, we need more information to disambiguate the true sense.

5. Conclusion

E-HowNet is a lexical sense representational framework and intends to achieve sense representation for all compounds, phrases, and sentences through automatic semantic composition processing. In this paper, we take DMs as an example to demonstrate how the semantic composition mechanism works in E-HowNet to derive the sense representations for all DM compounds. We analyze morphological structures of DMs and derive their morphological rules in terms of regular expression. Then we define the sense of all determinatives and measure words in E-HowNet definition exhaustively. We make some simple composition rules to produce candidate sense representations for DMs. Then we review development set to write some disambiguation rules. We use these heuristic rules to find the final E-HowNet representation and reach 88% accuracy.

The major target of E-HowNet is to achieve semantic composition. For this purpose, we defined word senses of CKIP lexicon in E-HowNet representation. Then we try to automate semantic composition for phrases and sentences. However there are many unknown or compound words without sense definitions in the target sentences. DM compounds are occurring most frequently and without sense definitions. Therefore our first step is to derive the senses of DM words. In the future, we will use similar methods to handle general compounds and to improve sense disambiguation and semantic relation identification processing. We intend to achieve semantic compositions for phrases and sentences in the future and we had shown the potential in this paper.

Acknowledgement:

This research was supported in part by the National Science Council under a Center Excellence Grant NSC 96-2752-E-001-001-PAE and Grant NSC96-2221-E-001-009.

References

[1] Zhendong Don & Qiang Dong, 2006, *HowNet and the Computation of Meaning*. World Scientific Publishing Co. Pte. Ltd.

[2] 陳怡君、黃淑齡、施悅音、陳克健，2005b，*繁體字知網架構下之功能詞表達初探*，第六屆漢語詞彙語意學研討會，廈門大學

[3] Shu-Ling Huang, You-Shan Chung, Keh-Jiann Chen, 2008, *E-HowNet- an Expansion of HowNet*, The First National HowNet Workshop, Beijing, China.

[4] Li, Shih-Min, Su-Chu Lin, Chia-Hung Tai and Keh-Jiann Chen, 2006. *A Probe into Ambiguities of Determinative-Measure Compounds*, International Journal of Computational Linguistics & Chinese Language Processing, Vol. 11, No. 3. pp.245-280.

[5] Tai, J. H-Y, *Chinese classifier systems and human categorization*, *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, ed. by M. Y. Chen

and O J.-L. Tzeng, Pyramid Press, Taipei, 1994, pp. 479-494.

[6] Chao, Y.-R., *A grammar of Spoken Chinese,* University of California Press, Berkeley, 1968.

[7] Li, C. N. and S. A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*,

University of California Press, Berkeley, 1981.

[8] 何杰(He, J.), *現代漢語量詞研究*, 民族出版社, 北京市, 2002.

[9] 黃居仁, 陳克健, 賴慶雄(編著), *國語日報量詞典*, 國語日報出版社, 台北, 1997.

[10] Mo, R.-P., Y.-J. Yang, K.-J. Chen and C.-R. Huang, *Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation*, In Proceedings of ROCLING IV (R.O.C. Computational Linguistics Conference), 1991, National Chiao-Tung University, Hsinchu, Taiwan, pp. 111-134.

[11] Chen Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, 2005a, *Extended-HowNet- A Representational Framework for Concepts*, OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop, Jeju Island, South Korea

[12] Duchier, D., Gardent, C. and Niehren, J. (1999a) *Concurrent constraint programming in Oz for natural language processing*. Lecture notes, http://www.ps.uni-sb.de/~niehren/oz-natural-language-script.html.

Appendix A. Determinative and measure word in E-HowNet representation

定詞(Determinative word)

　　定指

D1-> 這、那、此、該、本、貴、敝、其、某、諸 def: quantifier={definite|定指}；這些、那些 def: quantifier={definite|定指}, quantity={some|些}

D2-> 第、首 def: ordinal={D4}

D3-> 上、前 def: qualification={preceding|上次}、下、後 def: qualification={next|下次}、頭、首 def:ordinal={1}、末 def: qualification={last|最後}、次 def:ordinal={2}

不定指

D4-> 一、二、萬、雙... def: quantity={1、2、10000、2...} or def:ordinal={1、2、10000、2...}

D5-> 甲、乙... def: ordinal={1、2...}

D6-> 其他、其餘、別、旁、他、另、另外 def: qualification={other|另}

D7-> 每、任何、一、全、滿、整、一切 def: quantity={all|全}

D8-> 各 def: qualification={individual|分別的}

D9-> 若干、有的、一些、部份、有些 def: quantity={some|些}

D10-> 半 def: quantity={half|半}

D11-> 多少、幾多 def: quantity={.Ques.}

D12-> 何、啥、什麼 def: fomal={.Ques.}

D13->數、許多、很多、好多、好幾、好些、多、許許多多、多數、大多數、不少、泰半、半數、諸多 def: quantity={many|多}、少許、少數、幾許、個把 def: quantity={few|少}

D14->餘、許、之多 def: approximate()、足、整、正 def: exact()、出頭、好幾、開外、多 def: over();

D15->0、1、2、3、4、5、6、7、8、9 def: quantity={1、2、3、4...}

量詞(Measure word)

有語意量詞(Measures with content sense)

Nff-> 暫時量詞—身、頭、臉、鼻子、嘴、肚子、手、腳 def:{身,頭, …}

Nff-> 暫時量詞—桌子、院子、地、屋子、池、腔、家子 def: position={桌子, 院子...:quantity={all|全}}

Nfe-> 容器量詞—盒(子)、匣(子)、箱(子)、櫃子、櫥(子)、籃(子)、簍(子)、爐子、包(兒)、袋(兒)、池子、瓶(子)、桶(子)、聽、罐(子)、盆(子)、鍋(子)、籠(子)、盤(子)、碗、杯(子)、勺(子)、匙(湯匙)、筒(子)、擔(子)、籮筐、 杓(子)、茶匙、壺、盅、筐、瓢、鍬、缸 def: container={盒,匣,...}

Nfg-> 標準量詞—

表長度的，如：公厘、公分、公寸、公尺、公丈、公引、公里、市尺、營造 尺、台尺、吋(inch)、呎(feet)、碼(yard)、哩(mile)、 (海)浬、庹、噚、尺、里、釐、寸、丈、米、厘、厘米、海 哩、英尺、英里、英呎、英寸、米突、米尺、微米、毫米、 英吋、英哩、光年。 def: length={公分,...}

表面積的，如：公畝、公頃、市畝、營造畝、坪、畝、分、甲、頃、平方公里、平方公尺、平方公分、平方尺、平方英哩、英畝。def: size={公畝,...}

表重量的，如：公克、公斤、公噸、市斤、台兩、台斤(日斤)、盎司(斯)、磅、公擔、公衡、公兩、克拉、斤、兩、錢、噸、克、英磅、英兩、公錢、毫克、毫分、仟克、公毫。def: weight={公克,...}

表容量的，如：公撮、公升(市升)、營造升、台升(日升)、盎司、品脫(pint)、加侖(gallon)、蒲式耳(bushel)、公斗、公石、公秉、公合、公勺、斗、毫升、夸、夸特、夸爾、立方米、立方厘米、立方公分、立方公寸、立方公尺、立分公里、立方英尺、石、斛、西西。def: volume={公撮,公升,...}

表時間的，如：微秒、釐秒、秒、秒鐘、分、分鐘、刻、刻鐘、點、點鐘、時、小時、更、夜、旬、紀(輪, 12 年) 、世紀、天(日)、星期(禮拜、週、周) 、月、月份、季、年(載、歲) 、年份、晚、宿、。def:temporal={微秒,月…}, 週年、周歲 def:duration={年}

表錢幣的，如：分、角(毛)、元(圓)、塊、兩、先令、盧比、法郎(朗)、辨士、馬克、鎊、盧布、美元、美金、便士、里拉、日元、台幣、港幣、人民幣。def: role={分, …,money|貨幣, …盧布…}

其他：刀、打(dozen)、令、綸(十條)、籮(gross)、大籮(great gross)、焦耳、千卡、仟卡、燭光、千瓦、仟瓦、伏特、馬力、爾格(erg)、瓦特、瓦、卡路里、卡、仟赫、位元、莫耳、毫巴、千赫、歐姆、達因、兆赫、法拉第、牛頓、赫、安培、周波、赫茲、分貝、毫安培、居里、微居里、毫居里。def: quantity={刀,打,…,焦耳,...}

Nfh-> 準量詞—

指行政方面，如：部、司、課、院、科、系、級、股、室、廳。def: location={部, 司...}

指時間方面，如：世、輩、輩子、代、學期、學年、年代 def: time={學期, 年代,...} 會、會兒、陣(子) 、下子 def: duration={TimeShort|短時間}

指方向的，如：面(兒)、方面、邊(兒)、方。def: direction={EndPosition|端}、頭(兒) def: direction={aspect|側}

指音樂的，如：拍、板、小節。def: quantity={拍,板...}

指頻率的，如:回、次、遍、趟、下、遭、響、圈、把、關、腳、巴掌、掌、拳頭、拳、眼、口、刀、槌、槌子、板、版子、鞭、鞭子、棒、棍、棍子、針、槍矛、槍、砲、度、輪、周、跤、回合、票。Def:frequency={D4, D15} 分 def:role={ 分 數 :quantity={D4,D15}} 、 步 def:{ 步 } 、 箭 def:role={箭:quantity={D4,D15}}、曲 def:{曲:quantity={D4,D15}}

Nfc-> 群 體 量 詞 — 對 、 雙 def:quantity={double| 複 } 、 列 (系 列) 、 排 def:quantity={mass| 眾 :manner={ InSequence| 有 序 }} 、 套 def:quantity={mass| 眾 :manner={relevant| 相 關 }} 、 串 def:quantity={mass| 眾 :dimension={linear| 線 }} 、 掛、幫、群、伙(夥)、票、批 def: quantity={mass|眾}、組 def: quantity={mass| 眾 : manner={relevant|相關}}、窩 def: quantity={mass|眾:cause={assemble|聚集}} 、 種 、 類、樣 def: {kind({object|物體})} 、簇 def:quantity={mass| 眾 :cause={assemble| 聚 集}}、疊 def:quantity={mass|眾:cause={pile|堆放}}、紮 def:quantity={mass| 眾:cause={wrap|包紮}}、叢 def:quantity={mass| 眾 :cause={assemble| 聚 集 }} 、 隊 def:quantity={mass|眾:manner={ InSequence|有 序}}、式 def:{kind({object|物體})}

Nfd->部分量詞—些 def:quantity={some|些}、部分(份)、泡、綹、撮、股、灘、汪、帶、截、節 def: quantity={fragment|部}、團 def: quantity={fragment| 部:shape={round|圓}}、堆 def: quantity={ fragment|部:cause={pile|堆放}}、把 def: quantity={ fragment|部:cause={hold|拿}}、層、重 def: quantity={ fragment| 部:shape={layered|疊}}

無語意量詞(null-sense Measures)

Nfa-> 個體量詞—本、把、瓣、部、柄、床、處、期、齣、場、朵、頂、堵、道、頓、錠、棟(幢)、檔(檔子)、封、幅、發、分(份)、服、個(箇)、根、行、戶、件、家、架、卷、具、闕、節、句、屆、捲、劑、隻、尊、盞、張、枝(支)、樁、幀、只、株、折、炷、軸、口、棵、款、客、輛、粒、輪、枚、面、門、幕、匹、篇、片、所、艘、扇、首、乘、襲、頭、條、台、挺、堂、帖、顆、座、則、冊、任、尾、味、位、頁、葉、房、彎、班、員、科、丸、名、項、起、間、題、目、招、股、回。def: {null}

Nfc-> 群體量詞—宗、番、畦、餐、行、副(付)、蓬、筆、房、綑(捆)、胎、嘟嚕、部、派、路、壟、落、束、席、色、攤、項。def: {null}

Nfd-> 部分量詞—口、塊、滴、欄、捧、抱、段、絲、點、片、縷、坨、匹、疋、階、抔、波、道。def: {null}

Nfb-> 述賓式合用的量詞—通、口、頓、盤、局、番。def: {null}

Nfi-> 動量詞—回、次、遍、趟、下、遭、番、聲、響、圈、把、仗、覺、頓、關、手、(巴)掌、拳(頭)、拳、眼、口、槌(子)、板(子)、鞭(子)、棒、棍(子)、陣、針、箭、槍(矛)、槍、砲、場、度、輪、曲、跤、記、回合、票。def: {null}

Nfh-> 準量詞

指書籍方面，如：版、冊、編、回、章、面、小節、集、卷。def: {null}

指筆劃方面，如：筆、劃(兒)、橫、豎、直、撇、捺、挑、剔、鉤(兒)、拐、點、格(兒)。def: {null}

其他：

程、作(例:一年有兩作)、倍、成。def: {null}

厘(例:年利五厘、一分一厘都不能錯)。def: {null}

毫(萬分之一)、絲(十萬分之一)(例:一絲一毫都不差)。

圍、指、象限、度。def: {null}

開(指開金)、聯(例:上下聯不對稱)。def: {null}

軍、師、旅、團、營、伍、班、排、連、球、波、端。def: {null}

回合、折、摺、流、等、票、桿、棒、聲、次。def: {null}

Appendix B. The rule for candidate disambiguation

head-based rule

e.g.一, Neu, def:quantity={1}, while part-of-speech of head word is Na, except the measure word is 身”body” or 臉”face” or 鼻子”nose” or 嘴”mouth” or 肚子”belly” or 腔”cavity” .

e.g.塊,Nfg,def:role={money|貨幣}, while E-HowNet representation of head word is “{money|貨幣}” or {null}, or head word is 錢”money” or 美金”dollar” or the s uffix of word is 幣”currency” and previous word is not D1.

塊,Nfd,def:{null}, otherwise, use this definition.

e.g.面,Nfa,def:{null}, while part-of-speech of head word is Nab.

面,Nfh,def:direction={aspect|側}, otherwise use this one.

e.g.頭,Nfa,def:{null}, while head word is Nab and E-HowNet representation of head word is “動物{animate|生物}” and it’s subclass.

頭,Nfh,def:direction={EndPosition|端} , if part-of-speech of head word is Na, do not use this definition. The previous word usually are 這”this” or 那”that” or 另”another”.

e.g.All Nfi, def:frequency={}, while part-of-speech of head word is Verb, i.e. E-HowNet representation of head word is {event|事件} and it’s subclass. Except POS V_2 and VG.

All Nfi,def:{null}, while part-of-speech of head word is Noun, i.e. E-HowNet of head word is {object|物體} and it’s subclass.

e.g.部, 股…,Nfh,def:location={ }, if part-of-speech of head word is Na or previous word is 這”this” or 那”that” or 每”every”, do not use this definition.

部,股…,Nfa,def:{null}, otherwise use this definition.

e.g. 盤 ,Nfe,def:container={plate| 盤 },while head word is food, i.e. E-HowNet representation of head word is {edible|食物} and it’s subclass.

盤,Nfb,def:{null},otherwise use this one.

e.g. 分 ,Nfg, def:role={ 分 }, while head word is 錢 “money”, i.e. E-HowNet representation of head word is {money|貨幣} and it’s subclass.

分,Nfg, def:size={ 分 }, while head word is 地 “land”, i.e. E-HowNet representation of head word is {land|陸地} and it’s subclass.

分,Nfa, def:{null}, while part-of-speech of head word is Na or Nv. For example: 一分耕耘；十分力氣；五分熟.

e.g.點,Nfh;Nfd,def:{null}, while part-of-speech of head word is Nab. If part-of-speech of head word is V, Naa or Nad, do not use this definition.

collocation-based rule

e.g.分,Nfh,def:role={score|分數:quantity={D4,D15}}, while the sentence also contains the words 考 ”give an exam” (E-HowNet representation is {exam|考試}) or 得 ”get” (E-HowNet representation is {obtain|得到}) or 失”lose” (E-HowNet representation is {lose|失去}), then use this definition.

e.g.分,Nfg,def:time={minute|分鐘}, if the sentence contains the word 時”hour” or 鐘頭”hour”.

e.g.兩,Nfg,def:weight={兩}, if the sentence contains the word 重”weight” or 重量”weight”.

兩,Nfg,def:role={money|貨幣}, if the sentence contains the word 銀”sliver” or 錢”money” or 黃金”gold”

pre-determinant-based rule

e.g.頭, Nff,def:{head|頭}, while pre-determinant is 一(Neqa)”one” or 滿”full” or 全”all” or 整”total”.

e.g.腳, Nff,def:{leg|腳}, while pre-determinant is 一(Neqa)”one” or 滿”full” or 全 “all” or 整”total” and part-of-speech of head word is not Na.

腳, Nfi,def:frequency={}, while part-of-speech combination is V+D4,D15+腳.

e.g.點,Nfg, def:time={點}, while part-of-speech of pre-determinant is D4 or D15(1~24) and part-of-speech of previous word is not D1 or previous word is not 有”have”.

e.g.輪,Nfg,def:time={輪}, while pre-determinant is 第 ” a function word placed in front of a cardinal number to form an ordinal number” or 首”first”.

determinative-based rule

e.g.一、二...1、2...兩..., Neu, def:ordinal={}, the determinant of word is 第, 民國, 公元, 西元, 年號, 一九 XX or 12XX, (four digits number).

一、二...1、2...兩..., Neu,def:quantity={}, otherwise use this definition.

e.g.頭,Nes,def:ordinal={1},the word 頭”head” is determinant word.

e.g.兩,Neu,def:quantity={}, the word 兩”a unit of weight equal to 50 grams” is determinant word.

measure word based rule

e.g.一,Neqa,def:quantity={all|全}, the part-of-speech of the measure word behind 一 is Nff, or the suffix of the measure word is 子, (for example,櫃子” cabinet”, 瓶子”bottle”)or 籮筐” large basket”.

A Thesaurus-Based Semantic Classification of English Collocations

Chung-chi Huang, Chiung-hui Tseng, Kate H. Kao, Jason S. Chang
ISA, National Tsing Hua University
{u901571, smilet, msgkate, jason.jschang}@gmail.com

Abstract

We propose a new method for organizing the numerous collocates into semantic thesaurus categories. The approach introduces a thesaurus-based semantic classification model automatically learning semantic relations for classifying adjective-noun (A-N) and verb-noun (V-N) collocations into different categories. Our model uses a random walk over weighted graph derived from WordNet semantic relation. We compute a semantic label stationary distribution via an iterative graphical algorithm. The performance for semantic cluster similarity and the conformity of semantic labels are both evaluated. The resulting semantic classification establishes as close consistency as human judgments. Moreover, our experimental results indicate that the thesaurus structure is successfully imposed to facilitate grasping concepts of collocations. It might improve the performance of the state-of-art collocation reference tools.

Keywords: Collocations, Semantic classification, Semantic relations, Random walk algorithm, Meaning access index.

1. Introduction

Submitting queries (e.g., a search keyword "*beach*" for a set of adjective collocates) to collocation reference tools typically return many collocates (e.g., collocate adjectives with a pivot word "*beach*": "*rocky*", "*golden*", "*beautiful*", "*pebbly*", "*splendid*", "*crowded*", "*superb*", etc.) extracted from a English corpus. Applications of automatic extraction of collocations such as *TANGO* (Jian, Chang & Chang, 2004) have been created to answer queries of collocation usage.

Unfortunately, existing collocation reference tools sometimes present too much information in a batch for a single screen. With web corpus sizes rapidly growing, it is not uncommon to find thousands collocates for a query word. An effective reference tool might strike a balance between quantity and accessibility of information. To satisfy the need for presenting a digestible amount of information, a promising approach is to automatically partition words into various categories to support meaning access to search results and thus give a thesaurus index.

Instead of generating a long list of collocates, a good, better presentation could be composed of clusters of collocates inserted into distinct semantic categories. We present a robust thesaurus-based classification model that automatically group collocates of a given pivot word focusing on: (1) the adjectives in adjective-noun pairs (A-N); (2) the verbs in verb-noun pairs (V-N); and (3) the nouns in verb-noun pairs (V-N) into semantically related classes.

Our model has determined collocation pairs that learn the semantic labels automatically during random walk algorithm by applying an iterative graphical approach and partitions collocates for each collocation types (A-N, V-N and V-N mentioned above). At runtime, we start with collocates in question with a pivot word, which is to be assigned under a set of semantically

related labels for the semantic classification. An automatic classification model is developed for collocates from a set of A-N and V-N collocations. A random walk algorithm is proposed to disambiguate word senses, assign semantic labels and partition collocates into meaningful groups.

As part of our evaluation, two metrics are designed. We assess the performance of collocation clusters classified by a robust evaluation metric and evaluate the conformity of semantic labels by a three-point rubric test over collocation pairs chosen randomly from the results. Our results indicate that the thesaurus structure is successfully imposed to facilitate grasping concepts of collocations and to improve the functionality of the state-of-art collocation reference tools.

2. Related Work

2.1 Collocations

The past decade has seen an increasing interest in the studies on collocations. This has been evident not only from a collection of papers introducing different definitions of the term "collocation" (Firth, 1957; Benson, 1985; Lewis, 1997), but also from a number of research on collocation teaching/acquisition associating to language learning (Lewis, 2000; Nation, 2001). When analyzing Taiwanese EFL writing, Chen (2002) and Liu (2002) investigated that the common lexical collocational error patterns include verb-noun (V-N) and adjective-noun (A-N). Furthermore, with the technique progress of NLP, Word Sketch (Kilgarriff & Tugwell, 2001) or *TANGO* (Jian, Chang & Chang, 2004) became the novel applications as collocation reference tools.

2.2 Meaning Access Indexing

Some attention has been paid to the investigation of the dictionary needs and reference skills of language learners (Scholfield, 1982; Béjoint 1994), especially the structure for easy comprehending. According to Tono (1992 & 1997), menus that summarize or subdivide definitions into groups ahead of entries in dictionaries would help users with limited reference skills. The System "Signposts" of the *Longman Dictionary of Contemporary English*, 3rd edition, the index "Guide Word" of the *Cambridge International Dictionary of English*, as well as the "Menus" of the *Macmillan English Dictionary for Advanced Learners* all value the principle.

2.3 Similarity of Semantic Relations

The construction of practical, general word sense classification has been acknowledged to be one of the most ambitious and frustrating tasks in NLP (Nirenburg & Raskin, 1987), even *WordNet* with more significant contribution of a wide range of lexical-semantic resources (Fellbaum, 1998). Lin (1997) presented an algorithm for word similarity measure by its distributional similarity. Unlike most corpus-based word sense disambiguation (WSD) algorithms where different classifiers are trained for separate words, Lin used the same local context database as the knowledge sources for measuring all word similarities. Distributional similarity allows pair wise word similarity measure to deal with infrequent words or unknown proper nouns. However, compared to distributional similarity measure, our model by random walk algorithm has remarkable feature to deal with any kind of constraints, thus, not limited to pair-wise word similarities, and can be improved by adding any algorithm constraints available.

More specifically, the problem is focused on classifying semantic relations. Approaches presented to solve problems on recognizing synonyms in application have been studied (Lesk, 1986; Landauer and Dumais, 1997). However, measures of recognizing collocate similarity are not as well developed as measures of word similarity, the potential applications of semantic classification are not as well known. Nastase and Szpakowicz (2003) presented how to

automatically classify a noun-modifier pair, such as "laser printer", according to the semantic relation between the head noun (printer) and the modifier (laser). Turney (2006) proposed the semantic relations in noun pairs for automatically classifying. As for VerbOcean, a semi-automatic method was used to extract fine-grained semantic relations between verbs (Chklovski & Pantel, 2004). Hatzivassiloglou and McKeown (1993) presented a method towards the automatic identification of adjectival scales. More recently, Wanner et al. (2006) has sought to semi-automatically classify the collocation from corpora by using the lexical functions in dictionary as the semantic typology of collocation elements. Nevertheless, there is still a lack of fine-grained semantically-oriented organization for collocation.

3. Methodology

We focus on the preparation step of partitioning collocations into categories: providing each word with a semantic label and thus presenting collocates under thesaurus categories. The collocations with the same semantic attributes by the batch size are then returned as the output. Thus, it is crucial that the collocation categories be fairly assigned for users' easy-access. Therefore, our goal is to provide a semantic-based collocation thesaurus that automatically adopts characterizing semantic attributes. Figure 1 shows a comprehensive framework for our unified approach.

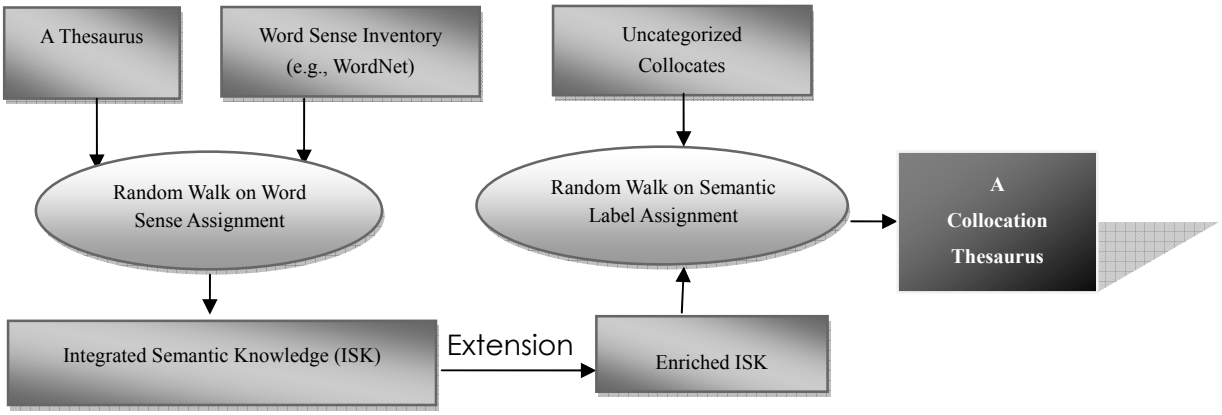


Figure 1.A comprehensive framework for our classification model.

3.1 Problem Statement

We are given (1) a set of collocates $Col = \{C_1, C_2, \ldots, C_n\}$ (e.g., *sandy, beautiful, superb, rocky, etc.*) denoted with a set of part-of-speech tags P, $\{P \in Pos \mid P =$ adjective P_{adj}, verb P_v, or noun $P_n\}$ for a pivot word X (e.g., *beach*) extracted from a corpus of English texts (e.g., *British National Corpus*); (2) a combination of thesaurus categories (e.g., *Roget's Thesaurus*), $TC = \{(W, P, L) \mid W \in Voc, P \in Pos, L \in Cat\}$, where *Voc* is the thesaurus vocabulary words W, ordered by general-purpose topics hereinafter called the semantic labels (e.g., feelings, materials, art, food, time, etc.), $Cat = \{L_1, L_2, \ldots, L_m\}$, with conceptual-semantic attributes as the basis for organization; and (3) a lexical database (e.g., *WordNet*) as our word sense inventory *SI* for semantic relation population. *SI* is equipped with a measure of semantic relatedness of W, REL(S, S') encoding semantic relations REL $\in$ *SR* holding between word sense S and S'.

Our goal is to partition *Col* into subsets *Sub* of similar collocates, $Sub \subseteq Col$, by means of an integrated semantic knowledge crafted from the mapping of *TC* and *SI* that is likely to express closely related meanings of *Col* in the same context of X mentioned herein *beach*. For this, we use a graph-based algorithm to give collocations a thesaurus index by giving each collocate in *Col* a semantic label L.

3.2 Learning to Build a Semantic Knowledge by Iterative Graphical Algorithms

Recall that we attempt to provide each word with a semantic label and partition collocations into thesaurus categories. In order to partition a large-scale collocation input and reduce the out-of-vocabulary (OOV) words occurred, automating the task of building an integrated semantic knowledge base is a necessary step, but also imposes a huge effort on the side of knowledge integration and validation. An integrated semantic knowledge (*ISK*) is defined to interpret a word in triples (*W*, *L*, *S*), i.e., the given word, a semantic label representing one of thesaurus categories, and its corresponding word sense, as cognitive reference knowledge. At this first stage, interconnection is still between words and labels from the given thesaurus category *TC* and not between word senses and semantic labels. For interpreting words in triples (*W*, *L*, *S*) as an *ISK* and corresponding to the fact that there's a limited, almost scarcely found, resource that is intended for such semantic knowledge, we proceeded as follows to establish one comprehensive *ISK* allowing concentrating on our task of populating it with new semantic relations between words and labels, overcoming the problem of constructing a resource from scratch.

3.2.1 Word Sense Assignment for Integrated Semantic Knowledge

In the first stage of the learning process, we used a graph-based sense linking algorithm which automatically assigns senses to all words under a thesaurus category by exploiting semantic relations identified among word senses. It creates a graph of vertices representing a set of words and their admissible word senses in the context of a semantically consistent list. The pseudo code for the algorithm is shown as Figure 2.

By adding synonymous words through semantic relations, it can broaden the word coverage of *TC*, which may reduce significantly the number of OOV words in *TC* and cope with the problem of collocates that form a group by itself. This strategy relies on a set of general-purpose topics as semantic labels *L* in a thesaurus category *TC* and a word sense inventory *SI* encoding semantic relations. *TC* and *SI* are derived from separate lexicographical resources, such as *Longman Lexicon of Contemporary English* and *WordNet*.

The algorithm assumes the availability of a word sense inventory *SI* encoding a set of semantic relations as a measure of semantic relatedness. Given a set of words with corresponding admissible senses in *SI*, we build a weighted graph $G = (V, E)$ for *SI* such that there is a vertex *V* for each admissible sense, and a directed edge *E* for each semantic relation between a pair of senses (vertices).

The input to this stage is a word sense inventory *SI* encoding a set of semantic relations *SR* attributing the senses of *SI*, and a set of words $W = \{w_1, w_2, \ldots, w_n\}$ listed under L_i in a set of semantic labels *Cat* used in a thesaurus *TC*. The semantic relations *SR* comprise REL(*S*, *S'*) where *S* and *S*' are admissible senses in *SI*, and REL is a semantic relation (e.g., synonyms, hypernyms, and hyponyms holding between senses) existing between *S* and *S'* and explicitly encoded in *SI*. Notice that semantic relations typically hold between word senses but not necessarily between words. We apply semantic relations to identify the intended senses for each word in the list. Accordingly these intended senses will form a semantically consistent set with maximal interconnecting relations

We use random walk on the weighted graph *G* encoding admissible senses as vertices *V* and semantic relations *SR* as edges *E* with a view to discovering the most probable sense S^* for *W*. The edges will be stepped through by imaginary walkers during the random walk in a probabilistic fashion. Through the random walk on *G*, the probability of intended senses will converge to a higher than usual level because of the influx via incoming edges representing semantic relations. All vertices in the weighted graph *G* start with a uniform probability distribution. The probability is reinforced by edges that participate in a *SR* until the reinforcement of probability converges for the given sense consistency, leading to a stationary

distribution over sense probability P_s, represented as scores Q_s attached to vertices in the graph. In all, the weights on G indicating the sense strength converge to arrive at the consistency of senses, which become the output of this learning stage. The procedure is repeated for all word lists in TC. Recall that these most probable senses are useful for extending the limited coverage of TC and reducing the number of OOV words effectively.

Algorithm 1. Graph-based Word Sense Assignment

Input: A word W from a set annotated with a semantic label L under a category Cat from a thesaurus TC; A word sense inventory SI with a measure of semantic relatedness of W, REL (S, S') encoding semantic relations REL $\in$ SR holding between word meanings S and S'.
S is one of the admissible senses of W listed in SI, and so as S' of W'.

Output: A list of linked word sense pairs (W, S*)

Notation: Graph $G = \{V, E\}$ is defined for admissible word senses and their semantic relations, where a vertex $v \in V$ is used to represent each sense S whereas an edge in E represents a semantic relation in SR between S and S'. Word sense inventory SI is organized by semantic relations SR, where REL (S, S'), REL $\in$ SR is used to represent one of the SR holding between word sense S of W and S' of W'.

PROCEDURE AssignWordSense(L,SI)

Build weighted graph G of word senses and semantic relations

(1) INITIALIZE V and E as two empty sets
FOR each word W in L
FOR each of n admissible word sense S of W in SI, $n = n(W)$
ADD node S to V
FOR each node pair (S,S') in $V \times V$
IF (S REL S') $\in$ SR and $S \neq S'$ THEN ADD edge E(S,S') to E
FOR each word W AND each of its word senses S in V
(2) INITIALIZE $P_s = 1/n(W)$ as the initial probability
(2a) ASSIGN weight (1-d) to matrix element $M_{S,S}$
(2b) COMPUTE $e(S)$ as the number of edges leaving S
FOR each other word $W' \neq W$ in L AND each of W' senses S'
(3) IF E(S,S') $\in$ E THEN ASSIGN Weight $d/e(S)$ to $M_{S,S'}$
OTHERWISE ASSIGN 0 to $M_{S,S'}$

Score vertices in G

REPEAT
FOR each word W AND each of its word senses S in V
(4) INTIALIZE Q_S to $P_S * M_{S,S}$
FOR each other word $W' \neq W$ in L AND each of W' senses S'
(4a) INCREMENT Q_S by $P_{S'} * M_{S',S}$
FOR each word W AND
Sum Q_S over $n(W)$ senses as N_w
FOR each sense S of W
(4b) Replace P_S by Q_S/N_w so as normalize to sum to 1
UNTIL probability P_S converges

Assign word sense

(5) INITIALIZE $List$ as NULL
FOR each word W
(6) APPEND (W,S*) to $List$ where S* maximizes P_s
(7) OUTPUT $List$

Figure 2.Algorithm for graph-based word sense assignment.

The algorithm (referring to Figure 2) for the best sense assignment S^* for W consists of three main steps: (1) construction of a word sense graph; (2) sense scoring using graph-based probability ranking algorithm; and (3) word sense assignment.

In Step 1, the weighted graph $G = (V, E)$ is built by populating candidate $n(W)$ admissible senses S of each given word W as vertices from SI, such that for each word W and its sense S, there is a vertex V for every intended sense S. In addition, the edge $E(S, S')$ in E, a subset of $V \times V$, is built up by adding a link from vertex S to vertex S' for which a semantic relation REL(S, S') between the two vertices is derived, where S is one of the admissible senses of W and S' of W'.

In Step 2, we initialize the probability P_s to a uniform distribution over each vertex S. And we set the weight of self-loop edge as (*1-d*) (Step 2a), and the weights of other outbound edges as $d/e(S)$, calculated as $Q_s = Q_s + \frac{d \times p_{s'}}{e(S')}$

In our ranking algorithm for the weighted graph, the decision on what edge to follow during a random walk considers the weights of outbound edges. One with a higher probability follows an edge that has a larger weight. The ranking algorithm is particularly useful for sense assignment, since the semantic relations between pairs of senses (vertices) are intrinsically modeled through weights indicating their strength, rather than a decision on binary 0/1 values.

As described in Step 3, the weights are represented as a matrix M for which the weights of all outbound edges from S are normalized to sum to 1. Our random walk algorithm holds that an imaginary walker who is randomly stepping over edges will eventually stop walking. The probability, at any step, that the walker will continue is a damping factor, a parameter usually denoted by d. The d factor is defined as the vertex ratio of the outgoing edges and the self-loop edge as the result of dividing the vertex weight of the damping constant. The damping factor is subtracted from 1. The value for (*1-d*) introduced is the principal eigenvector for the matrix M. The value of the eigenvector is fast to approximate (a few iterations are needed) and in practice it yields fairly optimal results. In the original definition of a damping factor introduced by PageRank (Brin and Page, 1998), a link analysis algorithm, various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85 whereas we use variant value for d in our implementation.

In Step 4 of vertex scoring, we compute the probabilistic values of each vertex at every iteration. The set of probabilities Q_s of each sense S for the next iteration is computed by multiplying the current probability P_s with the matrix $M_{s,s}$. For instance (Step 4a), suppose a walker is to start at one vertex of the graph. The probability of Q_s is the probability of a walker stands at a vertex of S forming a self-loop plus the sum of the influx of $P_{s'}$ weighted by $M_{s',s}$. In Step 4b, we normalize Q_s for the probability of all admissible senses with each word to sum to 1 and replace P_s by Q_s.

The normalized weighted score is determined as: $P_s(W) = \frac{Q_s(W)}{\sum_{l \in senses(W)} Q_l(W)}$

Subsequently, in Step 5, we calculate the ranking score of maximum probability P_s that integrates the scores of its start node. And thus the resulting stationary distribution of probabilities can be used to decide on the most probable set of admissible senses for the given word. For instance, for the graph drawn in Figure 3, the vertex on the vertical axis represented as the *sense #3* of "*fine*" will be selected as the best sense for "*fine*" under the thesaurus category "*Goodness*" with other entry words, such as, "*lovely*", "*superb*", "*beautiful*", and "*splendid*". The output of this stage is a set of linked word sense pairs (W, S*) that can be used to extend the limited thesaurus coverage. The overall goal of ranking admissible senses is to weight highly the senses that tend to arrive at the consistency of word senses.

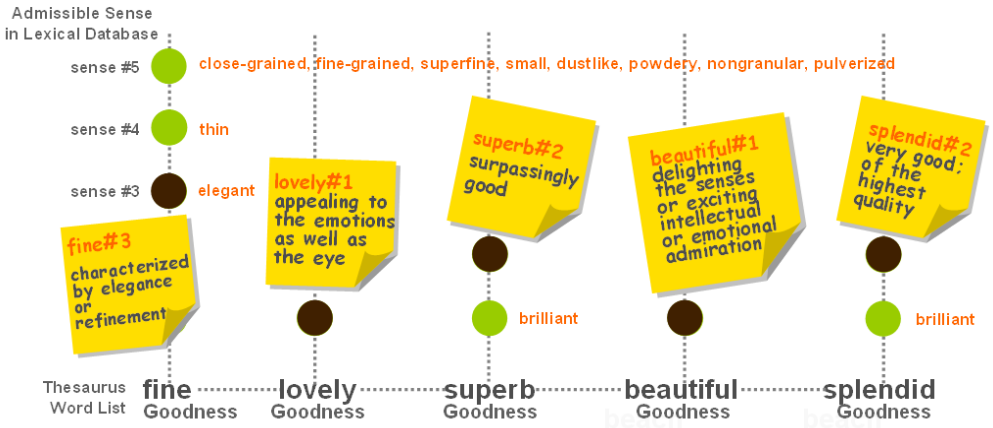


Figure 3.Highest scoring word sense under category "*Goodness*" assigned automatically by random walk.

Recall that our goal is to select the word senses for each specific collocate, categorized by the corresponding semantic label, for example, *sandy, rocky, pebbly beach* with label *Materials*; *beautiful*, *lovely*, *fine*, *splendid*, *superb* beach with *Goodness*. In order for the word coverage under thesaurus category to be comprehensive and useful, we need to expand the words listed under a label. This output dataset of the learning process is created by selecting the optimal linked word sense pairs (*W*, *S**) from each semantic relation in our word sense inventory where the specific semantic relation is explicitly defined.

Although alternative approaches can be used to identify word senses of given words, our iterative graphical approach has two distinctive advantages. First, it enables a principled combination of integrated similarity measure by modeling through a multiple types of semantic relations (edges). Secondly, it transitively merits local aggregated similarity statistics across the entire graph. To perform sense propagation, a weighted graph was constructed. On the graph, interconnection of edges is aggregated on a semantic relatedness level by random walk. The sense edge voltage is transitively propagated to the matching sense vertex. The effect depends on the reinforcement of the semantic relations (edges) and magnitude of the sense relations (vertices), creating a flexible amplitude-preserving playground like no other optional way of modeling a transcended graph propagation of senses. By doing so, our model is carved out to be a robust, more flexible solution with possible alternatives of combining additional resources or more sophisticated semantic knowledge. This approach is relatively computationally inexpensive for unsupervised approach to the WSD problem, targeting the annotation of all open-class words in lexical database using information derived exclusively from categories in a thesaurus. The approach also explicitly defines semantic relations between word senses, which are iteratively determined in our algorithm.

3.2.2 Extending the Coverage of Thesaurus

Automating the task of building a large-scale semantic knowledge base for semantic classification imposes a huge effort on the side of knowledge integration and validation. Starting from a widespread computational lexical database such as *WordNet* overcomes the difficulties of constructing a knowledge base from scratch. In the second stage of the learning process, we attempt to broaden the limited thesaurus coverage as the basis of our applied semantic knowledge that may induce to unknown words in collocation label assignment in Section 3.3. The sense-annotated word lists generated as a result of the previous step are useful for extending the thesaurus and reducing OOV words that may render words that form a group by itself.

In the previous learning process, "*fine*" with other adjective entries "*beautiful*, *lovely*,

splendid, *superb*" under semantic label "*Goodness*" can be identified as belonging to the word sense *fine#3* "*characterized by elegance or refinement or accomplishment*" rather than other admissible senses (as shown in Table 1). Consider the task of adding similar word to the set of "*fine#3*" in the thesaurus category "*Goodness*". We apply semantic relation operators for novel word extension for "*fine#3*". Some semantic relations and semantic operators available in the word sense inventory are shown in Table 2.

In this case, "***similar_to***", the semantic relation operator of "*fine#3*" can be applied to derive similar word "*elegant#1*" as the extended word for "*fine#3*" identified with the sense definition "*characterized by elegance or refinement*".

Table 1.Admissible senses for adjective "*fine.*"

Sense Number	Definition	Example	Synsets of Synonym
fine #1	(being satisfactory or in satisfactory condition)	*"an all-right movie"; "everything's fine"; "the passengers were shaken up but are all right"; "things are okay"*	all ight#1, o.k.#1,ok#1, okay#1
fine #3	(characterized by elegance or refinement or accomplishment)	*"fine wine" ; "a fine gentleman"; "fine china and crystal"; "a fine violinist"*	elegant#1
fine #4	(thin in thickness or diameter)	*"a fine film of oil"; "fine hairs"; "read the fine print"*	thin#1

Table 2.Some semantic operators in word sense inventory.

***SR* Operators**	**Description**	**Relations Hold for**
syn operator	synonym sets for every word that are interchangeable in some context	all words
sim operator	adjective synsets contained in adjective clusters	adjectives

3.3 Giving Thesaurus Structure to Collocation by Iterative Graphical Algorithms

The stage takes full advantage of the foundation built in the prior learning process, established an extended semantic knowledge to build a thesaurus structure for online collocation reference tools. We aim to partition collocations in groups according to semantic relatedness by exploiting semantic labels in a thesaurus and assign each collocate to a thesaurus category.

In this stage of the process, we apply the previously stated random walk algorithm and automatically assign semantic labels to all collocations by exploiting semantic relatedness identified among collocates. By doing so, our approach for collocation label assignment can cluster collocations together in groups, which is helpful for dictionary look-up and learners to find their desired collocation or collocations under a semantic label.

We use a set of corresponding admissible semantic labels L to assign labels under thesaurus category $L \in Cat$ to each collocate $C \in Col$, such that the collocates annotated with L can be partitioned into a subset corresponding to a thesaurus category, $Sub = \{ (C, L) \mid C \in Col, L \in Cat \in TC \}$, which facilitate meaning-based access to the collocation reference for learners. We define a label graph $G = (V, E)$ such that there is a vertex $v \in V$ for every admissible label L of a given collocate C, and there is an edge $e \in E$ between two vertices where the two vertices have the same label. Edge reinforcement of the label (vertex) similarity distance between pairs of labels is represented as directed edges $e \in E$, defined over the set of vertex pairs $V \times V$. Such semantic label information typically lists in a thesaurus.

Given such a label graph G associated with a set of collocates *Col*, the probability of each

label P_L can be iteratively determined using a graph-based ranking algorithm, which runs over the graph of labels and identifies the likelihood of each label (vertex) in the graph. The iterative algorithm is modeled as a random walk, leading to a stationary distribution over label probabilities P_L, represented as scores Q_L attached to vertices in the graph. These scores Q_L are then used to identify the most probable semantic label L_* for each collocate C, resulting in a list of annotations (C, L_*) for all collocates in the input set. The algorithm is quite similar to the one for graph-based word sense assignment shown in Figure 2. But note that the overall goal of ranking admissible labels is to weight highly the semantic labels that help arrange collocations in a thesaurus category and provide learners with a thesaurus index.

In other word, our goal is to assign corresponding semantic labels to each specific collocate, for example, "*sandy, rocky, pebbly beach* with label *Materials.*" In order for the semantic structure to be comprehensive and useful, we try to cover as much OOV words as possible by applying semantic relation operators (e.g., derivational relations). We propose the replacement of OOV words for their derivational words such as the replacement of "rocky" for "rock" and "dietary" for "diet". For a few number of derivationally substitutable OOV words occurred, such as *pebbly beach*, we apply the built-in vocabulary of words, i.e., *pebble*, as a substitution for *pebbly* by exploiting the derivational relations from the obtainable sense inventory as we will discuss in more detail in the section of experimental set-up.

The output of this stage is a list of linked label-annotated collocate pairs (C, L*) that can be used to classify collocations in categories.

4. Experimental Settings

4.1 Experimental Data

In our experiments, we applied random walk algorithm to partitioning collocations into existing thesaurus categories, thus imposing a semantic structure on the raw data. In analysis of learners' collocation error patterns, the types of verb-noun (V-N) and adjective-noun (A-N) collocations were found to be the most frequent error patterns (Liu, 2002; Chen, 2002). Hence, for our experiments and evaluation, we focused our attention particularly on V-N and A-N collocations.

Recall that our classification model starts with a thesaurus consisting of lists of semantic related words extended by a word sense inventory via random walk Algorithm. Then, the extended semantic knowledge provides collocates with topic labels for semantic classification of interest. Preparing the semantic knowledge base in our experiment consists of two main steps: (1) Integration, and (2) Extension. Two kinds of resources are applied as the input data of this learning process of semantic knowledge integration described below.

4.1.1 Input Data 1: A Thesaurus for Semantic Knowledge Integration

We selected the set of thesaurus categories from the dictionary of *Longman Lexicon of Contemporary English* (*LLOCE*). *LLOCE* contains 15,000 distinct entries for all open-class words, providing semantic fields of a pragmatic, everyday common sense index for easy reference. The words in *LLOCE* are organized into approximately 2,500 semantic word sets. These sets are divided into 129 semantic categories and further organized as 14 semantic fields. Thus the semantic field, category, and semantic set in *LLOCE* constitute a three-level hierarchy, in which each semantic field contains 7 to 12 categories and each category contains 10 to 50 sets of semantic related words. The *LLOCE* is based on coarse, topical semantic classes, making them more appropriate for WSD than other finer-grained lexicon.

4.1.2 Input Data 2: A Word Sense Inventory for Semantic Knowledge Extension

For our experiments, we need comprehensive coverage of word senses. Word senses can be

easily obtained from any definitive records of the English language (e.g. an English dictionary, encyclopedia or thesaurus). In this case, we applied *WordNet* to broaden our word coverage from 15,000 to 39,000. *WordNet* is a broad-coverage machine-readable lexical database, publicly available in parsed form (Fellbaum, 1998). *WordNet* 3.0 lists 212,557 sense entries for open-class words, including nouns, verbs, adjectives, and adverbs. In order to extend the sense coverage, we applied random walk Algorithm to match a significant and manageable portion of the *WordNet* sense inventory to the *LLOCE* thesaurus.

WordNet can be considered a graph over synsets where the word senses are populated as vertices and the semantic relations edges. *WordNet* is organized by the sets of synsets; a synset is best thought of as a concept represented by a small set of synonymous senses: the adjective {*excellent*, *first-class*, *fantabulous*, *splendid*}, the noun {*enemy*, *foe*, *foeman*, *opposition*}, and the verb {*fight*, *contend*, *struggle*} form a synset.

4.2 Experimental Configurations

We acquired all materials of the input data (1) and (2) to train and run the proposed model, using the procedure and a number of parameters as follows:

4.2.1 Step 1: Integrating Semantic Knowledge

To facilitate the development of integrated semantic knowledge, we organize synsets of entries in the first input data, *LLOCE*, into several thesaurus categories, based on semantic coherence and semantic relations created by lexicographers from *WordNet*. The integrated semantic knowledge can help interpret a word by providing information on its word sense and its corresponding semantic label, (i.e., "*fine*" tagged with "*Materials*").

Recall that our model for integrating word senses and semantic labels is based on random walk algorithm on a weighted directed graph whose vertices (word senses) and edges (semantic relations) are extracted from *LLOCE* and *WordNet* 3.0. All edges are drawn as semantic relatedness among words and senses, derived using the semantic relation operators (Table 3).

Table 3.The semantic relation operators used to link the lexical connection between word senses.

Relation Operators	Semantic Relations for Word Meanings	Relations Hold for
Syn operator	synonym sets for every word that are interchangeable in some context without changing the truth value of the preposition in which they are embedded	all words
hyp operator	hypernym/hyponym (superordinate/subordinate) relations between synonym sets	nouns verbs
vgp operator	verb synsets that are similar in meaning and should be grouped together when displayed in response to a grouped synset search.	verbs
Sim operator	adjective synsets contained in adjective clusters	adjectives
der operator	words that have the same root form and are semantically related	all words

In particular for all semantic relation operators, we construct a maximum allowable edge distance *MaxED*, informing a constraint over the edge path between words for which the word sense likelihood is sought. For our experiments, the *MaxED* is set to 4.

4.2.2 Step 2: Extending Semantic Knowledge

Once we have mapped the sense-label from the stationary distribution in the random walk graph, another step is taken to take advantage of the mapped semantic knowledge by adding

more novel words to the thesaurus categories. The word coverage in question is extended by more than twice as many *LLOCE* thesaurus entries. For the extension of our semantic knowledge, we need information on joint word sense and semantic label pairs, and semantic relation among words from the previous step. Various kinds of the above-mentioned semantic relation operators can be derived, depending on the type of semantic operators available for the word class at hand. In experiments, we focus on the synset operation provided in *WordNet*.

4.3 Test Data

We used a collection of 859 V-N and A-N collocation pairs for testing, obtained from the website, *JustTheWord* (http://193.133.140.102/JustTheWord/). *JustTheWord* clusters collocates into sets without understandable label. As a result, we will compare the performance of our model with *JustTheWord* in Section 5

We evaluated semantic classification of three types of collocation pairs, focusing on **A**-N, **V**-N and V-**N**. We selected five pivot words for each type of collocation pairs for their varying level of abstractness and extracted a subset of their respective collocates from the *JustTheWord*. Among 859 testing pairs, 307 collocates were extracted for **A**-N, 184 for **V**-N, and 368 for V-**N**.

To make the most appropriate selection from testing data in *JustTheWord*, we have been guided here by research into language learners' and dictionary users' needs and skills for second language learning, taking account especially of the meanings of complex words with many collocates (Tono, 1992; Rundell, 2002). The pivot words we selected for testing are words that have many respective collocations and are shown in boxes around each entry in *Macmillan English Dictionary for Advance Learners.*

5. Results and Discussions

Two pertinent sides were addressed for the evaluation of our results. The first was whether such a model for a thesaurus-based semantic classification could generate collocation clusters based on human-like word meaning similarities to a significant extent. Second, supposing it did, would its success of semantic label assignment also strongly excel in language learner collocation production? We propose innovative evaluation metrics to examine our results respectively in these two respects and assess whether our classification model can reliably cluster collocates and assign a helpful label in terms of language learning. In the first subsection, first we explain why we propose a new evaluation metrics in order to explore how the method results in simple, robust designs yet influences each facet of the question for lexicographic and pedagogical purposes. In the following subsections, the evaluation metrics are presented individually in two regards, for assessing the performance of collocation clusters, and for the conformity of assigned semantic labels.

5.1 Performance Evaluation for Semantic Cluster Similarity

The collection of the traditional evaluation (Salton, 1989) of clustering works best for certain type of *clustering* method but might not be well suited to evaluate our *classification* model, where we aim to facilitate collocation referencing and help learners improve their collocation production. In that case, for assessing collocation clusters, we propose a robust evaluation method by setting up the items to be evaluated as a test for semantic similarity to judge the performance of clustering results. For semantic labeling results, we developed a grading rubric with performance descriptions for the conformity of labels as a reference guide. Two human judges were asked to give performance assessment by scoring each item. The evaluation methodology is aimed at fostering the development of innovative evaluation designs as well as encouraging discussion regarding language learning by means of the proposed method.

Landauer and Dumais (1997) were first proposed using the synonym test items of the Test

of English as a Foreign Language (TOEFL) as an evaluation method for semantic similarity. Fewer fully automatic methods of a knowledge acquisition evaluation, one that does not depend on knowledge being entered by a human, have been capable of performing well on a full scale test used for measuring semantic similarity. An example provided by Landauer (1997) is shown below where "*crossroads*" is the real synonym for "*intersection*".

You will find the office at the main ***intersection***.

(a) place (b) crossroads (c) roundabout (d) building

For this experiment, we conducted the task of evaluating the semantic relatedness among collocation clusters according to the above-mentioned TOEFL benchmark to measure semantic similarity and set up target items out of our test data as sheet of clustering performance test. Our human judges performed a decision task similar to TOEFL test takers: They had to decide which one of the four alternatives was synonymous with the target word. A sample question is shown below where grouping "***sandy*"** and "***rocky*"** together with the target word "*beach*" because they belong to the same category of concept as the collocation is more appropriate than clustering "***sandy*"** and any of others together.

sandy beach

(a) long (b) rocky (c)super (4)narrow

There are 150 multiple choice questions randomly constructed to test the cluster validation, 50 questions for each 3 testing collocation types and therein 10 for each of **A**-N, **V**-N, and V-**N** testing collocation pairs. In order to judge how much degree our model ultimately has achieved in producing good clusters, two judges were asked to primarily choose the one most nearly correct answer. If the judges find one of the distracters to be also the plausible answer, giving collective answer options is allowed for our evaluation in order to test the cluster validation thoroughly from grey area among options given inadvertently. If the judges think no single correct answer is plausible enough, 0 point can be given for no satisfactory option considered. Table 4 shows the performance figures of collocation clusters generated by the two systems. As is evidence from the table, our model showed significant improvements on the precision and recall in comparison with *JustTheWord.*

Table 4.Precision and recall of our classification model and those of *JustTheWord*

Results / System	**Judge 1**		**Judge 2**		**Inter-Judge Agreement**
	Precision	**Recall**	**Precision**	**Recall**	
Ours	.79	.71	.73	.67	.82
JustTheWord	.57	.58	.57	.59	

Without doubt, subjectivity of human judgments interferes with the performance evaluation of collocation clusters, for inter-judge agreement is just above 80 %. The closer our precision (79% and 73%) is to the discrimination ratio, the more effectively that an automatic method distinguishes subjects in accordance with human judgment.

5.2 Conformity of Semantic Labels

The second evaluation task here focuses on whether the semantic labels facilitate users to

scan the entry quickly and find the desired concept of the collocations. From the experiments, we show that the present online collocation learning tools may not be an appropriate place to seek guidance on fine discrimination between near synonyms. This problem could be alleviated if the alphabetical frequency ordering of the learning tool could be supplemented by thematic treatment in our thesaurus-based semantic classification model. Our evaluation result will indicate the extent to which semantic labels are useful, to what degree of reliability. Only to the extent that evaluation scores are reliable and the test items are solidly grounded in its practical viewpoint can they be useful and fair to the assessment.

Two human informants were asked to grade collocation with label, half of them randomly selected from our output results. The assessment was obtainable through different judges that participated in evaluating all of the collocation clusters as described above. One native American graduate and a non-native PhD researcher specializing in English collocation reference tools for language learners were requested to help with the evaluation. We set up a three-point rubric score to evaluate the conformity of semantic labels. When earning two points on a three-point rubric, a label has performed well in terms of guiding a user finding a desired collocation in a collocation reference tool. If the assigned label is somewhat helpful in collocation look-up, a score of one is shown that labels are achieving at an acceptable level. To assign judgments fairly and to calculate a fair reflection of the conformity of the labels, a zero score can be given if the labels can be considerably misleading to what is more indicative of the concepts. We set up an evaluation guide to present judges with the description for each rubric point, and allow the judges to grade each question as “0”, “0.5” or “1” for the item.

Table 5 shows that 77% of the semantic labels assigned as a reference guide has been judged as adequate in terms of guiding a user finding a desired collocation in a collocation learning tool, and that our classification model provably yields productive performance of semantic labeling of collocates to be used to assist language learners. The results justify the move towards semantic classification of collocations is of probative value.

Table 5.Performance evaluation for assigning semantic labels as a reference guide

	Judge 1	**Judge 2**
Ours	.79	.75
JustTheWord	Not available	Not available

6. Conclusion

The research sought to create a thesaurus-based semantic classifier within a collocation reference tool limited to the collocates occurring without meaning access indexes. We describe a thesaurus-based semantic classification for a semantic grouping of collocates with a pivot word and the construction of a collocation thesaurus that is used by learners to enhance collocation production. The thesaurus-based semantic classification classifies objects into semantically related groups that can participate in the same semantic relation with a given word. Rather than relying on a distributional analysis, our model is resourced from an integrated semantic knowledge, which is then generalized to combat sparsity. The evaluation shows that this robustly designed classification model facilitates the existing computational collocation reference tools and provides users with the collocations they desire to make semantically valid choices. The thesaurus structure is successfully imposed to facilitate grasping concepts of collocations.

Given that there is very little precedent review for us to follow, this research offers insights into how such a collocation thesaurus could be structured and useful. The semantic labeling described here improves collocation reference tools and has given us a tool for studies of collocation acquisition. The final results convincingly motivate the move towards semantic

classification of collocations.

Many avenues exist for future research and improvement of our classification model. Another possibility would be to train more set of variables, each of which may take one among several different semantic relations for each collocation types. There is also a set of constraints which state compatibility or incompatibility of a combination of variable semantic relations.

To top it all off, existing methods for extracting the best collocation pairs from a corpus of text could be implemented. Domain knowledge, heuristics, and WSD techniques could be used to improve the identification of semantic label types. Semantic relations could be routed to classification model that performs best for more types of collocation pair (such as adverb-adjective pairs).

References

Béjoint, H. 1994. Tradition and Innovation in Modern English Dictionaries. Oxford: Clarendon Press.

Benson, M. 1985. Collocations and Idioms. In *Dictionaries, Lexicography and Language Learning*, R. Ilson (Ed.), 61-68. Oxford: Pergamon Press.

Brin, S. & Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.

Chen, P. C. 2002. A Corpus-Based Study of the Collocational Errors in the Writings of the EFL Learners in Taiwan. Unpublished master's thesis, National Taiwan Normal University, Taipei.

Chklovski, T. & Pantel, P. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of EMNLP*, 33-40.

Fellbaum, C.(Ed.) 1998. WordNet: An Electronic Lexical Database. MA: MIT Press.

Firth, J. R. 1957. The Semantics of Linguistics Science. Papers in linguistics, 1934-1951. London: Oxford University Press.

Hatzivassiloglou, V. & McKeown, K. R. 1993. Towards the Automatic Identification of Adjectival Scales: Clustering adjectives according to meaning. In *Proceedings of ACL*, 172–182.

Hindle, D. 1990. Noun Classification from Predicate-Argument Structures. In *Proceedings of ACL*,268-275.

Jian, J. Y., Chang, Y. C. & Chang, J. S. 2004. TANGO: Bilingual Collocational Concordancer. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

Kilgarriff, A. & Tugwell, D. 2001. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of ACL Workshop on Collocations*, 32-38.

Landauer, T. K. & Dumais, S. T. 1997. A Solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2):211-240.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC*, 24–26.

Lewis, M. 1997. Implementing the Lexical Approach. Hove : Language Teaching Publications.

Lewis, M. 2000. Teaching Collocation: Further Development in the Lexica1 Approach. Hove: Language Teaching Publications.

Lin, D. 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In *Proceedings of ACL*, 64-71.

Liu, L. E. 2002. A Corpus-Based Lexical Semantic Investigation of Verb-Noun Miscollocations in Taiwan Learners' English. Unpublished master's thesis, Tamkang University, Taipei.

Rundell, M. (Editor-in-Chief). 2002. The Macmillan English Dictionary for Advanced Learners. Oxford: Macmillan Publishers Limited.

Nastase, V. & Szpakowicz, S. 2003. Exploring Noun–Modifier Semantic Relations. In *Proceedings of International Workshop on Computational Semantics*, 285–301.

Nation, I. S. P. 2001. Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nirenburg, S. & Raskin, V. 1987. The Subworld Concept Lexicon and the Lexicon Management System. Computational Linguistics, 13(3/4): 276-289.

Salton, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. New York: Addison-Wesley Publishing.

Scholfield, P. 1982. Using the English Dictionary for Comprehension. TESOL Quarterly, 16(2):185-194.

Tono, Y. 1992. The Effect of Menus on EFL Learners' Look-up Processes. LEXIKOS 2: 229-253.

Tono, Y. 1997. Guide Word or Signpost? An Experimental Study on the Effect of Meaning Access Indexes in EFL Learners' Dictionaries. English Studies, 28: 55-77.

Turney, P. D. 2006. Similarity of Semantic Relations. Computational Linguistics, 32(3):379–416.

Wanner, L., Bohnet, B. and Giereth, M. 2006. What is beyond Collocations? Insights from Machine Learning Experiments. *EURALEX*.

Summers, D. (Director) 1995. Longman Dictionary of Contemporary English (3rd edition). Harlow:Longman.

Procter, P. (ed.) 1995. Cambridge International Dictionary of English. Cambridge: Cambridge University Press.

以Fujisaki模型驗證連續語流中字調及韻律詞對應於階層性韻律架構HPG的意義

鄭秋豫 蘇昭宇
中央研究院語言所語音實驗室
cytling@sinica.edu.tw, morison@gate.sinica.edu.tw

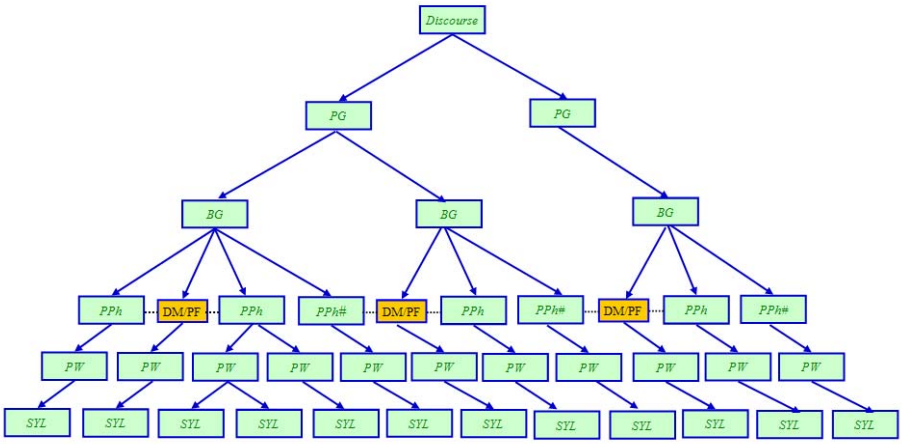
摘要

本文從台灣地區國語連續語流的字調及韻律詞基頻曲線模型，根據鄭秋豫所提出的「階層式韻律句群HPG架構」，由下層到上層，將基頻曲線模型參數與HPG韻律階層結合與驗證，探討（1.）句調成份與字調成分對應各韻律階層的貢獻度 與（2.）字調成分在各階層韻律單位管轄下，基頻模型參數如何變化。結果顯示（1.）HPG架構中各韻律層由上而下管轄制約，下層韻律單位必須承上層韻律訊息進行系統性的調整。（2.）字調接受韻律詞的上層制約，字調層及韻律詞層對基頻輸出均有貢獻，韻律詞與字調的基頻關係，不等於字調線性串接及平滑；表意語境所造成的制約，才是語流韻律的主要特徵。

關鍵詞：韻律詞，基頻曲線模型，階層式韻律句群HPG架構，句調成份，字調成分，表意韻律

一、 緒論

國語連續語流韻律一向被視爲充滿變異性且難以預測，本文主旨在討論韻律語境來自階層式管轄制約，韻律語境同時包括線性串接平滑及上層語篇的跨短語表意資訊，因此串接平滑不足以解釋語流中字調的變化。鄭秋豫由語段與語篇的角度切入，發現表面看似複雜的連續語流韻律，事實上有系統性規則可循，這些規則及階層式的關係，與串接共構表意韻律語境，並在物理信號上表現出特定基型，因此語者與聽者依據此基型產製與接收來自語篇的大範圍表意韻律訊息，結合區辨詞義的字調、區辨句法訊息的句調共同達到溝通的目的。鄭秋豫[1][2][3]於2004提出階層式語流韻律架HPG（Hierarchical Prosodic Phrase Grouping）指出，從聲學語音訊息而言，口語連續語流韻律的多短語階層架構是以感知爲基礎[2][3]，感知的最大成分是聽者預期，該架構主要精神在於將國語口語語流的韻律單位，定義爲多短語韻律短語組PG (Multi-phrase Prosodic Group）而非單一短語（phrase），表達上層語意訊息的連貫性，構成韻律短語組的相鄰及跨短語的語段韻律語境，從大範圍韻律單位表示特定語意段落的開始，延續與結束。因此此跨短語語段的韻律語境的基型，即爲語者溝通時語言即時產製與接收處理的模版，當短語形成句段時，各短語必須受上層語篇語意資訊管轄制約而調整，呈現表意韻律語境，才能成爲語意完整的語段。因此傳統語音信號分析或串接平滑皆無法解釋的句調變異，套用HPG架構後，其實可從韻律語境結構的角度得到解釋。我們先前也已提出基頻曲線、音節時長、能量分佈和停頓時長對應HPG架構的證據[1][2][3]。



圖一「階層式多短語韻律句群HPG」韻律單位架構圖。由下而上分別爲音節（SYL）、韻律詞（PW）、韻律短語（PPh）、呼吸組（BG）、韻律句組（PG）及語篇（Discourse）(Tseng 2006)。

以HPG韻律架構分析語流韻律的取向和傳統韻律分析最大的不同，在傳統韻律分析通常只分析單字字調和短語句調，卻忽略了連續語流裡，除了受詞義制約規範的字調和受句法結構制約規範的句調以外，還含有上層語篇語意造成的句調連貫訊息，表達語段訊息。亦即韻律語境構成句調以上語流全面性的連接性與連貫性，以致韻律短句間產生韻律關連性，各下級韻律單位必須依照更語篇韻律的上層訊息進行系統性的調整，提供語段從何處開始、維持到結束的韻律語境訊息；表意韻律語境不僅是相鄰韻律單位的關係，各語段下轄之各層次級韻律單位，均需依此規範做系統性的調整，同時表現相鄰單位及跨單位的韻律關係。鄭秋豫等稍早的韻律短語PPh基頻分析研究，已提出相鄰PPh及跨PPh調整的證據[4]。在HPG架構中，字調以上、韻律短語以下的韻律詞PW也是一個比詞彙詞略大的韻律單位，詞彙詞約佔韻律詞的80%, 構成韻律詞意範圍內的音節，同樣有著關聯性與連結性，也必須依據多音節韻律詞的詞意範圍對音節字調進行相鄰音節及跨音節的調整，使聽者不受音節個數限制，輕易的將韻律詞正確的歸類，並據此判斷是否爲另一個新韻律詞的開始。因此，由於HPG架構的提出，傳統語流韻律分析所視爲字調及孤立短句的種種變異性，其實皆可由階層式架構韻律短句上層的資訊加以釐清並解釋，而且可以預測。由此得知，語音信號必須兼顧HPG架構中各層的韻律效應所造成的相鄰及跨單位韻律關係，而不僅只是做字調與句調的線性串接及平滑，才能將完整地模擬出連續語流的特性。此篇論文將以單字及韻律詞基頻曲線爲物理特徵參數，由下至上提出對應HPG架構的證據。

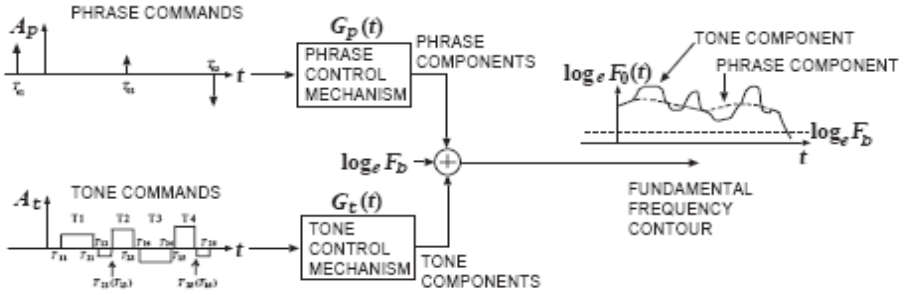
根據趙元任先生的說法[5]，短語句調與字調之間的關係，就好比波浪與連漪之間層層疊加關係，可以代數總和表示，相位相等時互相加成，相位相反時互相抵消。而在所有基頻曲線模型中最能表達這種疊加概念的，爲日本學者Fujisaki[6]於1984年提

出的Fujisaki Model。此模型的精神是非聲調語的一個句調單位IU (intonation unit)的基頻曲線，必能拆解成全面句調成分與局部強調二個成分，單位大小不同，而套用於聲調語言時，局部強調則被轉換成描述字調的成分[6, 7, 8]，此即証明且呼應了趙元任所指大波浪與小漣漪的關係。因此，綜合趙元任先生與Fujisaki教授的看法，國語短語句調的基頻曲線，其實是字調成分與句調成分疊加而成，亦即除相鄰字調的連接外，還有來自上層的句調的覆蓋，因此不僅是字調的串接。若根據HPG架構，連續語流裡還更不僅只有字調與句調兩種韻律層級與單位。我們先前依據聽感標註得出的HPG架構進行分析，已找出句調彼此間的關連性與連貫性，即爲所謂上層訊息，及對應HPG架構中的韻律句群關係[1][2][3] ，並在稍早的研究中，特別探討如何使用Fujisaki Model提取短語句調的成分在語段中的體現[9] 。同樣的，我們也希望在對應HPG架構中的韻律詞層，也找到字調間關係，並確認韻律詞在連續語流中作爲基本韻律單位的韻律意義。我們知道，音段成分相同的音節，會因爲字義不同而有不同的聲調，超音段成分基頻曲線的變化由詞義決定，因此嚴格說來，只能稱爲詞義韻律(lexical prosody)，而非語流韻律的特徵或成分。因此本論文除了證明HPG架構外，也將以韻律詞的基頻曲線變化爲主要分析參數，將強調字調以上韻律詞的在基頻曲線上的相對關係與特徵的確存在，且是連續語流中的基本韻律單位；短語也是連續語流的基本韻律單位而非終極韻律單位。

二、國語基頻曲線特徵參數自動擷取系統

(一) Fujisaki model 簡介

Fujisaki 等 1984 年[6]提出疊加式 command-response 基頻曲線模型，簡稱爲 Fujisaki model，此模型的特點在於拆解看似不規則的基頻曲線爲三個不同的元件函數（component）的疊加總合，並分別可找到相對應發聲器官的物理特性來解釋這些元件函數，此三元件函數分別爲(1.)短語元件 (Phrase Component Ap)，反應較大單位基頻曲線的控制與發聲限制；(2.)強調元件(Accent Component Aa)，反應較小單位基頻曲線的控制發聲限制；與(3.)基底頻率(base frequency Fb) 代表基本音高。原 Fujisaki model 中的 Accent component Aa 泛指強調、加重語氣對局部基頻曲線造成的影響，此模型應用到非聲調語言與英語、德語時，大單位指的是片語的語調或短語的句調，即陳述句的由高走低的下傾趨勢，小單位則用來表示局部的加重或加強（emphasis）；應用到聲調語言模擬國語時，大單位表示的成分與非聲調語相同，即片語的語調或短語的句調，而小單位被用來表示單音節的局部變化，亦即字調。本研究僅分析國語連續語流，所以 Aa 皆表示字調成分，因此下文均稱做字調元件。



圖二、短語元件、強調元件與基底頻率疊加後的基頻曲線（Fujisaki, 1984）

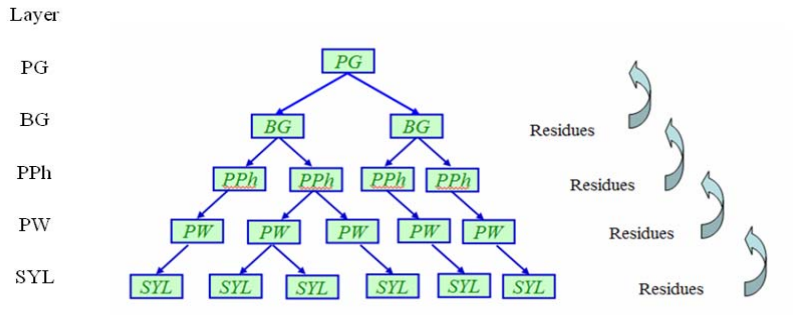
(二) Fujisaki 特徵參數的自動擷取

基於 Fujisaki Model 的國語基頻曲線，陳述句的句調可視爲開高後走低全面下傾、局部字調起伏變化與基本頻率三元件疊加所成，我們則採用了 Mixdorff 2000、2003 年[7][8]提出的方法來解析原有的基頻曲線：以一組截止頻率爲 0.5Hz 的高通濾波器(high-pass filter)來分離基頻曲線，自動提取出基頻曲線中變化劇烈的部份即爲語流的基本單位，可對應 Fujisaki model 中的字調元件；而變化和緩的部份，則爲語流中語調全面下傾的趨勢，可對應 Fujisaki model 中的短語元件。接著分別對三個元件進行逼近步驟：(1) 高通濾波器的輸出定義爲高通曲線(HFC)，爲字調元件逼近的目標曲線，(2) 扣掉高通部份剩餘平滑曲線則定義爲低通曲線(LFC)，找出此低通曲線的最低點並定義爲通過此最低點的直線爲基底直線(Fb)，(3) 扣掉基底直線後的曲線視爲短語元件的目標曲線，必須用短語元件函數來加以逼近。傳統人工逼近步驟(1)[10]中，遇到第三聲（字調三）跟第四聲（字調四）的音節，通常會指派一大一小的字調元件以逼近一音節內較複雜的基頻曲線，如圖二所示，然而爲了自動擷取大量語料的參數，在自動擷取 Fujisaki 特徵參數程式中，我們忽略了較小的字調元件，換句話說，不論字調，我們皆只採用一字調元件來逼近一音節內高頻的基頻曲線。

三、基於HPG架構之階層性多元迴歸分析

從自動擷取Fujisaki特徵參數系統中，我們可將原始的基頻曲線拆解成短語元件與字調元件，分別對應HPG架構中字調與句調的成分。然而原始的Fujisaki Model並無對應韻律詞層、呼吸句群層與韻律短語層的元件函數，因此我們利用階層性線性迴歸找出各層的貢獻度與特徵。階層性線性模型是簡單線性迴歸的衍生[11][12]，每個輸入都附有多層次的停頓標註，每一標註分別代表著在每一韻律層的參數與特性，我們利用每一層的標註資訊，可得到各層韻律單位的模型。接著依據此模型進行正規化。由於有更上層的標註，正規化後數值的變異並不被視爲實驗誤差，而是以更上

層的標註進行線性迴歸分析，分析及預測下一韻律層的輸入，因此可得到更上層的預測模型與貢獻度。 逐層分析、預測後計算出各韻律層的貢獻度。圖三以圖式表示階層式線性回歸逐層分析。



圖三、以階層式線性迴歸逐層分析示意圖 （Tseng et al, 2004）

本實驗以(1)Ap 爲分析韻律短語句調以上語段或與篇語意之特徵 (2)Aa 爲分析字調以上韻律詞意之特徵。Ap 是對應韻律短語的基頻特徵，因此分析的步驟由階層式架構中韻律短語層(PPh)開始、之後逐層向上，對上一層的呼吸句群層(BG)及更上一層韻律短語句群(PG)進行的線性迴歸，分析參數如下: (1)PPh 層: 以目前 PPh 長度(Current PPh Length)、前一 PPh 長度(Preceding PPh Length)與後一 PPh 長度(Following PPh Length)的組合做爲分析參數，進行分析，經過線性迴歸後的殘差定義爲 Delta1，並輸入 BG 層進行分析，(2)BG 層: 以目前 PPh 在 BG 中的位置(BG Sequence)做爲分析參數，若 BG Sequence=1，表示目前 PPh 爲此 BG 之起始 PPh，以此類推，進行線性迴歸，(3)PG 層: 與 BG 層輸入參數相同，其數學函數表示如下：

PPh

$Ap=f(FollowingPPh\_Length, PrecedingPPh\_Length, CurrentPPh\_Length)+Delta1$

BG

$Delta1=f(BGSequence)+Delta2$

PG

$Delta2=f(PGSequence)+Delta3$

其中 *f* 表示線性迴歸函數，迴歸係數與原始值 Ap 間的差值視爲上層貢獻 *Delta*，並以上層的輸入參數對 *Delta* 再執行一次線性迴歸，此時得到的迴歸正確率視爲上層的貢獻度，以此類推。

Aa 是對應字調的特徵，因此分析的步驟由階層式架構音節層(Syllable)開始，預測參數包括目前的字調(Current Tone)與前後字調的組合(Preceding Tone + Following Tone= Tone Context)，之後對上層的韻律詞層(PW)做分析，預測參數包括韻律詞邊界(PW Boundary Info)與此音節在韻律詞內的位置順序(PW Position Sequence)，由於在[12]中發現，在較高層級的韻律邊界常有邊界效應發生，因此我們也將邊界效應加入考慮，包括 1.韻律短語邊界訊息(PPh Boundary Info)、2. 韻律句群層邊界訊息(PG Boundary Info)，將大範圍韻律單位首和尾音節的類別標記出來，進行獨立的

Aa 類別分析，其數學函數表示如下：

Syl
Aa=f(FollowingTone, PrecedingTone, CurrentTone)+Delta1
PW
Delta1=f(PW Boundary Infor, PWSequence)+Delta2
Boundary effect above PPh
Delta2=f(PPh Boundary Infor, PG Boundary Infor)+Delta3

四、實驗語料

文本部份，採用 (1)古典文體 CL 與 (2)長篇敘事段落文本 CNA，共計(1)26 篇古典文體語篇段落（包含：4 篇古典散文，1 首賦，１首民歌，6 首古詩，6 首唐代樂府詩和 8 首宋詞），以及(2) 26 則白話敘事段落。

語料部分，由一男一女發音員（M056＆F054）朗讀古典文體文本，平均語速分別爲 202m/syl 和 265ms/syl；另外的一男一女發音員則（M051＆F051），負責朗讀白話敘事段落，平均語速分別爲 189m/syl 和 199ms/syl，錄製過程使用 Sony ECM-77B 迷你麥克風、以及 Cool Edit 2000 在隔音室進行錄音。表一統計兩種韻律格式加總後的 HPG 架構下韻律邊界以及相對應韻律單位的個數。

表一、古典文體 CL 與長篇敘事段落文本 CNA 韻律邊界以及相對應韻律單位個數

語料	語者	SYL/B1	PW/B2	PPh/B3	BG/B4	PG/B5
CL	F054	1444	599	290	135	58
	M056	1551	619	318	142	47
CNA	F051	6583	3468	1092	297	151
	M051	6661	3332	1207	270	129

五、實驗結果與分析

(一)句調成份與字調成分對應各韻律階層的貢獻度

依據第三節的階層性多元迴歸分析方法，我們由下層到上層進行線性迴歸，每一層的迴歸正確率視爲每一韻律層的分層貢獻度，而由最下層(音節層)累積到當層的迴歸正確率即被視爲累積到當層的累積貢獻度，累積正確率１００％表示最後所有韻

律層的迴歸係數的總和等於原始Fujisaki特徵參數值，Fujisaki特徵參數值完全可由階層性多元迴歸分析正確預測。

1. 字調元件Aa分層貢獻度

從階層性多元迴歸分析可求出字調元件Aa在音節層與韻律詞層的貢獻度，並加入韻律短語層以上邊界效應的考慮[13]，包括此音節前後是否有韻律短語邊界及韻律句群層邊界。最後結果顯示Aa的正確率可達73.80%到56.25%不等，其中大部份貢獻度來自音節層與韻律詞層，邊界效應的貢獻度則介於5~7%。

表二、字調元件在音節層與韻律詞層的累積正確率

語料	語者	Syl 層貢獻度		PW 層貢獻度	
		Tone	Tone Context	PW Boundary Info	PW Position Sequence
CL	F054	46.21%	54.74%	60.54%	66.61%
	M056	39.12%	47.86%	57.68%	61.45%
CNA	F051	38.40%	45.00%	48.43%	51.27%
	M051	41.61%	47.96%	51.33%	54.53%

表三、加上邊界效應後字調元件的累積正確率

語料	語者	PPh 層以上的邊界效應		邊界效應的貢獻度
		PPh Boundary Info	PG Bounary Info	
CL	F054	72.98%	73.80%	7.19%
	M056	64.13%	66.89%	5.43%
CNA	F051	54.41%	56.25%	4.98%
	M051	57.43%	59.32%	4.79%

2. 短語元件Ap分層貢獻度

從階層性多元迴歸分析可求出短語元件Ap在韻律短語層、呼吸句群層與韻律句群層的累積貢獻度，我們發現古文語料CL有許多貢獻度來自於韻律短語層以上的上層資訊，反之，白話長篇敘事語料CNA正確率則多來自目前PPh與前後PPh的長度資訊。

表四、短語元件Ap在韻律短語層、呼吸句群層與韻律詞層的累積正確率

語料	語者	PPh	BG	PG
CL	f054	58.79%	63.58%	76.66%
	m056	37.89%	48.99%	73.66%
CNA	F051	80.17%	81.46%	87.71%
	m051	81.53%	82.72%	88.20%

如前文所提，基頻曲線的變化由Ap與Aa構成，因此我們將Ap與Aa的預測率正確率平均作爲套用HPG架構後Fujisaki model對總體基頻曲線模型預測正確率。結果如表五。

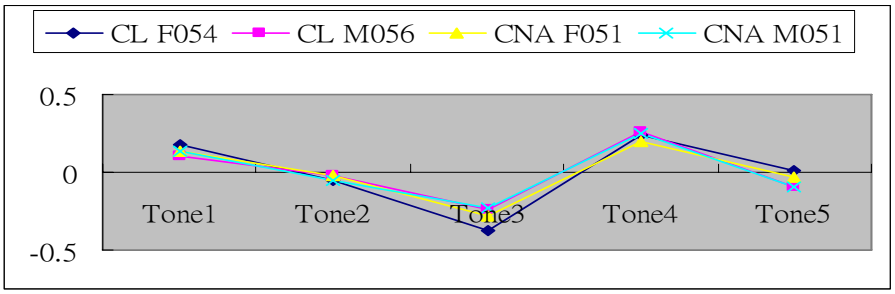
表五、套用HPG架構對總体基頻曲線模型預測正確率

語料	語者	Aa	Ap	Total
CL	f054	76.66%	73.80%	75.23%
	m056	73.66%	66.89%	70.28%
CNA	F051	87.71%	56.25%	71.98%
	m051	88.20%	59.32%	73.76%

(二) Aa對應音節層與韻律詞層模型

1. 字調(音節層)模型

在上表中我們得見字調對Aa的貢獻度爲最大，此結果與先前研究Fujisaki Model的結果一致。相對於每個字調的Aa模型列於下圖，我們可看見Aa的能量模型在不同語料間相當一致。



圖四、自動擷取出的字調元件對應各個字調之模型

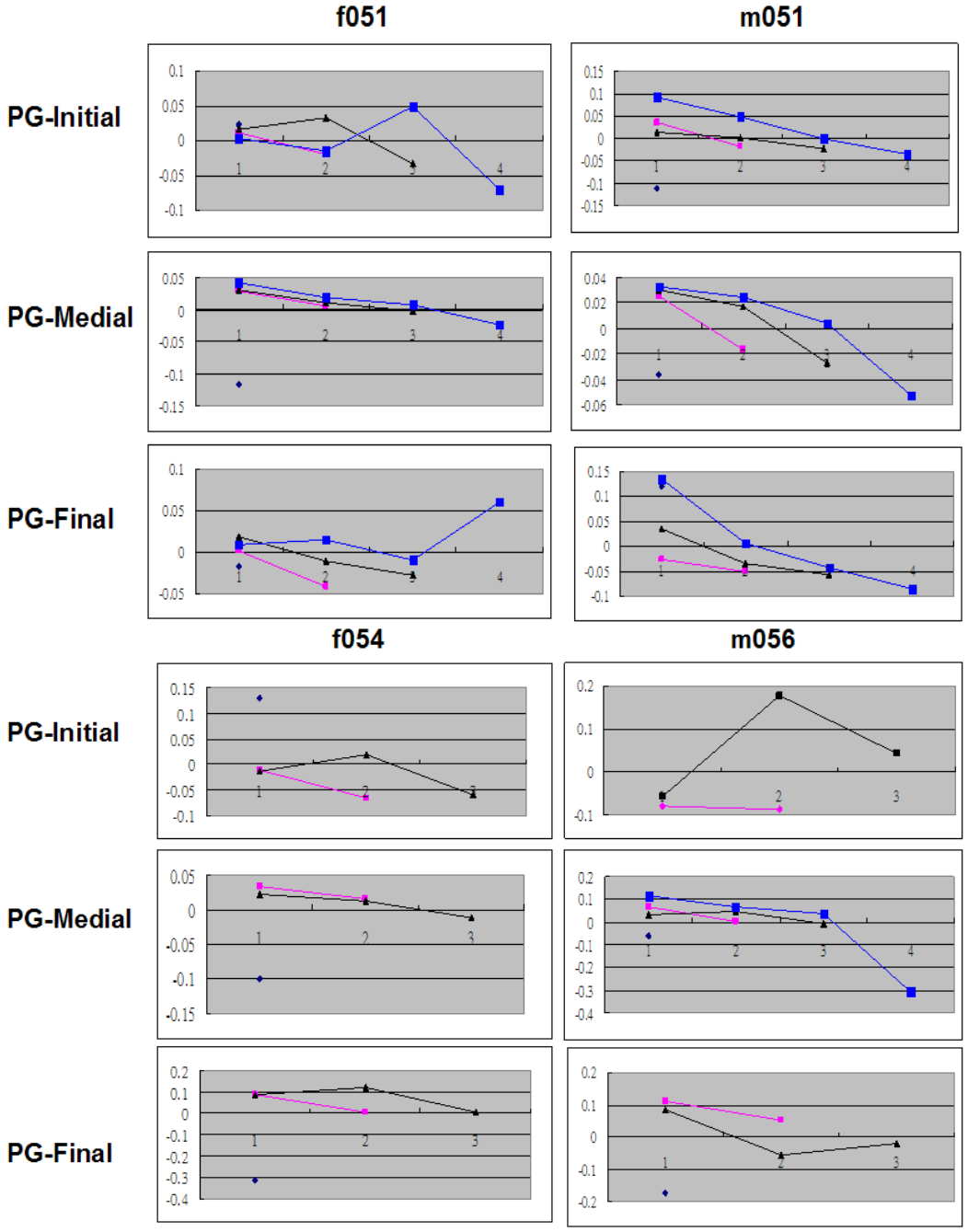
2. 韻律詞模型

2.1 韻律詞受韻律句群PG管轄之特徵

將消除字調效應後的韻律詞基頻模型依照PG位置分類後，我們發現在PG-Medial的韻律詞模型，在四筆語料間呈現一致性的特徵，而且韻律詞在基頻曲線上的邊界分隔，主要發生在韻律詞的詞尾，換言之，即便韻律詞間並沒有停頓，且韻律詞間受字調間平滑影響，並不易觀察到基頻重設，但經過消除字調效應的正規化處理後，可發現PW末的Aa強度不如其他PW音節，因此我們可從這詞尾的基頻能量衰減(decay)來辨別韻律詞單位邊界。在PG-Initial位置，我們也發現韻律詞在PW末存在相

同特徵，且PW末和倒數第二音節間相對於PG-Medial的對比也較大。

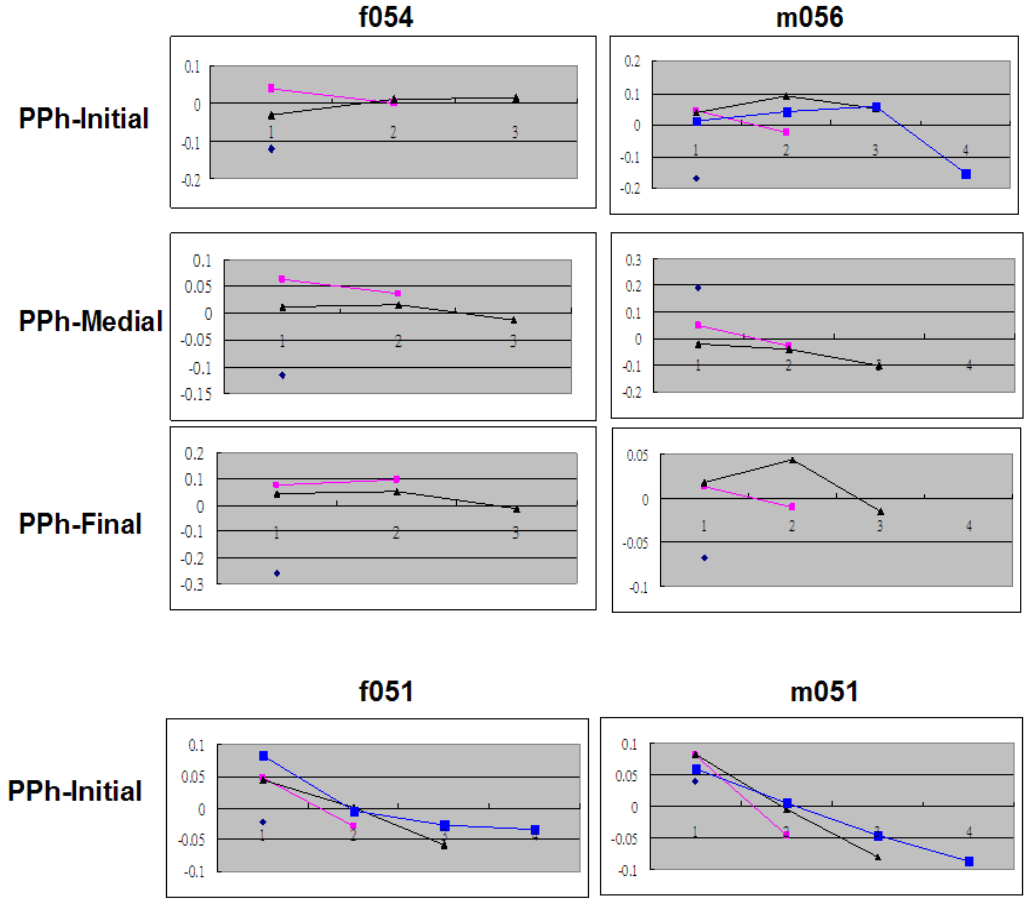
對照2007以Ap爲實驗，獲不同語體的階層貢獻度不同之結論[9]，經由本實驗結果可進一步發現，由於跨語體間的韻律詞特徵，在PG-Medial最爲穩定且最不明顯，因此大部分與語體變化相關的的貢獻度差異，應來自PG-Initial與PG-Final而非PG-Medial。

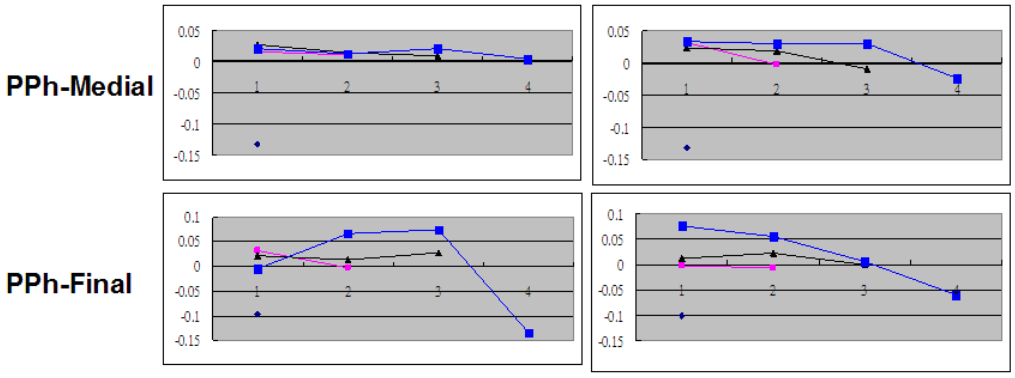


圖五、扣除字調成份後，韻律詞模型在不同 PG 位置的特徵，每條曲線表示特定長度的韻律詞模型，橫軸表示在此韻律詞內音節順序，縱軸表示扣除字調成份後的殘差。

2.2 韻律詞PW受韻律短句PPh管轄之特徵

同理我們將扣除字調成份的韻律詞PW模型依照韻律短句PPh位置分類後，得到各韻律詞的特徵，發現韻律詞詞尾的特徵與詞首、詞中最爲不同，尤其又以PPh-Medial最爲明顯，PPh–Initial、PPh-Medial最重要的特徵也大致發生在詞尾Aa。在PPh-Medial的位置的韻律詞的一致性表示，PPh-Medial爲最不受語體影響的相對位置，如圖六所示。我們也可發現在PPh-Final，Aa強度韻律詞的最末音節存在衰減的特徵，雖然PW模型不如PPh-Medial規則，但最末音節的Aa衰減程度較PPh-Medial明顯，顯示PPh在基頻區線上的邊界效應，主要發生在邊界前的韻律詞。這個特性，與音節時長在PPh句尾延長效應吻合[2] [12]，顯示在物理信號上，主要的PPh邊界效應，主要表現在PPh邊界前，PPh句末的物理信號上。





圖六、扣除字調成份後，韻律詞模型在不同 PPh 位置的特徵，每條曲線表示特定長度的韻律詞模型，橫軸表示在此韻律詞內音節順序，縱軸表扣除字調成份後的殘差。

六、結論與展望

由 Aa 實驗結果可知，字調仍爲影響音節內基頻曲線的主要因素，大約佔字調元件 Aa 預測正確率的 40~45%；韻律詞 PW 的階層性貢獻佔 15~20%。但以上數據亦顯示：從字調辨識的角度而言，字調的正確率不及一半，表示最終的語流韻律輸出中，字調的成分並非字字可辨。由扣除字調成份的韻律詞 PW 模型可發現，HPG 架構中的 PW 層存在一定特徵，也有一定的貢獻度，解釋了在基頻曲線中，並非只以字調爲基本單位進行串接；PW 對基頻曲線也有一定程度的影響，並在最終的語流韻律輸出中，佔有一定的比例。以上結果和稍早 Ap 實驗的 PPh 表現的結果相符，即 PPh 的 Ap 必須考慮呼吸句群 BG 層以及韻律句群 PG 層的上層效應[9]，本研究結果顯示 BG 層與 PG 層對句調元件 Ap 的貢獻度約佔 7~35%，語流最終的韻律輸出，各韻律層都有貢獻。不同語體的朗讀語料，因語體而產生的韻律輸出差異，只是階層式貢獻度的分佈差異而已，其韻律成份完全可由 HPG 架構解釋，同一基型只需調整階層式的韻律貢獻度，便可產生不同的韻律輸出[9]。結合以上實驗結果，我們因而得知，不論字調或句調，皆爲 HPG 的次級韻律單位，各自受到 HPG 架構的層層管轄，系統性的調整字調以及句調，以形成流暢的連續語流中的表意韻律語境，表達句調間的連續與跨句調的呼應。我們相信這些韻律語境的模型爲理解與產製語音的重要單位，口語產製時，人們使用這些韻律基型，因此依據這些基型，可輕易快速的將複雜的連續語流的語境歸類成最適合的語段、語篇，以進行上層語意的組織與分析。

我們期望這些證據及特徵能提供語音合成系統一個新的思維: 語流韻律是有架構的，因有系統性而有跡可循，所以連續語流並非充滿變異而難以預測。語流的流暢並非僅來自於字調與句調的線性串接的完美平滑，也必須同時包含多短語語段間的跨短語表意韻律語境,。我們所提出的多短語階層性 HPG 架構，完整的解釋語流韻律規範的來源及如何互動。合成的語音輸出聽起來不流暢、不自然的感覺，則大部

分由於未含上層資訊，以致韻律語境不足或未能體現韻律語境。同理，在聽者即時處理語流時，即便每一個單音節的字調訊息不完全，只要有相當成分的韻律語境，便能彌補並提供聽者切分韻律單位及向前預估的訊息。只不過，所謂的相當成分，必須同時含有相鄰下層韻律單位的連接及上層韻律單位的呼應。反之，如一段連續語流內的單位無法同時體現階層性 HPG 架構組織所規範之全面表意及局部連接的韻律語境時，將會違背聽者預期，造成聽者切分單位的錯誤，需一再修正，而延誤即時處理，由此亦可見，語音識別並不等同字調識別。未來研究方向包括將 HPG 架構套用至授課等自發性語料溝通意圖明顯、與篇及語段訊息分明的語料進行分析，一方面希望找出此規劃的基型也存在在看似不規則的自發性語料中，一方面更全面的解釋語流中字調的變異，解構字調即韻律的迷思。

參考文獻

[1] Tseng, C. "Prosody Analysis", *Advances in Chinese Spoken Language Processing*, World Scientific Publishing, Singapore, pp. 57-76, 2006.

[2] Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, Y. "Fluent Speech Prosody: Framework and Modeling", *Speech Communication, Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation*, Vol. 46:3-4, pp. 284-309, 2005.

[3] Tseng, C. and Lee, Y. “Speech rate and prosody units: Evidence of interaction from Mandarin Chinese”, *Proceedings of the International Conference on Speech Prosody 2004*, pp. 251-254, 2004.

[4]Tseng, C., Pin, S., Lee, Y., "Speech prosody: Issues, approaches and implications". *From Traditional Phonology to Modern Speech Processing*. ed. by Fant, G., Fujisaki, H., Cao, J., and Y. Xu, 417-437. Beijing: Foreign Language Teaching and Research Press, 2004

[5] Chao, Y., *A Grammar of Spoken Chinese*. University of California Press, Berkeley, 1968

[6] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J.Acoust, J.Acoust. Soc*. Jpn.(E), 1984; 5(4), pp. 233-242, 1984.

[7] Mixdorff, H. "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", *Proceedings of ICASSP 2000*, vol. 3, pp.1281-1284, 2000.

[8] Mixdorff, H., Hu, Y. and Chen, G. "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", *Proceedings of Eurospeech* 2003, pp. 873-876, 2003.

[9] 鄭秋豫、蘇昭宇. “從不同韻律格式驗證階層式韻律架構並兼論對語音科技的應用” 第十九屆自然語言與語音處理研討會，中華民國計算語言學會，台北，2007

[10] Wang, C., Fujisaki, H., Ohno, S. and Kodama, Tomohiro. "Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model", *Proceedings of EUROSPEECH'99*, pp. 1655-1658,

1999.

[11] Keller, E., and Zellner, K.,. “A Timing model for Fast French”, *York Papers in Linguistics*, *17*, University of York, pp.53-75, 1996.

[12] Zellner, K., and Keller, E., “Representing Speech Rhythm” *Improvements in Speech Synthesis*. Chichester: John Wiley, pp. 154-164, 2001.

[13] Tseng, C., Su, Z., “Boundary and Lengthening—On Relative Phonetic Information.”

The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers, Beijing, China., 2008

基於 ANN 之頻譜演進模型及其於國語語音合成之應用

An ANN based Spectrum-progression Model and Its Application to Mandarin Speech Synthesis

古鴻炎　　吳昌益
Hung-Yan Gu　and　Chang-Yi Wu

國立台灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
e-mail: guhy@mail.ntust.edu.tw

摘要

考量合成語音的流暢性不佳的問題，本文提出以動態時間校正(DTW)來匹配目標(句子發音)音節與參考(單獨發音)音節之間的頻演(頻譜演進)路徑，再將頻演路徑轉換成固定維度的頻演參數，用以去訓練頻演參數類神經網路(ANN)模型。之後，將文句分析、頻演參數、韻律參數、和信號合成模組的程式作整合，而成爲可實際運轉的系統。當把此系統合成出的語音，拿去作聽測評估，所得到的平均分數顯示，頻演參數 ANN 模型的確可明顯地改進合成語音的流暢性。

關鍵詞: 頻譜演進, 流暢性, ANN, DTW, 語音合成
Keywords: spectrum progression, fluency, ANN, DTW, speech synthesis

一、前言

由前人的研究成果可知，要合成出自然、流暢的國語語音，韻律(prosody)參數的塑模(modeling)及數值產生扮演重要的角色[1,2,3]。一般被歸屬爲韻律參數的語音特性，包括：音節的基週軌跡(pitch-contour)、時長(duration)、音強(amplitude)、及音節前停頓(pause)等。我們依據過去的研究經驗發現，當採取 model based 的研究方向時，也就是韻律參數產生和信號波形合成分開處理的作法，就算是我們的韻律模型已經可以產生出相當自然的韻律參數值，但是合成出的語音信號，聽起來就是不像人講的那麼順暢。所以會這樣地具有不錯的自然度(naturalness)而欠缺流暢度(fluency)，我們先前檢討時，認爲是因爲相鄰的合成單元(音節)串接時，邊界上的共振峰軌跡(formant trace)沒有平順轉移所造成，因此我們便研究了一種解決共振峰軌跡平順轉移問題的作法[4]。使用此作法後，由聆聽合成的語音發現，流暢性是可以獲得一些改進，但是距離人講話的流暢性，仍然存在著明顯的差距。

最近回顧一些文獻後發現，我們所關心的流暢性不足的問題，其實已經有其他研究者注意到了[5,6,7]，他們提出的一種作法是，以 HMM(hidden Markov Model)模型的數個狀態，來切割一個音節的時長成爲數個時間片斷，再分別去掌握各片段上的頻譜特性(例如頻譜包絡, spectrum envelope, 的形狀)，並且以特定的狀態駐留(state staying)機率分佈來掌握在各個狀態上所應停留的時間長度。這樣的作法，以我們的觀點來看，就是在於作更細緻的規劃，把一個音節的時長以某一種非均勻(或非線性)的方法作切割，而讓不同的狀態分配到不等的時間長度，造成不同的頻譜包絡形狀會佔據不同長度的時長，以便更細緻地模仿真人發音(articulation)時的頻譜隨著時間變化的關係。

前述頻譜(包絡形狀)隨著時間演變的關係，在本文裡簡稱之爲頻譜演進(spectrum progression)，而頻譜演進路徑(簡稱爲頻演路徑)指的是，當把欲合成的音節放在橫軸上，而把相同拼音的原始錄音音節放在縱軸上，此時橫軸上各時間點所應對應的縱軸時間點，需要一條曲線來描述此對映(mapping)關係，一個例子如圖 1 所示，這樣的對映曲線就是本文所謂的頻演路徑。過去很多的國語語音合成系統，

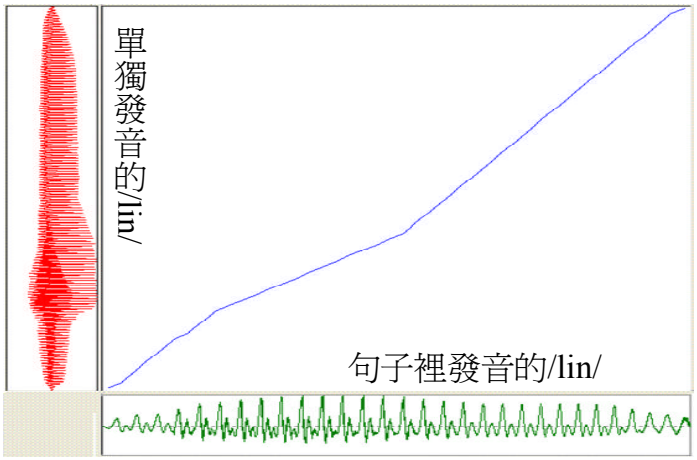


圖 1 頻譜對映曲線之例子

其合成出的語音的流暢性不佳的一個主要原因，我們認爲是因爲它們直接把頻演路徑設定爲直線，而沒有特別考慮頻演路徑的塑模(modeling)，再據以產生出逼近真人講話方式的頻演路徑。因此，我們便開始研究頻演路徑塑模及產生的問題，在此我們不追隨前人採取 HMM 來建立頻演路徑的模型，原因是 HMM 未去掌握時間上相鄰的觀測(observation)向量之間的依存(dependency)性，這相當於假設時刻 t 的觀測向量 O_t 和 O_{t+1} (或 O_{t-1})之間沒有依存關係，而只有去掌握 O_t 和它所停留的狀態之間的關連性，這樣的 modeling 方式令我們懷疑其是否可以滿足語音合成上的需求；此外，一個合成音節的頻演路徑並不會是只有固定的一條而已，而是會隨著左右鄰接

音節的不同，去行走不同的路徑(也就是 context dependent)，在此情形下，一個 HMM 的各個狀態如果只是各自去考慮 state duration 的機率分佈，而沒有考慮鄰接狀態和鄰接音節之間的相關性，則不免讓我們懷疑其完善性。

基於前述的考量，我們逐決定以 ANN (artificial neural network)來建立頻演路徑的模型，而模型的訓練步驟是: (a)逐一將整句發音裡的音節信號放在橫軸，而把相同拼音的單獨發音音節信號放在縱軸，再以 DTW(dynamic time warping)來匹配出一條頻演路徑；(b)將橫、縱軸上的音節信號的時間範圍各自正規化成 0 至 1 之間，然後在橫軸音節上均勻放 32 個正規化的時間點，各點再依頻演路徑對映至縱軸而得到介於 0~1 之間且隨著橫軸作非線性漸增的 32 的數值；(c)將各個句子發音裡的音節對映出的 32 個正規化的時間值(在本文裡稱爲頻演參數)作爲 ANN 模型學習的目標，並且把該音節及其前、後鄰接音節的資訊(也就是語境資料)作爲 ANN 的輸入資料，去訓練頻演參數的 ANN 模型。

得到頻演路徑參數的 ANN 模型後，就可將此模型和韻律模型、文句分析模組、及信號合成模組整合成一個文句翻語音系統，其結構如圖 2 所示。由圖 2 可知我們採取的是 model based 而非 corpus based 的研究方向，並且本文的焦點是在頻演參數模型的建構，而圖 2 裡其它的模組都是直接使用先前研究的成果[8,9]。我們傾向於不採取 corpus based 的研究方向，其考量是，corpus 的錄音、整理(標音、切音)需花費很大的人力、金錢(如購買 corpus)，並且 corpus 若不夠大，依然會發生音節之間基週軌跡銜接得不順暢，並且一個句子裡的音節時長會有太長及太短者，而造成發音速度忽快忽慢的不流暢情形。

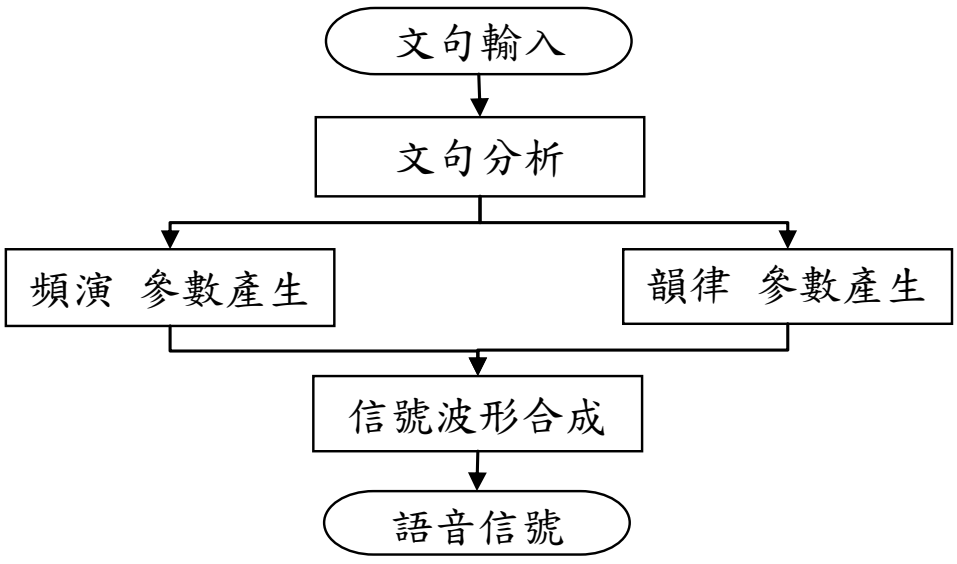


圖 2 整合頻譜演進之文句翻語音系統

二、頻演模型建造

2.1 訓練語料

我們所使用的訓練語料是由一位女性發音，先錄了 409 個單獨發音的國語基本音節，另外再錄 375 個句子的發音，總共 2,926 個音節，取樣率是 22,050 赫茲(Hz)。接著以音訊處理軟體 Wavesurfer 對語音檔進行標音(labeling)，在時間軸上標示出各個音節的音標、聲調和邊界點。標音後，我們寫了一個程式依據標籤檔來取出各音節的資訊，並將句子發音裡的各個音節切割成爲分別的音節檔案。

2.2 譜演路徑求取

動態時間校正(DTW) 是一種傳統的語音辨識方法[10]，尤其是在語者相關的語音辨識方面。DTW 的功用就是，它可以快速地找出參考音和測試音之間的一條具有最短距離的匹配路徑，當使用以頻譜差異爲依據的距離量測時，就可用 DTW 來找出頻譜上最匹配的路徑。如果我們將前述的測試音換成句子發音裡的音節，而把參考音換成單獨發音的相同拼音音節，則用 DTW 找出的頻譜上最匹配的路徑，就是本文所謂的譜演路徑。

令 $X=X_1、X_2、\ldots、X_n$ 表示，測試音(句子裡的音節)切割成音框後再求取特徵向量而得到的特徵向量序列，而 $Y=Y_1、Y_2、\ldots、Y_m$ 表示，參考音(單獨發音音節)切割成音框後再求取特徵向量而得到的特徵向量序列，在此使用的特徵向量包含 13 維度的 MFCC 係數和 13 維度的相鄰音框係數差值[10, 11]。在使用 DTW 來對 X、Y 兩序列作頻譜匹配之前，必須先選擇適當的局部路徑限制(local constrain)，過去被提出使用的局部路徑限制至少包括如圖 3 所示的三種[10]，其中 α 和 β 限制並不適合用於作頻演路徑的 DTW 匹配，原因是它們允許行走水平方向，這會使得放在橫軸的句子發音音節，要依據頻演路徑對映至縱軸放的單獨發音音節時發生混淆，也就是發生多對一(多個橫軸時間點對映同一個縱軸時間點)的情況。因此，我們選擇 γ 局部限制，以避免發生混淆的情況。

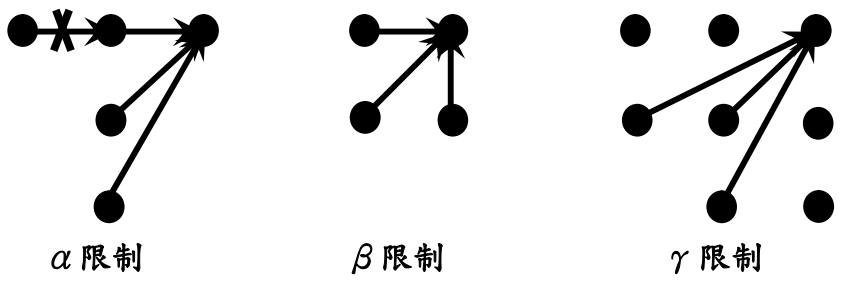


圖 3 DTW 之局部路徑限制

當採取圖 3 的 γ 局部限制時，DTW 的累積距離 Da(X,Y)的遞迴計算方式，就如公式(1)，

$$D_a\left(X_i,Y_j\right)=\min\begin{Bmatrix}D_a\left(X_{i-1},Y_{j-2}\right)+3\cdot D\left(X_i,Y_j\right)\\D_a\left(X_{i-1},Y_{j-1}\right)+2\cdot D\left(X_i,Y_j\right)\\D_a\left(X_{i-2},Y_{j-1}\right)+3\cdot D\left(X_i,Y_j\right)\end{Bmatrix} \quad (1)$$

其中，$D(X_i,Y_j)$表示以幾何距離量測特徵向量 X_i 和 Y_j 的距離，常數 3 和 2 則是我們爲了消除路徑偏好而設定的局部路徑權重值。

實際製作 DTW 程式後，進行初步測試時我們發現，如果作頻譜匹配的音節是含有無聲(unvoiced)聲母的(如/s,h,p/)，則時常會發生一個現象，就是某一軸(橫、縱軸)的聲母結尾部分會對映到另一軸的韻母起始部分，也就是匹配出的路徑不會如預期的聲、韻母的邊界剛好相互對應。因此，對於以無聲聲母開頭的音節，我們先作基週偵測[12]以找出無、有聲之邊界點，不過基週偵測也不保證會 100%正確，所以所發展的程式介面上，允許使用者去作邊界點的調整，有了邊界點之後，再將音節分割成兩段，分別去作 DTW 頻譜匹配。一個例子如圖 4 所示，橫軸放的是目標(句子發音)音節/siang/的波形，縱軸放的是參考(單獨發音)音節/siang/的波形，橫軸格線表示目標音節的音框，縱軸格線則表示參考音節的音框。進行 DTW 頻譜匹配時，分別對區域 A 作無週期性聲母的頻譜匹配，及對區域 B 作週期性部分的頻譜匹配。

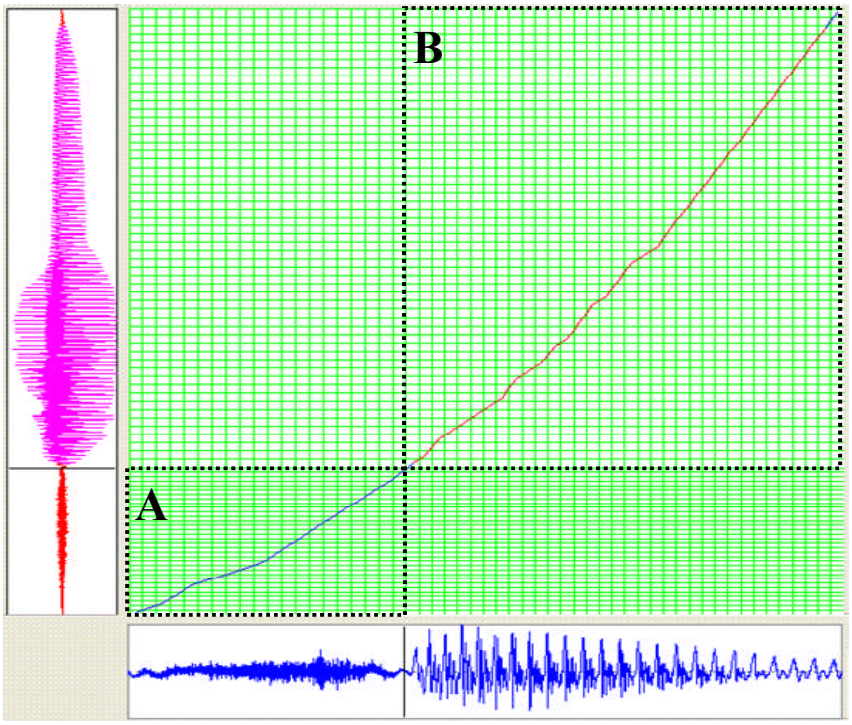


圖 4 兩段式 DTW 頻譜匹配

在音框位移(shift)和音框數量方面，原先內定的音框大小是 20ms，而音框位移是 5ms。不過我們採用的局部路徑限制的先天限制，必須將兩個音節的音框數量比例限制在 0.5 ~ 2 之間，以確保能夠滿足起點對起點、終點對終點的要求。因此當參考音音框數與目標音音框數比例超過 1.5 倍之門檻時，我們就將音框數較多的音節的音框位移作調整，也就是把音框位移乘上一個倍率，使導得出的音框數落在限制的範圍內，作法如公式(2)，

$$F_a = S_a \times \frac{2 \cdot N_a}{3 \cdot N_b} \tag{2}$$

公式(2)裡，Na 爲音框數較多之音節的原始音框數，Nb 爲音框數相對較少音節的原始音框數，Sa 爲音框數較多音節的原始音框位移，Fa 則爲調整後的音框位移。

2.3 ANN 頻演參數模型

本文採取的類神經網路結構如圖 5 所示，輸入層用以輸入 8 種語境資料，使用一層的隱藏層和一層的遞迴隱藏層，輸出層則有 32 個節點，用以輸出 32 個頻演參數。關於 ANN 權重值的訓練，採用的是最陡坡降學習法，此外遞迴隱藏層的權重值也是經由學習來決定，使用的是遞迴學習演算法。

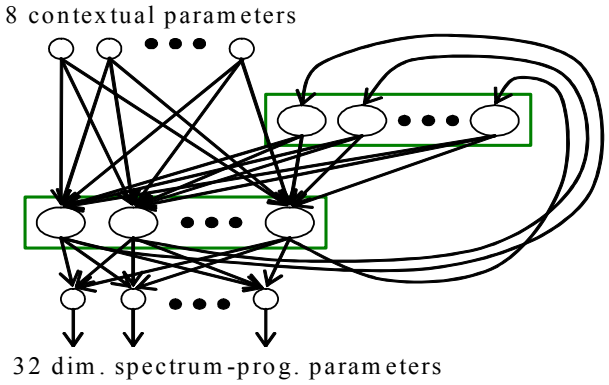


圖 5 頻演參數 ANN 之結構

輸入給 ANN 的語境資料，本研究使用“音節”作爲分析單位，由於語音是時序性資訊的傳遞，所以除了本音節的聲調種類、聲、韻母類別以外，也要考量到前一個音節的聲調和韻母類別，以及後一個音節的聲調和聲母類別。此外，考量到音節在句中的位置（例如句首、句中以及句末）也會對音節的韻律狀態產生影響，因此我們也使用一個句子時間比例之數值，來代表本音節在整句話中的時間位置。本研究所用到的 8 種語境資料，共需要以 27 個 bits 及一個浮點數來表示，詳細的配置情形如表 1 所列。

表 1 ANN 輸入之語境資料表示

項目	前音節聲調	前音節韻母類別	本音節聲調	本音節聲母	本音節韻母	句中位置	後音節聲調	後音節聲母類別
bits 數	3	4	3	5	6	浮點數	3	3

由於一個國語音節有 5 種聲調，因此聲調都以 3bits 表示。對於本音節的聲、韻母，由於國語有 22 種聲母和 39 種韻母，因此分別以 5bits 和 6bits 來表示。在前音節的韻母與後音節的聲母方面，我們考量到所準備的訓練語料較少，可能會因分類太多而造成 ANN 模型訓練語料嚴重不足，因此我們根據音節發音上的特性，將國語音節的聲母粗分爲 6 類，而韻母則粗分爲 9 類，詳細的分類方式如表 2 和表 3。依據粗分類類數 6 及 9，所以在表 1 中的前音節韻母和後音節聲母，分別使用 4 和 3bits。

表 2 國語聲母之粗分類

類別	聲母	類別	聲母
1	空聲母、ㄇ、ㄋ、ㄌ、ㄖ	4	ㄐ、ㄓ、ㄗ
2	ㄒ、ㄕ、ㄏ、ㄈ、ㄙ	5	ㄆ、ㄊ、ㄎ
3	ㄅ、ㄉ、ㄍ	6	ㄑ、ㄔ、ㄘ

表 3 國語韻母之粗分類

類別	韻母	類別	韻母
1	空韻母	6	ㄞ、ㄧㄞ、ㄨㄞ、ㄧ、ㄟ、ㄨㄟ
2	ㄚ、ㄧㄚ、ㄨㄚ	7	ㄠ、ㄧㄠ、ㄨ、ㄩ、ㄡ、ㄧㄡ
3	ㄛ、ㄧㄛ、ㄨㄛ	8	ㄢ、ㄧㄢ、ㄨㄢ、ㄩㄢ、ㄣ、ㄧㄣ、ㄨㄣ、ㄩㄣ
4	ㄜ、ㄦ	9	ㄤ、ㄧㄤ、ㄨㄤ、ㄥ、ㄧㄥ、ㄨㄥ、ㄩㄥ
5	ㄝ、ㄧㄝ、ㄩㄝ		

關於隱藏層節點數的設定，我們分別實驗了 14, 16, 18, 20 等四種數值，ANN 模型訓練的誤差值，量測後的結果如表 4 所示，其中 RMS 誤差表示 2,926 個音節的均方根誤差值的平均值，STD 誤差表示音節均方根誤差值的標準差，而 MAX 誤差表示最大的音節均方根誤差值。依據表 4 裡的誤差數值，我們最後選擇設定隱藏層的節點數爲 16。

表 4 不同節點數之 ANN 訓練誤差值

隱藏層 節點數	RMS 誤差	STD 誤差	MAX 誤差
14	0.05063	0.02443	0.18564
16	0.04906	0.02330	0.18038
18	0.04891	0.02343	0.18227
20	0.04946	0.02405	0.20055

三、系統製作

依據圖 2 所示之系統結構圖，當輸入一個中文文句後，首先會進行文句分析的處理，它經由長詞優先之查詞典過程，把文句中各個字的國語音節拼音查出，再去作三聲變調和”一”、”不”變調的處理。得到各個字的音節拼音、聲調、聲韻母、及詞邊界等資料後，接著就進行頻演參數的產生，及分別進行各項韻律參數數值的產生。由於頻演參數之 ANN 模型，已在第二節裡說明其建造過程，所以在合成階段就可以分別送各個字的語境資料給 ANN 模型，來得到各個字的頻演參數。至於韻律參數的產生和信號波形的合成，則分別在 3.1 和 3.2 節作說明。

在此值得一提的是，依據頻演模型產生出的頻演參數，我們可用以估計一個以無聲聲母(如/s/)開頭的欲合成的音節內，該無聲聲母所應分配的音節時長之比例。方法是，先將頻演參數作片段線性(piece-wise linear)內差，來形成如圖 4 裡所示的對映函數，然後以縱軸所放的參考音節的聲、韻母邊界點作爲參考點，再經由對映函數來找出橫軸上的對應點，而此點之正規化時間值，就是聲母的時長分配之比例。

3.1 韻律參數產生

韻律參數中的時長、音強、和基週軌跡等參數，我們同樣是使用如圖 5 所示的 ANN 模型結構，來對這三項韻律參數分別作訓練，也就是這三項韻律參數各自有一個獨立的 ANN 模型。我們所以必須對各項韻律參數和頻演參數分開作訓練，主要是因爲訓練語料的數量不夠多，如第 2.1 節裡所說的只有 375 句的 2,926 個音節。雖然訓練語料不是很足夠，但是經由適當地對語境資料作分類，如表 2 和表 3 之粗分類，所訓練出的模型仍可表現出不錯的效能，這可由實際的聽覺試聽來作驗證，我們爲此建立的一個網頁在 http://guhy.csie.ntust.edu.tw/spmdtw/。

另外可以一提的是，在本研究裡我們只使用 2.1 節所說的語料，去訓練時長和

音強的模型，而基週軌跡的 ANN 模型，則是直接使用先前研究所建造的模型[8]，也就是基週軌跡模型的建造，使用的是另一位男性所錄製的語料[13]，而語料數量是一樣的。所以會使用不同人的語料來訓練不同的韻律參數，其原因是我們想嘗試，當結合不同人的說話方式時，合成出的語音會有什麼不一樣的地方? 不過，作過實驗後並沒有感覺到什麼特殊的地方，初步的判斷是，各項韻律模型應可以使用不同來源的語料來作獨立的訓練。

3.2 信號波形合成

關於信號波形的合成，我們選擇採取 HNM (harmonic-plus-noise model)爲基礎的合成方法[14]，其原因是，一般熟知的 PSOLA 合成法，它所合成出的語音信號品質並不穩定，尤其是當音高(pitch)或時長(duration)作較大幅度的改變時。這裡所指的信號品質是，愈少迴音(reverberant)和愈清晰，則品質愈好。爲了能夠在大幅度改變韻律特性的情況下，仍然能夠合成出高信號品質的語音，我們在二年前就開始研究以 HNM 爲基礎的國語語音合成方法[15]，並且對它作了一些改進[9]。

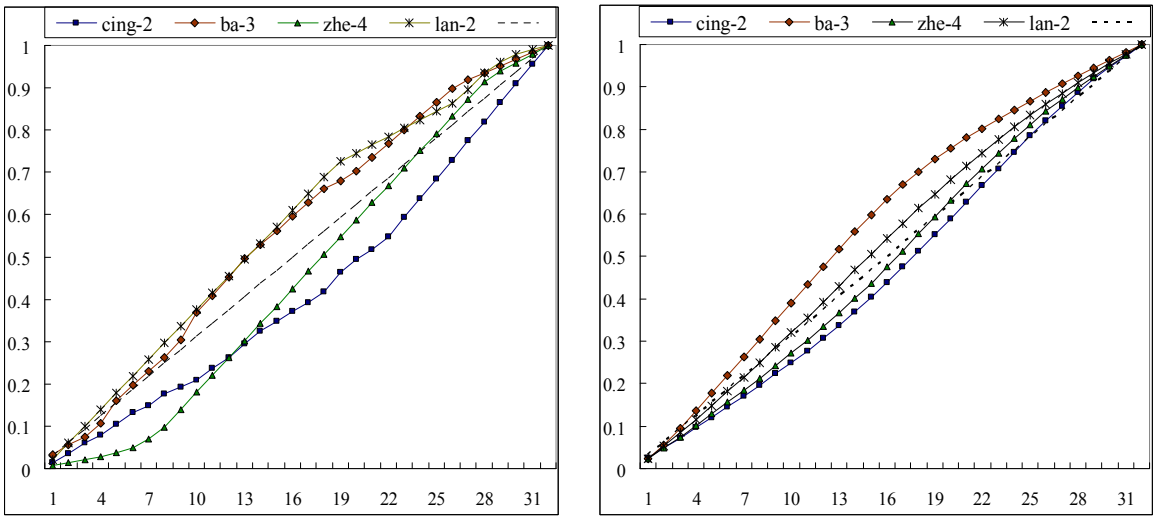
在參數分析階段，我們先對 2.1 節提到的單獨發音音節作 HNM 分析，以求得國語各音節各自的 HNM 參數，包括各個音框內表示雜音頻譜包絡的 cepstrum 係數、和各個諧波的頻率、振幅、相位參數。由於各個國語音節都只錄、存一次原始信號，所以在此只有一份 HNM 參數可供使用，也就是不能作單元選擇(unit selection)。在 HNM 的信號合成階段，首先要依據欲合成音節的時長數值(由韻律單元產生)，來決定在合成音的時間軸上要佈放多少個控制點(control point)。然後，對於各個控制點，要依據它所在的時間位置，去決定它上面的 HNM 參數數值，一個簡單作法是先以線性的時間軸對映方式，來找出原始音節上的對應的音框，再將該音框的 HNM 參數複製到該控制點上。由於本研究的重點是，探討頻演參數模型對於合成語音的流暢性的影響，系統製作上就是要去控制合成音和原始音時間軸的對映(mapping)關係，所以，我們要先把頻演參數模型產生的 32 個正規化時間的頻演參數，作片段線性內差以形成一個對映函數，然後依據這個對映函數，就可找出一個控制點所應對應的原始音上的兩個相鄰音框，再將此二音框的 HNM 參數作線性內差，並且複製到該控制點上。

當各個控制點上都有 HNM 參數之後，下一步還需考慮如何調整各控制點上的 HNM 參數數值，以使合成音節的音調能夠符合基週軌跡參數的規定。對於 HNM 合成法來說，音調高低的更改，不是只改動諧波頻率的數值就好了，因爲會牽動到音色(timbre)，發生音色隨著音高(pitch)在變動的不穩定現象。要在維持音色一致性的條件下，作音節音調的更改，其詳細方法可參考我們先前的研究成果[9]。

四、頻演路徑產生及聽測實驗

4.1 頻演路徑產生

我們可以一個訓練用的文句爲例，如”請把這藍兔子送走”，把它帶入第 2 節所建造的頻演模型，來產生出頻演路徑，然後和該訓練語句原始的頻演路徑作比較，以觀察兩者之間的異同。圖 6(a)畫的是/cing2 ba3 zhe4 lan2/四個音節原始錄音的頻演路



(a) 原始音節的頻演路徑　　(b) 合成音節的頻演路徑

圖 6 頻演路徑比較

徑，而圖 6(b)畫的是相同四音節以模型所產生出的頻演路徑，圖形中橫軸表示正規化的時間點，而縱軸表示頻演參數數值。比較圖 6(a)與圖 6(b)可發現，它們的相同點是，對應音節的頻演路徑具有相同的走向趨勢，例如圖 6(a)裡/cing2/的頻演路徑行走中間線(左下角至右上角之直線)的下方，而/ba3/則行走中間線的上方，這樣的現象也可在圖 6(b)裡看到。至於不同點方面，圖 6(a)裡原始音節的頻演路徑，其斜率變化較大，而圖 6(b)裡的斜率變化較小，因此會感覺較平順；此外，圖 6(a)裡的頻演路徑會偏離中間線較遠，而圖 6(b)裡的頻演路徑則偏離得較少。所以一般來說，ANN 頻演模型所產生出的頻演路徑，可以保有路徑的走向趨勢，但是路徑有向中間線靠攏的現象。

4.2 聽測實驗

合成語音的一種評估方法是主觀的聽覺測試。在此我們選擇了一篇短文，然後令圖 2 的”頻演參數產生”方塊，先直接產生出線性比例之頻演參數，而以此種頻演參數

合成出的語音檔案，以 Va 表示；另外，再令”頻演參數產生”方塊，以 ANN 頻演模型來產生出頻演參數，再拿去作語音合成，而得到的語音檔案以 Vb 表示。以前述兩種方式產生出的語音檔案，我們也已經放在網頁上 http://guhy.csie.ntust.edu.tw/spmdtw/，而可讓有興趣者來作試聽和比較。

接著，我們將合成的語音檔案 Va 和 Vb，分別讓 9 位受測者來進行聽測評估。分數是以比較的方式來評定，由受測者就前後兩個播放的語音音檔(先播 Va 再播 Vb)，評出那一個比較順暢，在此”順暢”的定義是，整句話的多個音節聽起來，連接得很緊密而沒有顆顆粒粒(形容音節像是各自獨立地在發音)的感覺。評分的方式是給一個-2 到 2 之間的整數值，正值代表第二個播放的語音檔比第一個播放的好，1 和 2 代表程度上的差別，2 表示第二個音檔比第一個明顯的好，1 則是表示稍好一些，而-1 和-2 代表第二個播放的語音音檔比第一個播放的較差，-2 表示明顯的較差，-1 表示稍差一些，至於 0 則代表聽不出兩個語音音檔的差異。聽測實驗後，將受測者的評分作平均，結果得到了 1.33 之平均分數，這表示 Vb 的確比 Va 流暢，並且流暢度的提升是感覺得出來的，至於提升的程度還不算很大，我們推測有幾位受測者，可能還未把自然度和流暢度的定義區分出來。

五、結論

本文提出了一個以 DTW 加上 ANN 來建立頻演參數模型的方法。對於以 DTW 匹配目標音節和參考音節的最佳路徑時，無聲聲母開頭的音節，常常會發生兩音節的聲、韻母邊界點無法正確對齊的問題，這可經由音節分段和兩段式 DTW 的作法來解決。此外，兩個要作匹配的音節，若發生時間長度相差太多的問題，這可以調整 frame shift 長度的作法來解決。

當建立頻演參數模型之後，將它和文句分析、韻律參數產生、信號波形合成等模組作整合，用以合成出國語語音信號，再把合成出的語音拿去作聽測實驗，初步結果顯示，我們的頻演參數模型的確可用以提升合成語音的流暢度。所以，本文提出的頻演參數模型，可說是 HMM 之外的一種可行的頻譜演進模型，並且它不需要和信號波形合成模組所用的聲學特性參數(如 HNM 的諧波參數)之間有依附的關係存在，不過我們無意和 HMM 爲基礎的合成方法，去比較誰好誰壞。此外，使用本文的語音合成系統所獲得的語音流暢性，可以驗證音節內頻譜演進的掌握，會比相鄰音節之間共振峰軌跡連續性的掌握，能夠獲得超過許多的效益。

參考文獻

[1] Wu, C.-H. and J.-H. Chen, “Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis”, *Speech Communication*, Vol. 35. pp. 219-237, 2001.

[2] Yu, M. S., N. H. Pan, and M. J. Wu, “A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-to-Speech System”, *International Symposium on Chinese Spoken Language Processing* , Taipei, pp. 21-24, 2002.

[3] Chen, S. H., S. H. Hwang, and Y. R. Wang, “An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech”, *IEEE trans. Speech and Audio Processing*, Vol. 6, No.3, pp. 226-239, 1998.

[4] Gu, Hung-Yan and Kuo-Hsian Wang, "An Acoustic and Articulatory Knowledge Integrated Method for Improving Synthetic Mandarin Speech’s Fluency", *International Symposium on Chinese Spoken Language Processing*, Hong Kong, pp. 205-208, 2004.

[5] Qian, Y., F. Soong, Y. Chen, and M. Chu, “An HMM-Based Mandarin Chinese Text-to-Speech System”, *International Symposium on Chinese Spoken Language Processing, Singapore*, Vol. I, pp. 223-232, 2006.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System", *International Conference on Spoken Language Processing*, Vol. 2, pp. 29–32, 1998.

[7] Yeh, Cheng-Yu, *A Study on Acoustic Module for Mandarin Text-to-Speech*, Ph.D. Dissertation, Graduate Institute of Mechanical and Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan, 2006.

[8] Gu, Hung-Yan, Yan-Zuo Zhou and Huang-Liang Liau, ”A System Framework for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech”, *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 12, No. 4, pp. 371-390, 2007.

[9] 古鴻炎、周彥佐，「基於 HNM 之國語音節信號的合成方法」，*第十九屆自然語言與語音處理研討會* (ROCLING 2007)，台北，第 233-243 頁, 2007。

[10] Rabiner, L. and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, 1993.

[11] O`Shaughnessy, D., *Speech Communication: Human and Machine*, 2’nd ed., IEEE Press, 2000.

[12] 古鴻炎、張小芬、吳俊欣，「仿趙氏音高尺度之基週軌跡正規化方法及其應用」，*第十六屆自然語言與語音處理研討會* (ROCLING XVI)，台北，第 325-334 頁, 2004。

[13] Gu, Hung-Yan and Chung-Chieh Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", *International Symposium on Chinese Spoken Language Processing*, Beijing, pp. 125-128, 2000.

[14] Yannis Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. Dissertation, Ecole Nationale Sup erieure des Telecommunications, Paris, France, 1996.

[15] 古鴻炎、廖皇量，「用於國語歌聲合成之諧波加噪音模型的改進研究」，*WOCMAT 國際電腦音樂與音訊技術研討會*，台北，session 2 (音訊處理 I)，2006。

一個結合 SVM 與 Eigen-MLLR 新的多語者線上調適架構應用於泛在語音辨識系統

A New On-Line Multi-Speaker Adaptation Architecture Combining SVM with Eigen-MLLR for Ubiquitous Speech Recognition System

施伯宜　Po-Yi Shih
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
hanamigi@gmail.com

林苑寧　Yuan-Ning Lin
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
yukinaco@hotmail.com

王駿發　Jhing-Fa Wang
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
wangjf@mail.ncku.edu.tw

摘要

本論文提出了一個以結合 SVM 和 Eigen-MLLR 爲基礎的線上多語者調適架構，應用於 ubiquitous 環境的語音辨識系統。語者獨立式的辨識系統相較於傳統的辨識系統有著更好的辨識效果，而語者調適方法便是其關鍵所在。本論文應用 SVM 和 Eigen-MLLR 的特性作爲調適技術的基礎，對於每個訓練語者的個別訓練語料做分類以及建立特徵參數向量空間。在語音辨識時，使用 SVM 找出測試語者所屬的類別，再找出類別相對應的 MLLR 特徵參數矩陣，並將其與非語者獨立模型結合成語者獨立模型。最後再利用辨識結果與原本的 MLLR matrix 和 Eigenspace 採取比重運算，並將運算結果更新原本的 MLLR matrix。相較於非語者獨立的辨識系統可以增加了 5~8%的辨識率。

Abstract

This work presents a novel architecture using SVM and Eigen-MLLR for rapid on-line multi-speaker adaptation in ubiquitous speech recognition. The recognition performance in speaker independent system is better than in conventional speaker dependence system, and the key point is speaker adaptation techniques. The adaptation approach is on the basis of

combine SVM and Eigen-MLLR, generating a classification model and building parameters vector-space for all speakers' individual training data. While in recognition, to find test speaker classification by SVM and look for MLLR parameters matrix correspond to speaker classification, then the MLLR parameters matrix and original acoustic model will integrate into speaker dependent model. Last, we estimate the adapted MLLR transformation matrix set by weighting function with recognition result, the present MLLR matrix, and Eigenspace. The estimate result will be used to update the MLLR matrices in adaptation phase. The experimental results show that the proposed method can improve 5% to 8% speech recognition accuracy with speaker adaptation.

關鍵詞：ubiquitous，語者調適，SVM，MLLR

Keywords: ubiquitous, speaker adaptation, SVM, MLLR,

一、緒論

（一）、研究動機

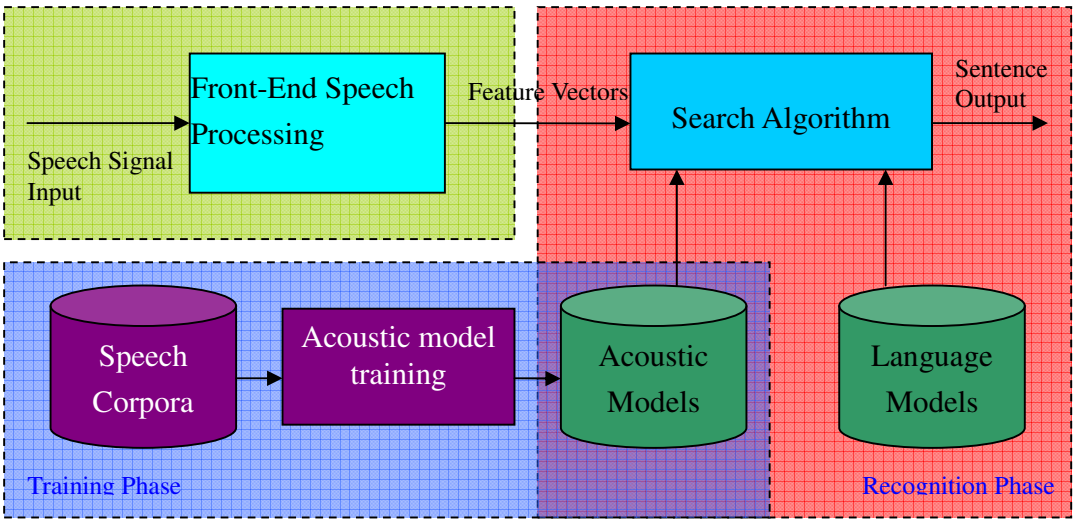
科技的發展始終來自於人類的需求，而以人爲本的生活應用科技一直是各方面研究的重點所在。在人的生活中，溝通，是人類這個社會生活中做重要的環節。人跟人之間的溝通模式之一，語言，創造出許多的生活價值，拉近人與人之間的距離，但在這個電腦化時代，資訊爆炸時期，語言是否可以在人跟電腦之間的溝通當中創造出另一種價值，因此語音科技便爲此提供有效的解決方法。

語音技術的發展已經有數十年，且累積了相當豐富的成果，對於這個時代的進步以及人們的需求也有相當的貢獻。在現有的語音技術中，能夠扮演人跟機器之間互動角色當屬語音辨識技術。語音辨識的研究當中，可以分爲聲音處理以及語言處理兩大方面，在聲音處理的研究上，語音模型是一切的根本，也可說是整個辨識系統的靈魂，聲音處理研究方面的重點之一。此外，語音辨識系統在應用方面也會碰到所謂強健性（robustness）的問題，一個系統如何在一般使用下，去學習適應環境以及新使用者的語音特性而來提升對此語者的語音辨識能力。而在這個課題中，語者調適（speaker adaptation）技術的研究變成爲一個重點課題。在傳統的語音辨識系統使用中，都是處於單一環境下和單一輸入，也就是只可以在特定地點使用辨識系統，但這對於現在的人類生活模式來說，這種模式已經無法滿足人們的生活需求。在這個 e 化的社會當中，所謂的數位生活（digital life）已經充斥在我們的生活之中，特別是在家庭中的數位生活應用，更希望可以隨時隨處都使用語音來控制、操作家中的各種電器生活用品，於是便發展出泛在語音辨識技術（Ubiquitous Speech Recognition Techniques），讓人們在家中的任何一個地方都可以透過語音辨識技術的應用來享受科技所帶來的生活便利。而語音辨識系統面對家中成員的不同，便更需要一套適合的且可以即時更新的多語者調適方法，讓語音辨識系統對家中每個成員的語音都有良好的辨識能力。

（二）、研究方向

在語音辨識系統中，如圖一所示，需要一套聲學模型（acoustic model）來模擬各式的語音特性，而這語音特性同時也包含了語者的特性。每個聲音模型初期則是使用大量的訓

練語料（training corpus）經由統計的方式建立而成。爲了避免聲學模型在使用時環境因素跟訓練時的環境因素不同而造成便是效能大打折扣，如背景雜訊，所以便盡量收集許多不同環境因素下的語料，以維持聲音模型應有的準確性。但對聲音模型影響最大的並非只有環境因素，要隨時面對不同語者的語音訊號也是一大挑戰。在語音訊號中，因爲每個語者天生的物理特性差異，如共振腔發聲習慣、說話腔調，所以這也歸類爲一種環境因素。一個基本的聲音模型，便是透過多位訓練語者提供的語料經由統計的方式而建立的語者不特定模型（speaker independent model, SI model），對於每個語者的訊號模擬程度只能算是中等，不可能各方面適合每個人的語音特性。所以爲了對不同語者都提供有良好的辨識效能，系統必須配合語者所提供的語料來做適當的修正、調整，這個工作便稱爲語者調適（speaker adaptation）。在使用語者提供的語料對原本的聲音模型來做加以修正之後，這個聲音模型便稱爲語者特定模型（speaker dependent model, SD model）。但這又便會碰到許多延伸的問題：如何在短暫的時間內以及語者提供少數的語料下可以快速的修正聲音模型，使辨識系統辨識能力提升也達到好的服務品質，這便是目前語者調適技術需要克服的一個問題。



圖一、語音辨識系統流程圖

在許多語者調適研究當中，早期是以貝氏調適法（Bayesian adaptation）爲基礎，這個方法的的優點是可以將語音模型完全配合語者做很精確的調整，但缺點便是它對需求的語料相當的大，這並不能滿足語音辨識系統在快速調適上的要求；此外，若是在未知語料內容的非監督式調適（unsupervised adaptation ）的情形下，反而只會大幅降低整個模型的精確度。近年來最熱門的語者調適方法是使用最大相似度線性迴歸（Maximum Likelihood Linear Regression, MLLR），許多的研究實驗結果也都證實了此方法可以在快速調適以及非監督式調適上有良好的表現。然而在 MLLR 中，需要估測大量的參數，然而如果在語料稀少的情況下，有時也會發生將聲音模型調整的更差。在 [1] 所提出的使用 Eigenspace-MLLR 爲基礎的快速調適法中，便考慮到這方面的兩個特點，不僅可以在語料甚少的情形下保有參數估測值的可靠性，並且在非監督式調適的情形下表現良好。本論文便以此方法來加以延伸在新的語者調適架構中。SVM 系統近年來都被使用在解決關於分類（classification）、回歸（regression）、以及快速的偵測（novelty detection）這些相關的問題上，且應用層面廣泛，不只在關於機器學習（machine learning）上，也有許多應用於語者鑑別（speaker verification）。在 [2] [3] 的研究便使用以 MLLR 爲基

礎核心的 SVM 來實現語者辨識與鑑別。在本論文中便以 SVM 爲來做基礎的語者分類，使得可以找出在 MLLR 中相對應的語者估測參數，並且可以更新原本的語者估測參數。

（三）、研究主題與主要成果

本篇論文的主題主要是著重在兩大方向上，第一個是如何將 SVM 與 Eigen-MLLR 兩個不同性質的演算法結合使用，亦即如何將所有調適語者在有限的少量調適語料下同時使用 SVM 對語者做分類，以及 Eigen-MLLR 建立所有語料的特徵向量參數空間，並且將 SVM 中的個別分類與 Eigen-MLLR 中的特徵向量參數空間建立一個相對應的關係，這稱爲訓練階段（Training phase）。之後在每一位測試語者使用時，便會先透過 SVM 將語料找出相近的分類，並從分類中找出相對應的 MLLR 特徵參數，在使用此特徵參數與原本的聲音模型結合成語者特定模型，再進行語音辨識，這稱爲辨識階段（Recognition Phase）。第二個便是結合辨識結果，語者相對應的 MLLR 特徵參數以及特徵參數估測三者來使用比重取決，並且將結果來即時更新原本語者相對應的 MLLR 特徵參數，這稱爲調適階段（Adaptation phase）。根據實驗，我們結合使用了 SVM 以及 Eigen-MLLR 的方法，在對於泛在的語音辨識中辨識率提升確實有加成的作用，在整體效能上又向上提升。

在第二章中主要是介紹本論文中所主要使用的調適方法以及語者調適技術在實行及 SVM 的技術和應用簡介。第三章便會詳細的介紹研究主題的架構以及實行情況，第四張便會繼續介紹整個實驗環境和過程以及結果，並且會加以探討。第五章則是本論文報告的結論以及未來的相關研究遠景。

二、相關技術回顧

在這一章節將對語者調適的實行以及性質上作簡介，以及對於 SVM 和 MLLR 演算法做概要的敘述。語者調適法的角度可以從系統、語料以及本質等三方面探討，如表一所列。SVM 系統近年來都被廣泛的使用在分類（classification）、回歸（regression）、以及快速的偵測（novelty detection）相關的問題上。SVM [4] [7] 的所有相關技術也可以視爲是使用在分類和回歸的監督式學習演算法。MLLR 的基本意義是在於假設調適後的參數群組與現有的基準參數群組存在著一個線性迴歸函數的關係，而可以藉由使用最大相似度（Maximum Likelihood, ML）測法來求得現有參數群組間的函數關係。

（一）、語者調適的簡介

語者調適技術的目的在於利用語者所提供的有限語料，來改善辨識系統對於與使用者的辨識能力。表一爲現有調適法類別的分類。從系統的角度來看，當系統要進行語者調式工作時，須先獲得語者提供的語料，此稱爲訓練語料（training data）或是調適語料（adaptation data）。假如系統是以每收到一句語料便調適一次的話，稱爲循序調適法（sequential adaptation）；如果是一次獲得所有的語料再做調適的話，便稱爲批次調適法（batch adaptation）。如果從調適語料的角度來看，如果系統事先得知語料的內容，也就是清楚知道語料每一句的標音，系統及可以找出語音訊號和相對應的語音參數做精確的調整，這稱爲監督式調適法（supervised adaptation）；相反的，若不知道語料的內容，須先對語料辨認過後，才將辨認結果當作語料的內容來調適，則稱爲非監督式調適法

（unsupervised adaptation）。

辨識系統中原有的聲音模型（acoustic model）可以稱爲語者不特定模型（speaker independent model, SI model），又稱爲初始模型（initial model）。若以模型爲基礎的調適方式，則須先使用初始模型對語音訊號進行分析（切音），並找出每個音框（frame）所相對應最有可能的模型狀態，甚至還必須先進行一次辨認，因此，在辨識階段的搜尋演算法（Search Algorithm，如 Viterbi）中所獲得的結果也會與初始模型有相當大的關係。所以調適演算法的精確度也跟初使模型的切音有相當大的關係。

若回到語者調適的本質來看，又可分爲兩種基礎調適法：一是利用語音訊號的特徵向量爲基礎的調適法（feature-based adaptation），主要是以調整語音訊號的特徵向量，使得現有模型參數更能精確描述變化情形；一是調整模型的參數爲基礎的的調適法（model-based adaptation），使得可以更有效模擬語者的特性。近來相當多的研究顯示，以模型爲基礎的方法優於以特徵爲基礎的調適法，其在實作上的要求與系統複雜度也較爲簡單，而許多的熱門調適技術都是採用此性質的調適方式，如最大相似度線性迴歸（MLLR）、貝氏調適法（又稱爲最大事後機率法，Maximum a Posterior，MAP）。

表一、語者調適法的類別

語者調適法分類	調適法類別
以系統分類	循序調適法（sequential adaptation），批次調適法（batch adaptation）。
以語料分類	監督式調適法（supervised adaptation），非監督式調適法（unsupervised adaptation）。
以本質分類	特徵向量爲基礎的調適法（feature-based adaptation），模型參數爲基礎的調適法（model-based adaptation）

（二）、支援向量機（Support Vector Machine, SVM）

支援向量機（SVM）是目前經常使用來做爲分類（classification）或回歸（regression）的方法。當給予一群已經分類好的資料之後，支援向量機可以經由訓練（training）獲得一組模型（model）。之後，若有未分類的資料加入時，支援向量機便可以依據先前訓練出的模型去做預測（predict），並決定這筆資料所屬的分類。而因爲在建立起模型時，必須要有已經分類好的資料做爲訓練，所以是屬於監督式學習（supervised learing）的方法。支援向量機也是一種線性分類（Linear Classification）的方法，目的在於找出一個超平面（hyperplane）而可以將在特徵空間中（feature space）已經分類好的資料清楚的分開成不同的類別。

支援向量機最重要的特性便是確定模型參數符合最佳化的問題，特別是能把區域性的最佳化當成全域性的最佳化，因爲 SVM 使用 Lagrange multipliers 來探討延伸整個系統的最佳化問題。在參考文獻 [4] 表示，SVM 的基礎是一個利用核心函數（Kernel function）$K(\cdot,\cdot)$的總和來建立一個兩類別的分類器，其基本數學式如下：

$$f(\mathrm{x}) = \sum_{i=1}^{L} \gamma_i t_i K(\mathrm{x}, \mathrm{x}_i) + \xi \tag{1}$$

t_i 表示爲理想的輸出值，$\sum_{i=1}^{L} \gamma_i t_i = 0$ ，並且 $\gamma_i > 0$ 。向量群 Xi 爲藉由最佳化處理從訓練集合中獲得的支援向量群（Support Vectors）。理想的輸出值藉著相關的支援向量是落在類別 1（其值爲＋1）或類別 2（其值爲－1）。對於分類來說，數值 $f(x)$所屬類別的決定是在於所定的標準（threshold）之上或之下。

（三）、特徵式最大相似度線性迴歸（Eigen-Maximum Likelihood Linear Regression, Eigen-MLLR）

首先我們從原始的 MLLR 演算法來看 [3] [5]，此方法的原理背後是透過假設改變後的參數會和原本的基本參數間唯一線性迴歸的函數關係：如下

$$Y = AX + B \tag{2}$$

若將此原理使用在語者調適中，語料的每一個特徵向量接代表著語者與因空間中的其中一個樣本，而我們也假設了待求參數和現有參數之間的函數關係，因此可以便可以使用最大相似度（Maximum Likelihood, ML）估測法來求得 A、B 的值，如下所示：

$$\begin{aligned} A &= \arg\max_{A} f(x \mid \theta, A, B) \\ B &= \arg\max_{B} f(x \mid \theta, A, B) \end{aligned} \tag{3}$$

在這裡 θ 指的是所有語音參數模型的集合，x 則表示調適語料的觀測值。

在許多的語者調適法當中，最大相似度線性迴歸的方法被廣泛的應用在快速語者調適，例如它只需要些許的語料便可以對模型參數做調適。在最大相似度線性迴歸中，非語者獨立的模型參數可以根據一個或多個仿射轉換方式（affine transformations）來達到調適的目的。最大相似度線性迴歸調適法是使用仿射轉換方式來調適高斯混合模型（Gaussian mixture model, GMM）中所有混合元件的平均值（mean），而同一個仿射轉換法則可以提供所有的混合元件共享，表示法如下：

$$\hat{}_i = \mathrm{A}\ _i + \mathrm{b} \quad \forall i \tag{4}$$

$_i$表示在 GMM 中還沒被調適過的平均值，而 $\hat{}_i$表示已經調適過的平均值。
在眾多的資料量當中，混合的元件可以被歸類成多個類別，且不同的仿射轉換方式可以在不同的類別中被共用，延伸(4)的表示法：

$$\hat{}_i = \mathrm{A}_1\ _i + \mathrm{b}_1 \quad \forall \hat{}_i \in \mathrm{class}_{1,} \tag{5}$$

$$\hat{}_i = \mathrm{A}_2\ _i + \mathrm{b}_2 \quad \forall \hat{}_i \in \mathrm{class}_{2,} \tag{6}$$

在單一和多個類別的情況下的轉換方式是透過選擇最大的相似度來決定的。由於在這個調適方法中是利用到語料共享的觀念，當有一模型在沒有任何的調適語料情形下也可以藉助同一類別中有語料的模型來求出 A, b 值，所以在每次調適時，只要選到適當的類別，則所有的模型參數都可以調整到，這也是最大相似度線性迴歸法可以應用於快速調適語者的主要原因。

特徵式-MLLR [8] [10]爲一改良式的 MLLR，其目的是以特徵向量空間計算 MLLR 回歸矩陣（Eigen-MLLR），主要是採取了以向量空間爲基礎之語者調適技術以及傳統 MLLR

的優點。向量空間爲基礎之語者調適技術的優點，就是將所有訓練語者的聲學參數向量化，並利用這些向量撐出一片可信賴的聲學向量空間。此後，在執行語者調適時，所做的工作便只是找出測試者在這聲學空間上的最可能的位置，因此所處的位置一旦被決定，測試者最後整套的聲學參數也就呼之欲出了。其向量空間的建立步驟如下：

Step1.模型參數向量化：首先，我們先將每位訓練語者的語者特定模型參數向量化，即是對於整套 SD model 裡所有高斯混合的平均值向量，像火車一節一節地串接起來。如此，我們便可得到一個高維度的向量，亦即若高斯混合的平均值向量之維度爲 n，一套 SD 模型中有 M 個高斯混合，則此新向量的維度 D 便是 $M \times n$。

Step2.將模型參數向量組成矩陣：接下來，有了這些由所有 SD 模型參數所組成的向量後，扣除向量中各維度的平均值，隨後再將這些向量以 row vector 的姿態組成一個矩陣 Z，若語者的個數爲 K，則 Z 的大小即爲 $K \times D$。

Step3.計算空間基底 最後再計算相關矩陣 $Z^T \cdot Z$(correlation matrix ,Z^T 爲 Z 的轉置矩陣)的特徵向量(eigenvector ,維度亦等於 D)後便大功告成。這些求得的特徵向量即做爲向量空間的基底(basis)，在語者調適中成爲一具代表的特徵，所以求得的空間便稱做特徵向量空間(Eigenspace)。

當 Eigen-MLLR 在實現時，建構特徵向量空間的取材已由模型參數改爲 MLLR 回歸矩陣(還包含偏移向量 b)，所以我們可以從中發現，除了矩陣 A 或著是向量 b 中的參數各自有不同的意義外，兩者的參數在物理意義上更是截然不同。因此在這情形下，所有相異的參數在事前做正規化的處理就愈顯得格外的需要。Eigen-MLLR 相較於傳統的方式，在有適當的訓練語料下，是更加提高了矩陣估測值的準確性。

三、所提出的演算法架構

在本段中我們將會介紹一些實驗背景所採用的方法及設定，並且敘述架構中每個部份。

（一）、實驗相關背景

在建立語音辨識系統的基本聲學模型，又稱爲語者不特定模型或是初始模型，我們是使用台灣口音中文語料庫(Mandarin Across Taiwan, MAT-400)來做爲訓練語料庫，裡面包含了約 400 個語者所錄製的共 5,000 個不同的語音檔，以及 77,324 個詞語和 5,353 個句子。其詳細內容如下表：

表二、台灣口音中文語料庫簡介

Databases	Number of Files	Prompting Item Numbers	Speaking Style	Description
MATDB-1	3600	1-9	spontaneous	Short answering statements
MATDB-2	2000	10-14	read	Numbers pronounced in five different ways
MATDB-3	4800	15-26	read	Mandarin syllables
MATDB-4	12000	27-56	read	Words of 2 to 4 syllables
MATDB-5	4000	57-66	read	Phonetically balanced sentences

語音訊號的特徵參數中採用 13 維的梅爾倒頻譜系數(MFCC)及一階和二階的回歸係數 Δ (delta)，設定如表三：

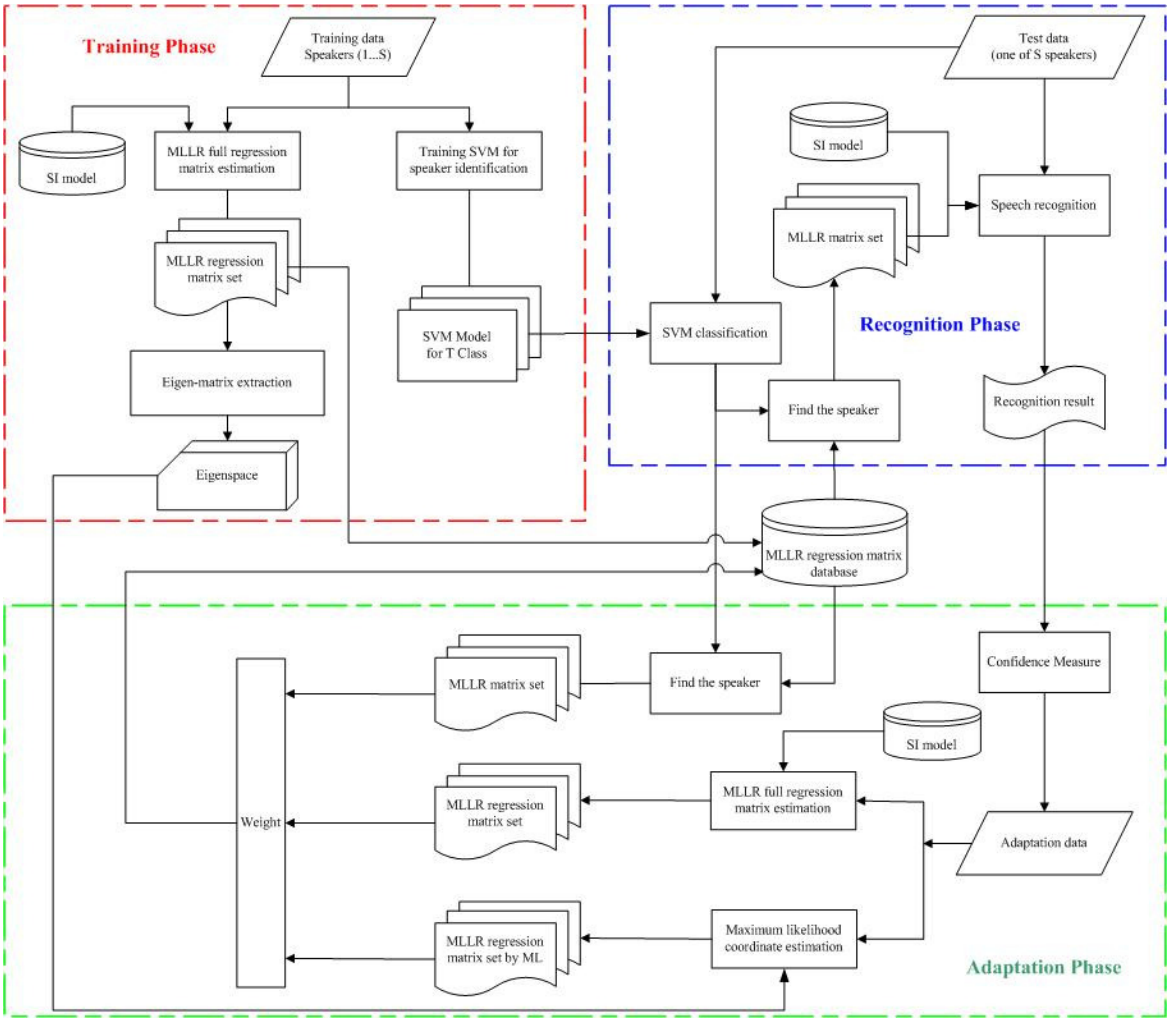
表三、本論文中所使用的參數擷取設定

取樣頻率	8 kHz
預強濾波器	$1\text{-}0.97z^{-1}$
分析視窗	Hamming Window
視窗長度	20ms
音框平移	10ms
梅爾濾波器個數	23
特徵向量參數	13 MFCC + Δ + ΔΔ & log energy + Δ + ΔΔ

在聲學模型架構中，則是採用最通用的連續密度馬可夫模型(continues density hidden Markov models, CDHMM)，採用由左至右(left-to-right)的型態，也就是狀態轉移上只允許從抹一狀態跳至鄰近的下一狀態或是停留在原本的狀態上。而在模型單位的選取，則採用音節右相關聯音素模型，每個音素模型包含 5 個狀態，3 個音素、1 個靜音、1 個短暫停。整個聲學模型是利用劍橋大學所提供的 HTK [6] 工具來建立。

（二）、演算法架構

整個演算法的架構可分爲三個階段，訓練階段（Training Phase），辨識階段（Recognition Phase）和調適階段（Adaptation Phase）。當系統一開始預設時，會先使用訓練語料來建立初始的模型，當建立之後，便會在每次作辨識的時候再針對個別語者的語音模型做調適。整個流程的框架結構如下圖：



圖二、語者調適演算法架構圖

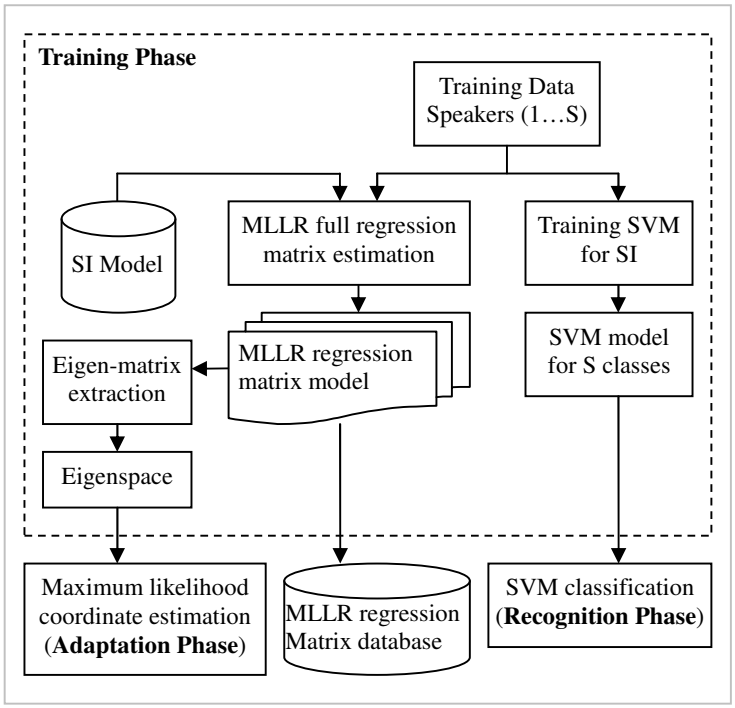
我們在這針對個別的階段來做詳細介紹。

訓練階段（Training Phase）：

在訓練階段，首先必須讓所有的測試語者都說一定量的語句來當做調適訓練語料，使用的方式是監督式調適學習。假設共有 S 個訓練語者，對每一個語者必須使用訓練語料和非語者獨立模型參數計算總共 C 個的傳統最大相似度線性迴歸完全回歸矩陣（MLLR full regression matrices），也對每一位語者用其全部的語料做一次最大相似度線性迴歸的調適計算以取得回歸矩陣，並且在訓練語料中找出足夠可以描述語者特徵的數量。而對於每一個訓練語者來說，所有 C 個的最大相似度線性迴歸迴歸矩陣可以當成單一語者特有的矩陣集合。接著從 S 個語者特有的矩陣集合當中攫取出 S 個根本構成要素，這便稱爲特徵矩陣（eigen-matrices），最後我們便取 S 個特徵矩陣。在這個階段中，傳統最大相似度線性迴歸完全迴歸矩陣的計算和以特徵空間爲基礎迴歸矩陣估測是分開始實現的。另一方面，這階段同時將調適訓練語料使用支援向量機（SVM）來做語料分類，共有 S 個類別，並且在建立語料分類的同時也建立起跟迴歸矩陣之間的相關性，使得每一分類都有其相對應的特徵矩

陣，而這相對應關係會是在辨識時最重要的。

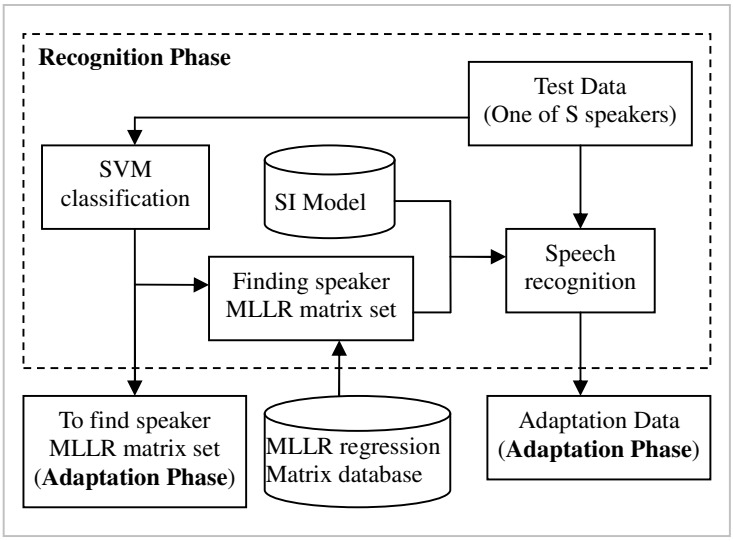
在有別於傳統的最大相似度線性回歸，我們是利用建立一個迴歸矩陣特徵空間來尋找特定語者的最大相似度線性回歸回歸矩陣，而不需要透過語者的語料求得迴歸矩陣中的每個參數，只需要利用語料找出回歸矩陣最有可能在特徵向量空間中的所在位置。也由於這個特徵空間是使用足夠的語料來估測得到的 MLLR 回歸矩陣，所以能替回歸矩陣估測提供相當足夠的資訊。如圖三所示。



圖三、訓練階段架構圖

辨識階段（Recognition Phase）：

圖四爲辨識階段的流程圖。在辨識階段中，從提供訓練語料的語者中挑選一個做測試，當語者說話經由語音特徵擷取後，便會傳送至 SVM 中做分類，若發現與第 n 個類別最相近，便從最大相似度線性回歸回歸矩陣集合中挑出相對應第 n 個迴歸矩陣，再將此迴歸矩陣跟原本的非語者特定模型結合成語者特定模型，再以這個模型進行語音辨識，然後輸出結果。這個辨識結果同時也提供給在調適階段中更新回歸模型的一個主要參考。



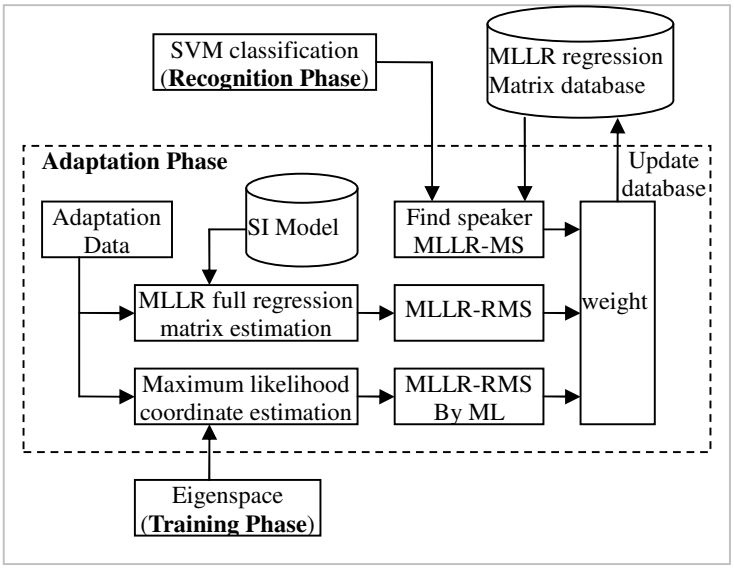
圖四、辨識階段架構圖

調適階段（Adaptation Phase）：

在經過辨識階段取得辨識結果後，會將辨識結果當成爲調適語料（adaptation data），對於每一個測試語者所提供的調適語料會使用最大相似度估測（Maximum Likelihood estimate, ML）方法來將語者定位在特徵空間中的回歸矩陣。最大相似度估測和最大相似度線性回歸會使用調適語料來個別做一個新的估測。個別估測後所獲得的結果將會和利用 SVM 所找出對應語者的 MLLR 回歸矩陣，三者做一個 weighting 的運算，並將獲得的結果對原本語者的 MLLR 回歸矩陣作更新。於是我們修改了 [10] 中的 Equation (5) 增加了一個調適信心度的估算。下列爲一個更新值的運算。

$$\hat{\mathrm{W}}_c = \xi_{conf} \cdot \left(\frac{\lambda \cdot \mathrm{W}_c^{\,EIGEN} + \sum_{m=1}^{M} \sum_{n=1}^{N_c} \gamma_n(m) \mathrm{W}_c}{\lambda + \sum_{m=1}^{S} \sum_{n=1}^{N} \gamma_n(m)} \right) + (1 - \xi_{conf}) \cdot \mathrm{W}_c^{\,present} \qquad (7)$$

M 表示爲特徵向的數目，n 表示爲在 N_c 中的一個混合元件，$r_n(m)$表示在時間爲 t 時的觀測機率，$\mathrm{W}_c^{\,present}$，$\mathrm{W}_c^{\,EIGEN}$，$\mathrm{W}_c^{\,estimate}$ 則分別爲類別 c 目前的回歸矩陣，由 ML 估測出來的特徵空間矩陣，以及由 MLLR full regression 估測出來的回歸矩陣。$\hat{\mathrm{W}}_c$ 爲更新過後的回歸矩陣。ξ_{conf} 爲信心比重，信心比重則是依照辨識結果而得的。增加信心比重參數是避免當發生辨識結果產生錯誤而接受，也就是對的語音辨識成錯的，或錯的語音辨識成對的，來導致整個調適模型越來越差。整個調適過程如圖五。



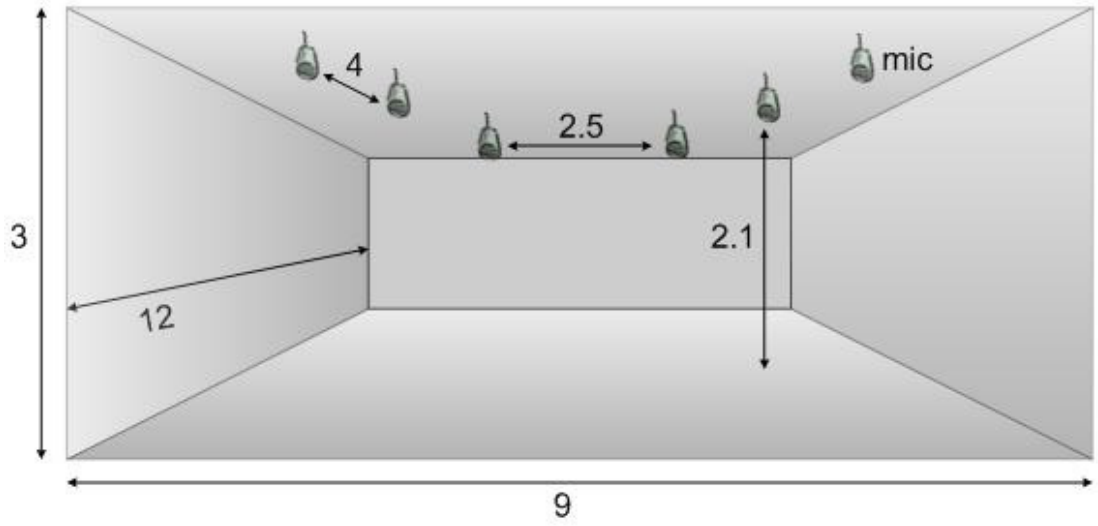
圖五、調適階段架構圖

四、實驗結果

我們實現實驗的結果是以台灣口音中文語料 MAT-400 爲基礎的語音模型。調適語料則是以日常生活用語大約 7~10 個字，而在測試時則以人名或是日常生活用語爲主。語音特徵的擷取設定則定義在表三。我們使用 MLLR 調適製造出擴張每一個語句的基底 SVM 特徵，對於高斯混合模型的每一個具體的類別都要做調適。我們便使用 HTK[6] 來建立上述的工作並且產生一遞迴式 MLLR 來建立轉換式。

對於已經提供訓練語料的目標語者，每句話都會用一個 SVM 特徵向量來代表，而選擇 SVM 的類別則會透過目標語者的特徵向量和原有的 SVM 特徵向量類別來作一比重選擇。

在我們的實驗中，使用了泛在語音辨識系統（ubiquitous speech recognition system）來測試，總共使用了 6 支全向性麥克風佈在房間的天花板，其收音範圍可以涵蓋了整個房間內部，如圖六。由於其排列方式並非傳統的麥克風陣列方式，無法使用傳統的雜訊消除法來達到較好的效果，所以便將 6 支麥克風使用多通道混音（multi-channel mixer）的方式成單一輸入，再使用子空間式語音增強演算法 [9]（Subspace Speech Enhancement, Using SNR and Auditory Masking Aware Technique）在語音訊號的前處理來消除語音雜訊，再開始取語音特徵。



圖六、泛在麥克風陣列示意圖

實驗過程總共使用了 10 個人爲提供訓練語料以及調適的語者，每個語者的訓練語句都爲 15 句。所採取的辨識率計算以正確率爲主，皆以百分比表示，計算方式如下：

$$\text{辨識正確率}=\frac{\text{辨識正確語句句數}}{\text{全部語句句數}}\times 100\% \tag{8}$$

表四爲在尚未有任何語者調適方法的泛在語音辨識系統下所測試的結果，可以發現初始聲音模型確實是對任何一個語者都提供相同的辨識準確效能，總平均準確率爲 85.7%。

表四、未有任何語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	130	124	126	133	129	126	127	131	134	125	1285
錯誤語句	20	26	24	17	21	24	23	19	16	25	215
辨識正確率(%)	86.6	82.6	84	88.7	86	84	84.7	87.3	89.3	83.3	85.7

接著我們以傳統的 MAP 語者調適方式來使用於系統，表五爲使用 MAP 語者調適後的結果。可以發現對於每個語者做調適之後，辨識系統對於每個人的準確率平均都有 2%~3%的準確率提升，總平均準確率爲 88.2%。

表五、使用 MAP 語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	135	126	131	136	134	129	132	135	137	128	1323
錯誤語句	15	24	19	14	16	21	18	15	13	22	177
辨識正確率(%)	90	84	87.3	90.7	89.3	86	88	90	91.3	85.3	88.2

表六則使用傳統 MLLR 語者調適法使用於辨識系統的辨識結果，相較於 MAP 調適法，有著更高的辨識率，會造成這樣的結果最有可能的原因是 15 句的調適語料對 MAP 來

說是非常的少量，而使得 MAP 真正的優點根本還來不及發揮，相較於 MLLR，反而能在少量語料時就顯示出相當出色的表現，比未調適的正確率多出 5%~8%，比 MAP 多出 3%~5%的辨識正確率。

表六、使用 MLLR 語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	139	136	137	142	142	135	142	141	142	138	1394
錯誤語句	11	14	13	8	8	15	8	9	8	12	106
辨識正確率(%)	92.6	90.6	91.3	94.6	94.6	90	94.6	94	94.6	92	92.9

表七的實驗結果則爲本論文所提出應用特徵式 MLLR 與 SVM 的語者調適架構。與 MAP 調適法來比較，由於藉著 MLLR 的特性關係，還是可以藉著少量語料而達到明顯的效果，且利用 SVM 來直接選擇相對應的特徵參數矩陣，少掉了重新由語料計算參數的運算量，也比傳統的 MLLR 平均提升了 1%左右辨識率，相較於未調適的辨識系統更提升了 8%平均辨識率，相較於 MAP 平均提升了 4%~5%辨識率。

表七、本論文所提出的語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	142	137	139	144	141	138	140	144	145	138	1408
錯誤語句	8	13	11	6	9	12	10	6	5	12	92
辨識正確率(%)	94.7	91.3	92.7	96	94	92	93.3	96	96.7	92	93.9

五、結論

在語音辨識系統中，語者調適的工作是整體辨識效能的一個很重要的環節，而且對於在多人使用的環境下，語者調適技術在快速調適以及非監督式調適的情況下就更必須加強，這兩方面也會是今後語者調適技術發展的重點。一個調適技術不僅要能夠達成快速調適，且必須能夠在只有少量語料的情形下將現有原始的聲學模型調整至更適合當下語者的狀態，特別是對於在一個空間中的固定成員更是有所需要，如家庭成員。在本論文提出的架構中，利用特徵式最大相似度線性回歸（Eigen-MLLR）所建立的多語者特徵向量空間結合支援向量機（SVM）的分類來達成快速多語者調適。測試語者的語句經過 SVM 分類完畢之後，便會在 MLLR 回歸矩陣群中找出 SVM 類別相對應的回歸矩陣並且與初始模型結合成語者特定模型，再進行語音辨識。然後將語音辨識的結果利用 MLLR 回歸矩陣估測（MLLR regression matrix estimate）以及最大相似度估測（Maximum Likelihood estimate）三者來使用比重取決重新計算，並且將結果來即時更新測試語者相對應的 MLLR 回歸矩陣參數。在本論文中也發現，若可以再加強語音增強處理的演算法降低更多雜訊以及提升訊號強度，則對整個語者調適和辨識效能再進一步提升。在未來的語音辨識環境中，希望能夠增加更多的麥克風，使的能達到寬容度更高的泛在語音使用環境，也可以隨時加入新的語者讓系統可以自我更新以及作調適，也希望這項技術能結合其他的應用到更廣泛的層面，讓人們生活可以藉著數位化更便利。

參考文獻

[1] K. Chen et al, “Fast speaker adaptation using eigenspace-based maximum likelihood linear regression,” in Proc. ICSLP, Beijing, Oct. 2000.

[2] P. C. Woodland, “Speaker Adaptation: Techniques and Challenges”, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.85-90, 2000.

[3] Brian Kan-Wing Mak, and Roger Wend-Huu Hsiao, “Kernel Eigenspace-Based MLLR Adaptation”, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, Mar 2007.

[4] Nello Cristianini and John Shawe-Taylor, *Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[5] M.J.F.Gales and P.C.Woodland, “Mean and variance adaptation within the MLLR framework”, *Computer Speech and Language*, vol. 10, no. 4, pp. 249-264, 1996.

[6] S.Young et al, “The HTK book,” http://htk.eng.cam.ac.uk.

[7] Zahi N. Karam and William M. Campbell, “A multi-class MLLR kernel for SVM speaker recognition,” in ICASSP 2008.

[8] Nick J.C.Wang, Wei-Ho Tsai and Lin-shan Lee, "Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification," 2001 European Conference on Speech Communication and Technology, Aalborg, Denmark, Sept 2001

[9] Wang Jia-Ching, Lee Hsiao-Ping, Wang Jhing-Fa, and Yang Chung-Hsien, “Critical Band Subspace-Based Speech Enhancement Using SNR and Auditory Masking Aware Technique”, IEICE Transactions on Information and Systems. vol 90; number 7, pages 1055-1062, July 2007

[10] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, “Fast speaker adaptation using eigenspace-based maximum-likelihood linear regression,” in Proc. ICSLP, 2000, vol. 3, pp. 742–745.

調變頻譜正規化法使用於強健語音辨識之研究
Study of Modulation Spectrum Normalization Techniques for Robust Speech Recognition

王致程 Chih-Cheng Wang
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
s95323553@ncnu.edu.tw

杜文祥 Wen-hsiang Tu
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
aero3016@ms45.hinet.net

洪志偉 Jeih-weih Hung
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
jwhung@ncnu.edu.tw

摘要

自動語音辨識在實際系統應用中，語音信號經常受到環境雜訊的影響而降低其辨識率。爲了提升系統的效能，許多研究語音辨識的學者歷年來不斷地研究語音的強健技術，期望能達到語音辨識系統的最佳化表現。在本論文中，我們主要是受時間序列結構正規化法觀念所啓發，進而探討並發展出更精確有效的調變頻譜正規化技術。我們提出了三種新方法，包含了等連波時間序列濾波器法、最小平方頻譜擬合法與強度頻譜內插法。這些方法將語音特徵時間序列的功率頻譜密度正規化至一參考的功率頻譜密度，以得到新的語音特徵參數，藉此降低雜訊對語音之影響，進而提升雜訊環境下的語音辨識精確度。同時，我們也將這些新方法結合其他特徵強健化的技術，發現這樣的結合能帶來更顯著之辨識率的提升。

Abstract

The performance of an automatic speech recognition system is often degraded due to the embedded noise in the processed speech signal. A variety of techniques have been proposed to deal with this problem, and one category of these techniques aims to normalize the temporal statistics of the speech features, which is the main direction of our proposed new approaches here.

In this thesis, we propose a series of noise robustness approaches, all of which attempt to normalize the modulation spectrum of speech features. They include equi-ripple temporal filtering (ERTF), least-squares spectrum fitting (LSSF) and magnitude spectrum interpolation (MSI). With these approaches, the mismatch between the modulation spectra for clean and noise-corrupted speech features is reduced, and thus the resulting new features are expected to be more noise-robust.

Recognition experiments implemented on Aurora-2 digit database show that the three new approaches effectively improve the recognition accuracy under a wide range of noise-corrupted environment. Moreover, it is also shown that they can be successfully

combined with some other noise robustness approaches, like CMVN and MVA, to achieve a more excellent recognition performance.

關鍵詞：語音辨識、調變頻譜正規化、強健性語音特徵參數

keyword: speech recognition, modulation spectrum, robust speech features

一、緒論

自動語音辨識系統(automatic speech recognition systems, ASR)，藉由多年來各方學者的研究發展，逐漸達到實際應用的階段，而爲人類生活帶來更多方便與幫助，雖然還不能達到一個完美的地步，但是這方面的技術仍一直不斷地進步當中。

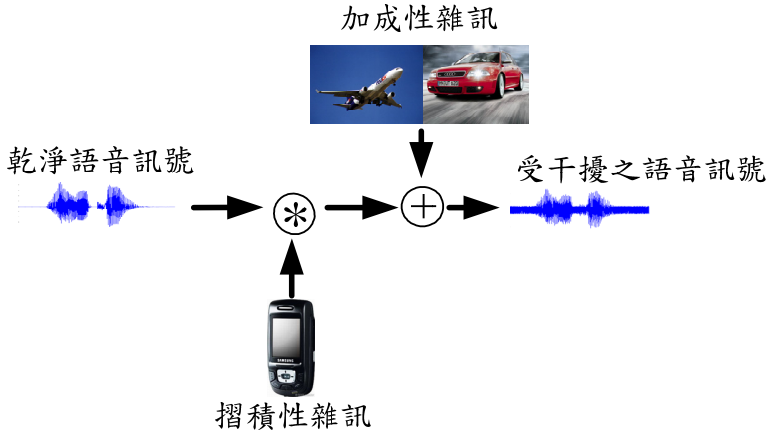
自動化語音辨認仍有許多相當具有挑戰性的研究課題，由於語音的變異性太多，例如每位語者說話的方式與口氣都不一樣、不同語言有不同的特性、語者當時說話的情緒、語者所處的環境是否有其他雜訊干擾等，這些變異對於語音辨識效果都有影響。在真實應用環境下，語音辨識系統所遇到的主要問題其中兩個，分別爲：

（一）語者不匹配(speaker mismatch)

語者不匹配的問題是因爲說話者先天條件（如口腔形狀）與後天習慣（如說話腔調）的差異所產生的變異性，因此當以特定語者所訓練出來的聲學模型來辨識不屬於此特定語者的語音時，辨識效果常會明顯下降，而要克服這一類問題的方法，通常是使用所謂的語者調適(speaker adaptation)技術。也就是將原本訓練出來的聲學模型調適成接近當下語者之語音特性的模型[1]，如此便可提高辨識率。

（二）環境不匹配(environment mismatch)

環境不匹配的問題是因爲語音辨識系統訓練環境與我們實驗或應用時的環境不同所致，其變異因子主要包含了加成性雜訊(additive noise)，如車站四周的雜訊、嘈雜街道的人聲或車聲等，及摺積性雜訊(convolutional noise)，如不同的有線或無線電話線路或麥克風所造成的通道效應等，語音辨識系統常會因這些雜訊的影響使辨識率降低。下圖一爲乾淨語音受雜訊干擾之示意圖。



圖一、乾淨語音受雜訊干擾之示意圖

在諸多降低雜訊影響、改進語音特徵的強健性技術中，有一大類的方法其目標是找出一強健語音特徵表示式(robust speech feature representation)，降低語音特徵對雜訊的敏感度，使雜訊產生的失真變小。此類著名的方法包括了倒頻譜平均消去法(cepstral mean subtraction, CMS)[2]、倒頻譜平均與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]、相對頻譜法(RelAtive SpecTrAl, RASTA)[4]、倒頻譜平均與變異數正規化化結合自動回歸動態平均濾波器法(cepstral mean and variance normalization

plus auto-regressive-moving-average filtering, MVA)[5]、倒頻譜增益正規化法 (cepstral gain normalization, CGN) [6]、資料導向時間序列濾波器法(data-driven temporal filter design)[7]等。以上這些方法皆是在語音特徵的時間序列域(temporal domain)作處理，根據語音訊號與雜訊在時間序列域上不同的特性，強調出語音的成分，而抑制雜訊的影響。近來，新加坡大學之李海洲博士研究團隊，新推出了一套時間序列濾波器設計的新方法，稱爲『時間序列結構正規化法』(temporal structure normalization, TSN)[8]，此方法的目的，在於將語音特徵序列之功率頻譜密度(power spectral density)正規化，使其輪廓逼近於一參考功率頻譜密度，此方法所得的時間序列濾波器，可以因應不同雜訊環境的語句特徵而加以調適，在其文獻[8]可知，當此新方法所得的時間序列濾波器作用於 CMVN 與 MVA 處理後的梅爾倒頻譜特徵參數時，在各種雜訊環境下所得到的語音辨識精確率都能有大幅改進。

雖然 TSN 法對語音特徵具有優異的強健化效果，且執行複雜度極低，但根據我們的觀察，此法仍然有幾點可以改進之處，首先，TSN 所得的初始濾波器係數是參考頻率響應之反傅利葉轉換求得，然後將這些係數會乘上一個漢寧窗(Hanning window)以減緩不當高頻成份的產生，此求取濾波器的方法未必是最佳化的，所得之濾波器係數其頻率響應可能與參考頻率響應之間的誤差較大。其次，在 TSN 法中，濾波器係數和被正規化爲 1，代表其直流增益爲一定值，此步驟使正規化後的特徵參數其功率頻譜密度並不會趨近參考功率頻譜密度，只在輪廓上大致相同。最後一點，則是 TSN 法皆是根據 MVN 或 MVA 處理後的梅爾倒頻譜特徵所設計，進而得到良好的效能，我們希望能探討 TSN 法單純作用於未經任何處理的梅爾倒頻譜特徵時，其效果是否也一樣明顯。

根據以上對 TSN 法的分析與觀察，在本論文中，我們提出了三種語音特徵時間序列之調變頻譜正規化(modulation spectrum normalization)的新方法，分別爲等漣波時間序列濾波器法(equi-ripple temporal filtering, ERTF)、最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)與強度頻譜內插法(magnitude spectrum interpolation, MSI)，這三種方法之目的與 TSN 類似，皆爲了正規化語音特徵時間序列的功率頻譜密度，但我們會在後面章節的實驗結果發現，這三種方法之效能皆比 TSN 法來得好，且並不需要與 MVN 或 MVA 法結合，即可以十分有效地處理梅爾倒頻譜特徵因雜訊干擾所造成的失真。然而，當它們與 MVN 或 MVA 相結合時，也可以得到更佳的辨識精確率，此代表它們與 MVN 或 MVA 有良好的加成性。

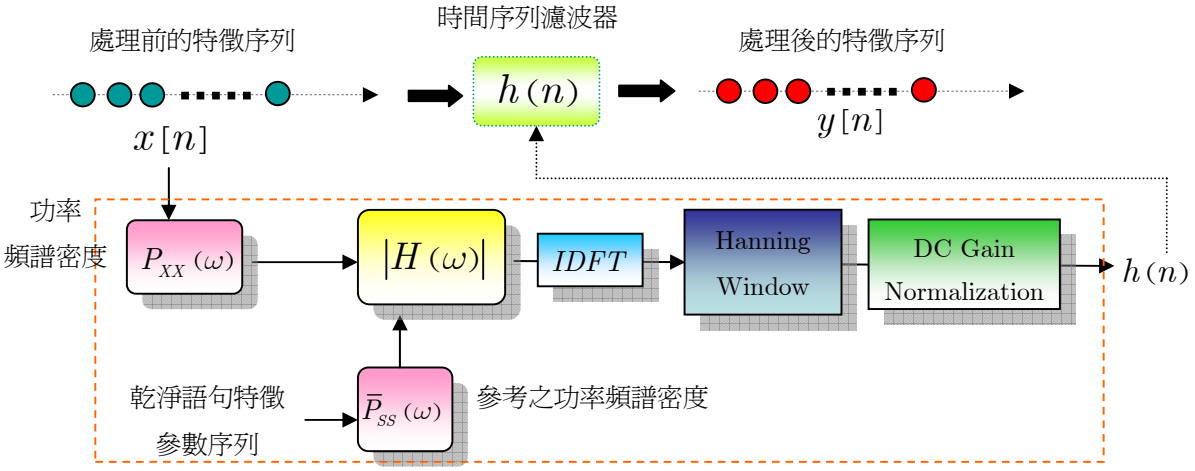
本論文其餘的章節概要如下：在第二章，我們將討論時間序列結構正規化法，包括其執行程序及初步效果，第三章爲本論文的重點，我們將在此章中針對時間序列結構正規化法作改進，而提出三種新的調變頻譜正規化法，並對其初步效果加以介紹。在第四章，我們將執行一系列的語音辨識實驗，來驗證所提之新方法足以有效提昇語音特徵在雜訊環境下的強健性，最後，第五章則爲結論及未來展望。

二、時間序列結構正規化法(temporal structure normalization, TSN)

（一）TSN 處理簡介

本章節主要介紹時間序列結構正規化法(temporal structure normalization, TSN)[8]，在下一章中，我們將以 TSN 法之觀念爲基礎，提出一系列的調變頻譜正規化的演算法。TSN 是屬於一種時間序列濾波器(temporal filter)設計之強健性語音技術，原始的 MFCC 語音特徵參數序列經過 CMVN 法[3]或 MVA 法[5]處理後，先求取其功率頻譜密度(power spectral density)，接著藉由此功率密度與事先定好的參考功率密度來決定一濾波器的強度響應(magnitude response)，此強度響應經反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)、漢寧窗化(Hanning window)處理與直流增益正規化處理後，產生一組

濾波器係數，此即爲 TSN 法所求得的時間序列濾波器，將語音特徵序列通過此濾波器後，則預期可達到調變頻譜正規化的效果，而增加語音特徵之其強健性。圖二爲 TSN 法的處理程序示意圖：



圖二、TSN 法處理程序示意圖

在 TSN 法中，每一句訓練語料之某一維特徵序列 $\{s[n]\}$ 與測試語料同一維特徵序列 $\{x[n]\}$，先求取其功率頻譜密度，分別以 $\{P_{SS}(\omega_k)\}$ 與 $\{P_{XX}(\omega_k)\}$ 表示。接著將訓練語料所有句子同一維的功率頻譜密度作平均，所得即爲參考功率頻譜密度，如下所示：

$$\bar{P}_{SS}(\omega_k) = E\{P_{SS}(\omega_k)\}, \qquad \text{(式 2.1)}$$

在 TSN 法中所使用的濾波器，其初始的強度頻譜設定如下式所示：

$$|H(\omega_k)| = \sqrt{\bar{P}_{SS}(\omega_k)/P_{XX}(\omega_k)}, \qquad \text{(式 2.2)}$$

其上式明顯看出，當任一測試語料 $x[n]$ 通過上式之濾波器時，其原始功率頻譜密度 $P_{XX}(\omega_k)$ 會被正規化爲 $\bar{P}_{SS}(\omega_k)$。

爲了進一步求取濾波器的脈衝響應(impulse response)，上式(2.2)中的 $|H(j\omega_k)|$ 先經過反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)，之後再乘上一個漢寧窗(Hanning window)，並將濾波器係數總和正規化爲 1，以達到直流增益正規化的目的。其數學表示式如以下數式所示：

1、反離散傅立葉轉換：

$$h[m] = \frac{1}{M}\sum_{k=0}^{M-1} H(j\omega_k)e^{-j\omega_k m}, \quad 0 \le m \le M-1. \qquad \text{(式 2.3)}$$

2、漢寧窗化處理：

$$\hat{h}[m] = h[m]\cdot w[m], \qquad \text{(式 2.4)}$$

其中

$$w[m] = 0.5\left(1-\cos\left(2\pi\frac{m}{M-1}\right)\right), \quad 0 \le m \le M-1.$$

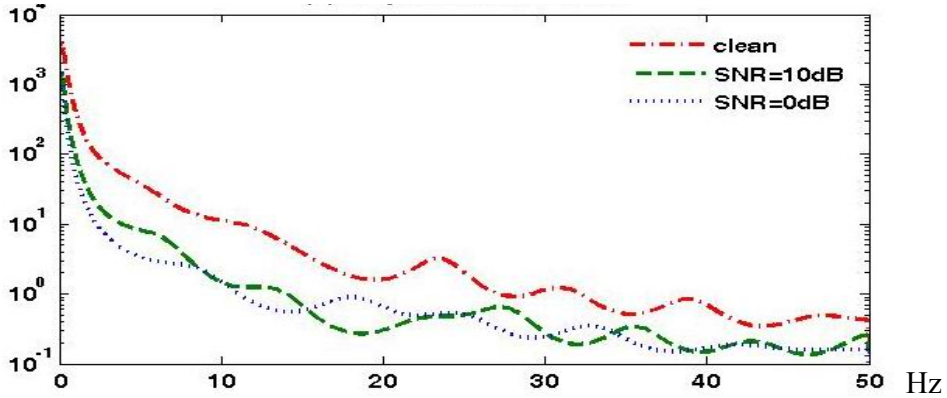
3、直流增益正規化：

$$\tilde{h}[m] = \frac{\hat{h}[m]}{\sum_{m'=0}^{M-1} h[m']}. \qquad \text{(式 2.5)}$$

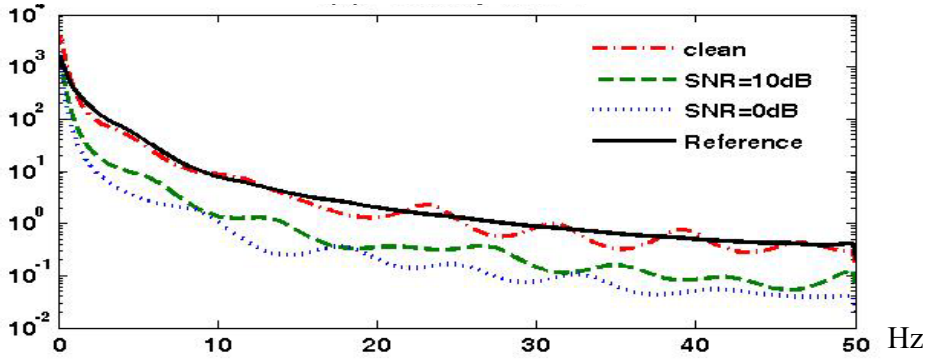
其中 M 爲濾波器長度。式(2.5)之 $\tilde{h}[m]$ 即爲 TSN 所求得之時間序列濾波器的脈衝響應。

（二）TSN 法效果相關討論

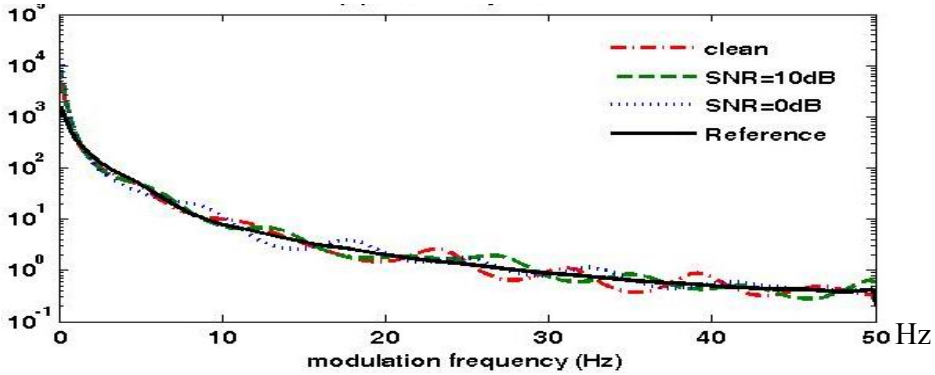
在 TSN 之文獻[8]中，所用的原始特徵參數皆爲經過 CMVN 法或 MVA 法所處理後之梅爾倒頻譜特徵參數(MFCC)。這裡我們特別將 TSN 法運用在未經處理之梅爾倒頻譜特徵參數上，觀察其改進效果。其中我們把原始 TSN 法命名爲 TSN-1，而把省略了直流增益正規化步驟的 TSN 法，命名爲 TSN-2。圖三爲原始第一維梅爾倒頻譜係數(c_1)序列的功率頻譜密度曲線圖，圖四爲原始 c_1 序列經 TSN-1 法處理後的功率頻譜密度曲線圖，圖五爲原始 c_1 序列經 TSN-2 法處理後的功率頻譜密度曲線圖。這些圖都使用了 AURORA 2 資料庫[9]裡的 MAH_4625A 語音檔，加入不同訊雜比的地下鐵雜訊。其中參考功率頻譜密度爲訓練語料庫之所有 c_1 序列之功率頻譜密度平均而得。



圖三、不同訊雜比之下，原始 c_1 序列之功率頻譜密度曲線圖



圖四、不同訊雜比之下，原始 c_1 序列經 TSN-1 處理後之功率頻譜密度曲線圖



圖五、不同訊雜比之下，原始 c_1 序列經 TSN-2 處理後之功率頻譜密度曲線圖

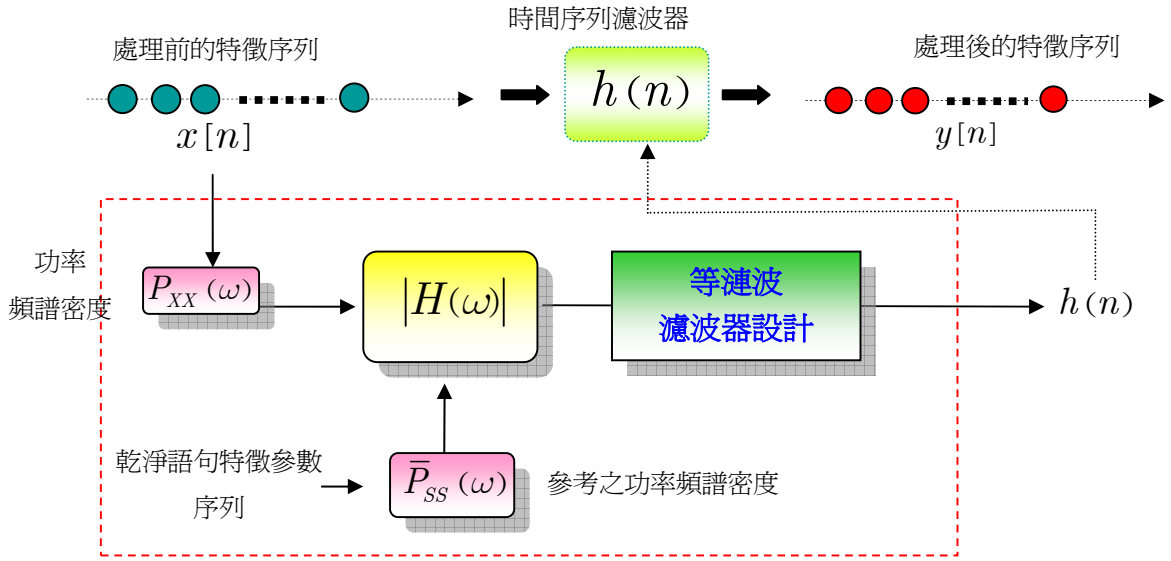
首先，從圖三可以明顯看出，雜訊會造成 c_1 特徵序列在功率頻譜密度上的失真，此是造成雜訊環境下，語音辨識精確率下降的原因之一。接著，我們從圖四觀察到原始 TSN 法（TSN-1）作用於原始 c_1 序列時，原本在圖三所看到之功率頻譜密度的失真並未被有效地改善，亦即其正規化效果並不理想，受到雜訊影響的 c_1 序列，當訊雜比(SNR)越低時，偏移參考功率頻譜密度的量越明顯。最後，從圖五可以看出，經過省卻直流增益正規化步驟的 TSN-2 法處理後，不同訊雜比下的 c_1 特徵序列其功率頻譜密度彼此十分接近，亦即 TSN-2 法可以有效正規化受雜訊干擾之原始 c_1 序列的功率頻譜密度，其降低失真的效能遠比 TSN-1 來的好。由此我們推論，原始 TSN 法中直流增益正規化的步驟並不是十分恰當，而其可能原因是，此步驟無法有效處理加成性雜訊對語音調變頻譜所造成的直流增益失真的效應。在下一章中，我們將提出一系列的方法，相較於 TSN 法而言，這些方法能更精確地正規化語音特徵的功率頻譜密度。

三、調變頻譜正規化的新方法

在第一章與第二章中，我們探討到時間序列結構正規化法(temporal structure normalization, TSN)可能有些可以改進的地方，同時藉由 TSN 法之觀念啓發，因此在本章節中，我們提出一系列的調變頻譜正規化的新方法。這些新方法分別爲等漣波時間序列濾波器法(equi-ripple temporal filtering, ERTF)、最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)與強度頻譜內插法(magnitude spectrum interpolation, MSI)，這些方法分別在本章的前三節中作介紹，而最後第四小節則爲這些方法簡要的效能評估與特性討論。

（一）等漣波時間序列濾波器法(equi-ripple temporal filtering, ERTF)

在等漣波時間序列濾波器法(ERTF)中，我們使用等漣波濾波器設計法 (equi-ripple filter design)[10]來設計濾波器的脈衝響應，以取代原始 TSN 法中，反傅立葉轉換與窗化處理的步驟，同時，我們也挪去原始 TSN 法中正規化濾波器直流增益的步驟。圖六爲等漣波時間序列濾波器法處理程序圖，在 ERTF 法中，我們所提出的兩更新步驟之目的正是求取更精確的濾波器係數，以趨近正規化特徵序列之調變頻譜強度成份的目標。



圖六、等連波時間序列濾波器法處理程序圖

ERTF 法中所使用的 $P_{XX}(\omega_k)$、$\bar{P}_{SS}(\omega_k)$ 和 $H(\omega_k)$ 求取方式都和前一章所述之原始

TSN 法相同，但是濾波器係數$\{h[n]\}$是以等漣波濾波器設計法[10]求得，此方法是利用所謂的最小化最大誤差準則(minimax criterion)來求取一最佳的濾波器頻率響應，如下式所示：

$$\breve{H}(\omega_k) = \arg\min_{H(\omega_k)} \left(\max_{\omega} W(\omega_k) \left| H(\omega_k) - D(\omega_k) \right| \right) , \qquad \text{(式 3.1)}$$

其中$W(\omega_k)$爲權重值，$\breve{H}(\omega_k)$爲最佳化濾波器之頻率響應，$D(\omega_k)$爲參考的頻率響應，$D(\omega_k)$可表示如下式：

$$D(\omega_k) = \sqrt{\frac{\overline{P}_{SS}(\omega_k)}{P_{XX}(\omega_k)}} \qquad \text{(式 3.2)}$$

由此法得到的濾波器係數$\{h[n]\}$，會自動符合前後對稱(symmetric)的性質，因此其相位響應是線性的(linear phase)[10]，並不會使原始特徵序列的調變頻譜產生相位失真的情形，同時，因爲濾波器本身是根據最佳化準則設計，所以我們預期它會比 TSN 法所得之濾波器效果來的好。

（二）最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)

在這方法裡，我們針對每一個待正規化的N點特徵時間序列$\{x[n] | 0 \le n \le N-1\}$先定義一$2P$點的參考調變頻譜，作爲此特徵序列的調變頻譜正規化的目標：

$$\hat{Y}(\omega_k) = \left| Y(\omega_k) \right| \exp\left(j\theta_X(\omega_k) \right), \quad 0 \le k \le 2P-1, \qquad \text{(式 3.3)}$$

其中的強度成份$\left| Y(\omega_k) \right|$以下式表示：

$$\left| Y(\omega_k) \right| = \left| X(\omega_k) \right| \sqrt{\overline{P}_{SS}(\omega_k) / P_{XX}(\omega_k)} \qquad \text{(式 3.4)}$$

其中，$\overline{P}_{SS}(\omega_k)$與如前章的式(2.1)中所定義，即$\overline{P}_{SS}(\omega_k)$爲所有訓練語料特徵與$\{x[n]\}$同一維序列的功率頻譜密度平均而得，$P_{XX}(\omega_k)$爲原始特徵序列$\{x[n]\}$的功率頻譜密度。而強度成份$\left| X(\omega_k) \right|$和相角成份$\theta_X(\omega_k)$爲$\{x[n]\}$經過$2P$點之離散傅立葉轉換(discrete Fourier transform, DFT)所得到。值得注意的是，特徵長度N會隨著不同的語句而不同，但是這裡的 DFT 取樣點數$2P$則設爲一固定值，也就是參考調變頻譜的長度對於每一個語句都是相同的。

由式(3.3)與式(3.4)可知，我們希望每一個更新後的特徵序列，其調變頻譜的強度成份能趨於一致，而相位成份則由原始的特徵序列$\{x[n]\}$而來。接下來，我們利用最小平方化(least-squares)[10]的最佳化準則求取一新的特徵參數序列，使新的特徵序列$\{y[n]\}$的調變頻譜逼近如式(3.3)的參考調變頻譜，如下式所示：

$$y[n] = \min_{\{\tilde{y}[n] | 0 \le n \le N-1\}} \sum_{k=0}^{2P-1} \left| \sum_{n=0}^{N-1} \tilde{y}[n] e^{-j\frac{2\pi nk}{2P}} - \hat{Y}(\omega_k) \right|^2, \qquad (2P \ge N) \qquad \text{(式 3.5)}$$

其中$2P$爲 DFT 取樣點數，N爲此特徵序列的點數。

藉由矩陣與向量表示法，我們可將式(3.5)改寫爲下式：

$$\mathbf{y} = \min_{\tilde{\mathbf{y}}} \left\| W\tilde{\mathbf{y}} - \hat{\mathbf{Y}} \right\|^2 \qquad \text{(式 3.6)}$$

其中W是$2P \times N$的矩陣，其第(m,n)項如下所示：

$$W_{mn} = \exp\left(-j\frac{2\pi mn}{2P} \right),$$

而$\mathbf{y}$、$\tilde{\mathbf{y}}$與$\hat{\mathbf{Y}}$則定義爲：

$$\mathbf{y} = \begin{bmatrix} y[0] & y[1] & \cdots & y[n-1] \end{bmatrix}^T ,$$

$$\widehat{\mathbf{y}} = \left[\widehat{y}[0] \quad \widehat{y}[1] \quad \cdots \quad \widehat{y}[N-1]\right]^T,$$
$$\widehat{\mathbf{Y}} = \left[\widehat{Y}(\omega_0) \quad \widehat{Y}(\omega_1) \quad \cdots \quad \widehat{Y}(\omega_{2P-1})\right]^T,$$

由於$\widehat{\mathbf{y}}$爲實數向量，故式(3.6)可改寫爲：

$$\begin{aligned}\mathbf{y} &= \min_{\tilde{\mathbf{y}}} \left\| \left(W_R\widehat{\mathbf{y}} - \widehat{\mathbf{Y}}_R\right) + j\left(W_I\widehat{\mathbf{y}} - \widehat{\mathbf{Y}}_I\right)\right\|^2 \\ &= \min_{\tilde{\mathbf{y}}} \left(\left\|W_R\widehat{\mathbf{y}} - \widehat{\mathbf{Y}}_R\right\|^2 + \left\|W_I\widehat{\mathbf{y}} - \widehat{\mathbf{Y}}_I\right\|^2 \right)\end{aligned} \qquad \text{(式 3.7)}$$

其中矩陣W_R與W_I分別爲矩陣W的實部與虛部，而向量$\widehat{\mathbf{Y}}_R$與$\widehat{\mathbf{Y}}_I$則分別爲向量$\widehat{\mathbf{Y}}$的實部與虛部。

由式(3.7)明顯看出，此爲一典型的最小平方化(least-squares)的求解問題，故其精確的封閉解(closed-form solution)可由下式表示：

$$\mathbf{y} = \left(W_R^T W_R + W_I^T W_I\right)^{-1}\left(W_R^T\widehat{\mathbf{Y}}_R + W_I^T\widehat{\mathbf{Y}}_I\right) \qquad \text{(式3.8)}$$

所以，式(3.8)中的$\mathbf{y}$即爲 LSSF 法所求得之新特徵參數序列$\{y[n]\}$，其$2P$點之 DFT 和式(3.3)的參考調變頻譜之間具有最小平方誤差的良好性質。

（三）強度頻譜內插法(magnitude spectrum interpolation, MSI)

在此方法中，我們爲每一個待正規化的N點特徵序列$\{x[n] | 0 \le n \le N-1\}$，定義了一個$N$點的參考調變頻譜，作爲此特徵序列之調變頻譜正規化的目標，如下式所示：

$$\tilde{Y}(\omega_{k'}) = \left|\tilde{Y}(\omega_{k'})\right| \exp\left(j\theta_X(\omega_{k'})\right), \quad 0 \le k' \le N-1 \qquad \text{(式 3.9)}$$

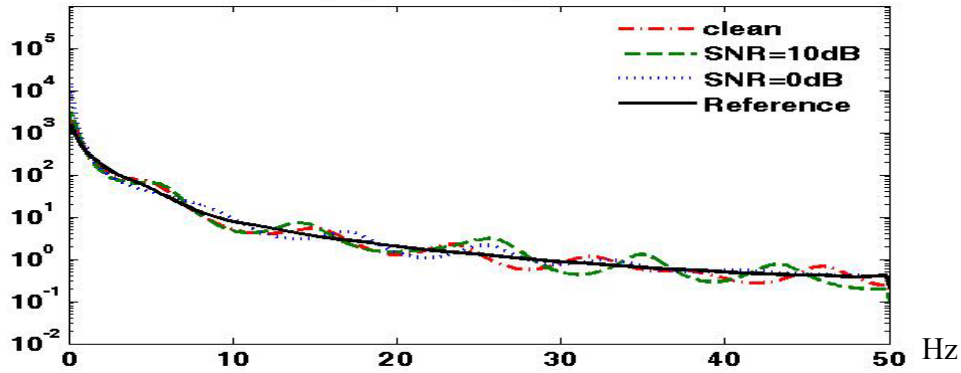
其中相位成份$\theta_X(\omega_{k'})$爲$x[n]$取N點的 DFT 所得。MSI 法跟前節之 LSSF 法的最大不同之處，在於此時我們是使用一個跟原始特徵序列長度相同的參考調變頻譜，而由於不同語句的特徵序列，其點數N也隨之不同，我們不能如前面的 LSSF 法中，直接拿$2P$點的參考功率頻譜密度$\left\{\overline{P}_{SS}(\omega_k) \middle| 0 \le k \le 2P-1\right\}$(如式(2.1)所示)來求取式(3.9)中的$N$點頻譜強度$\left|\tilde{Y}(\omega_{k'})\right|$。然而，由於原始$2P$點的參考頻譜其涵蓋頻率範圍與欲求的$\left|\tilde{Y}(\omega_{k'})\right|$頻率範圍相同，在這裡，我們使用線性內插(linear interpolation)[10]的方法，藉由式(3.4)中所示的之$2P$點的$\left\{\left|Y(\omega_k)\right| \middle| 0 \le k \le 2P-1\right\}$來求取式(3.9)中$N$點的$\left\{\left|\tilde{Y}(\omega_{k'})\right| \middle| 0 \le k' \le N-1\right\}$之近似值。但是式(3.9)的$\left\{\tilde{Y}(\omega_{k'})\right\}$爲一實數序列之離散傅立葉轉換，其強度成份$\left\{\left|\tilde{Y}(\omega_{k'})\right|\right\}$必須符合左右對稱的性質，即$\left|\tilde{Y}(\omega_{k'})\right| = \left|\tilde{Y}(\omega_{N-k'})\right|$，因此我們先利用$\left\{\left|Y(\omega_k)\right|\right\}$的左半部執行內插法，求取$\left\{\left|\tilde{Y}(\omega_{k'})\right|\right\}$的左半部$\left\{\left|\tilde{Y}(\omega_{k'})\right| \middle| 0 \le k' \le \left\lfloor\frac{N}{2}\right\rfloor\right\}$，再利用左右對稱的性質，求取$\left\{\left|\tilde{Y}(\omega_{k'})\right|\right\}$右半部$\left\{\left|\tilde{Y}(\omega_{k'})\right| \middle| N-1-\left\lfloor\frac{N}{2}\right\rfloor \le k' \le N-1\right\}$。在得到$\left\{\left|\tilde{Y}(\omega_{k'})\right| \middle| 0 \le k' \le N-1\right\}$後，我們就可以直接對式(3.9)的$\left\{\tilde{Y}(\omega_{k'})\right\}$做$N$點的反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)，以求得新的特徵序列$\{y[n]\}$，如下式所示：

$$y[n] = \frac{1}{N}\sum_{k'=0}^{N-1}\tilde{Y}(\omega_{k'})e^{j\frac{2\pi nk'}{N}}, \quad 0 \le n \le N-1. \qquad \text{式(3.10)}$$

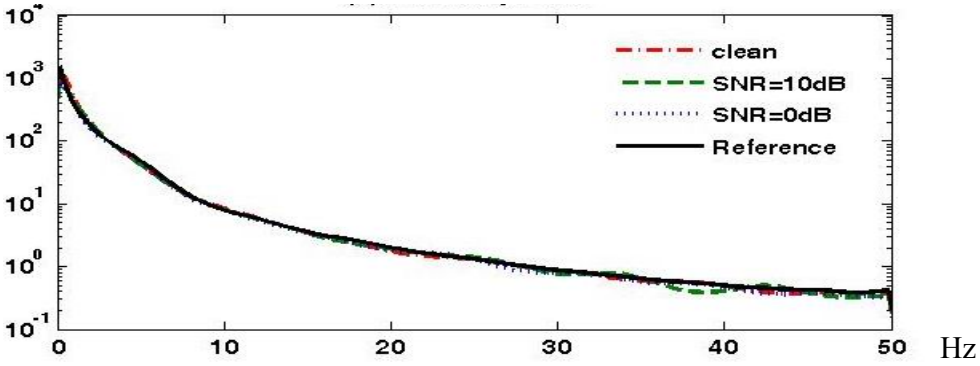
以上的方法，即稱爲強度頻譜內插法(magnitude spectrum interpolation, MSI)。

（四）調變頻譜正規化之新方法的效果討論

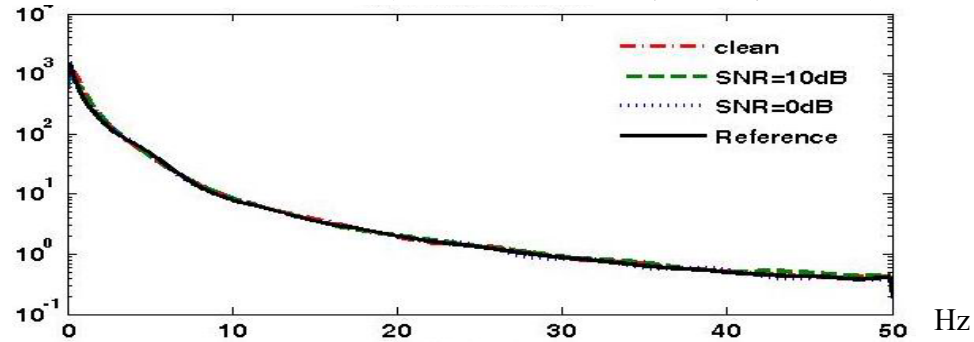
這小節將簡單展示本章節所提出的三種新方法對原始 MFCC 特徵序列之調變頻譜正規化的效果，圖七、圖八與圖九分別爲原始第一維梅爾倒頻譜係數(c_1)序列分別經 ERTF 法、LSSF 法與 MSI 法處理後的功率頻譜密度曲線圖。與前一章的圖三、圖四和圖五相同，這裡我們所使用的是 AURORA 2 資料庫[9]裡的 MAH_4625A 語音檔，然後加入不同訊雜比(SNR)的地下鐵(subway)雜訊。



圖七、不同訊雜比之下，原始 c_1 序列經 ERTF 法處理後之功率頻譜密度曲線圖



圖八、不同訊雜比之下，原始 c_1 序列經 LSSF 法處理後之功率頻譜密度曲線圖



圖九、不同訊雜比之下，原始 c_1 序列經 MSI 法處理後之功率頻譜密度曲線圖

將圖七、圖八、與圖九配合前一章之圖三、圖四與圖五相比較，我們有以下兩點討論：

① 由於 ERTF 法和 TSN 法同樣是設計一時間序列濾波器，作用於特徵參數序列上，我們先比較這兩種方法的效能。從圖七中可看出 ERTF 法能同時使得乾淨語音與受雜訊干擾的語音的功率頻譜密度曲線，逼近參考的功率頻譜密度曲線，有效降低圖三所顯示之不同訊雜比下特徵序列之功率頻譜密度的失真，相較於圖四所顯示之原始 TSN 法的效果有明顯改善，且與圖五之 TSN-2 法的效果十分接近，此代表我們使用等漣波濾波器設計法 (equi-ripple filter design)來設計時間序列濾波器，可以有效地正規化不同

雜訊比下的語音特徵之調變頻譜。

② LSSF 法和 MSI 法都是直接在特徵的調變頻譜域(modulation spectral domain)上正規化其強度成份，從圖八和圖九可看出這兩種方法與 ERTF 法類似，能將受雜訊干擾的語音之功率頻譜密度曲線，逼近參考的功率頻譜密度曲線，使這些曲線之間的差異明顯較低，代表了這兩個調變頻譜強度正規化法也能有效地強健語音特徵。其中，MSI 法是三個方法中計算複雜度最低的技術，因此有更大的應用價值。

從上述三個圖中，可看出我們所提的三個新方法都能有效地降低雜訊所造成之語音特徵在調變頻譜上失真的現象，我們在下個章節，將會以辨識實驗數據證實這些方法的效能。

四、調變頻譜正規化法與各種特徵時間序列正規化技術法之辨識實驗結果與討論

本章節主要是將我們提出的三種調變頻譜正規化法：等漣波時間序列濾波器(equi-ripple temporal filtering, ERTF)、最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)和強度頻譜內插法(magnitude spectrum interpolation, MSI)運用於雜訊環境下的語音辨識，藉此觀察分析其結果，同時我們也會將它們與其他特徵時間序列正規化法的效果作比較。最後，我們嘗試將這些新方法與其他方法互相結合，來觀察這樣的結合是否能來更進一步的效能提升。

（一）實驗環境與實驗架構設定

本論文中所採用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行之語料庫：AURORA 2.0[9]，內容是以美國成年男女所錄製的一系列連續的英文數字字串，語音本身並加上各種加成性雜訊與通道效應的干擾。加成性雜訊共有八種，分別爲地下鐵、人聲、汽車，展覽會館、餐廳、街道、飛機場和火車站雜訊等，通道效應則有兩種，分別爲 G712 與 MIRS[11]。雜訊含量的大小包含了乾淨無雜訊的狀態，以及六種不同訊雜比(signal-to-noise ratio, SNR)狀態，分別是 20dB、15dB、10dB、5dB、0dB 與-5dB，因此我們可以觀察不同的雜訊環境對於語音辨識的影響。因雜訊特性的不同，測試環境可分爲 Set A、Set B 與 Set C 三組[9]。

聲學模型是執行隱藏式馬可夫模型工具(hidden Markov model tool kit, HTK)[12]訓練所得，包含 11 個數字模型(zero, one, two, …, nine 及 oh)以及靜音(silence)模型，每個數字模型包含 16 個狀態，各狀態包含 20 個高斯密度混合。

（二）調變頻譜正規化法作用於梅爾倒頻譜特徵參數之實驗結果

本章節所有實驗所使用的語音特徵爲梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)，我們採用的 MFCC 特徵參數爲13維(c0~c12)，加上其一階差量(delta)和二階差量(delta-delta)，總共爲39維特徵參數。基本實驗(baseline experiment)是以原始 MFCC 特徵參數作爲訓練與測試，TSN-1法爲第二章中所介紹之原始 TSN 法，而 TSN-2 法則是將原始 TSN 法中直流增益正規化步驟省略所得的修正法，TSN-1與 TSN-2所得之時間序列濾波器長度皆設爲21，此值是直接參考 TSN 法的文獻[8]而來。ERTF 法所得的時間序列濾波器長度爲21，而 LSSF 與 MSI 法所用的 DFT 點數2^P (如式(3.3)所示)則固定爲1024。下表一中，我們綜合了 TSN-1、TSN-2、ERTF、LSSF、MSI，及著名的特徵正規化技術 CMVN[3]和 MVA[5]，其各別作用於原始 MFCC 特徵參數所得的平均辨識率（20dB、15dB、10dB、5dB 與0dB 五種訊雜比下的辨識率平均），其中 AR 與 RR 分別爲相較於基本實驗結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。

由表一的數據，我們可看出以下幾點現象：

表一、各種特徵序列處理技術之辨識率(%)

Method	Set A	Set B	Set C	average	AR	RR
Baseline	72.46	68.31	78.82	73.20	-	-
TSN-1	73.61	70.44	77.19	73.75	0.55	2.05
TSN-2	80.29	82.36	75.82	79.49	6.29	23.47
ERTF	85.45	86.92	85.34	85.90	12.70	47.39
LSSF	84.37	86.21	84.72	85.10	11.90	44.40
MSI	83.61	85.36	84.28	84.42	11.22	41.87
CMVN	85.03	85.56	85.60	85.40	12.20	45.52
CMVN+ARMA(MVA)	88.12	88.81	88.50	88.48	15.28	57.01

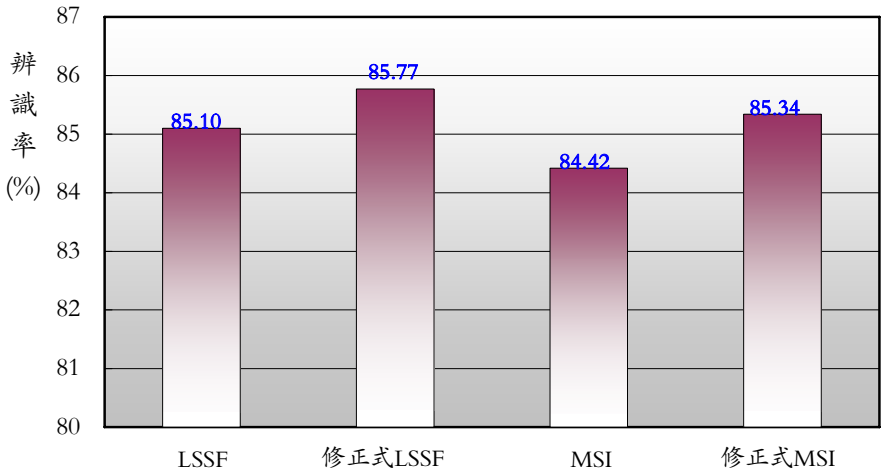
①原始 TSN 法(TSN-1)對 MFCC 特徵在雜訊環境下的辨識率的改進並不是很明顯，只進步0.55%，然而 TSN-2法帶來十分明顯的辨識率提升(Set C 除外)，在 Set A 和 Set B 環境下，平均辨識率相對於 TSN-1而言分別改進了8%與14%左右。如此看出，藉由省略直流增益正規化的步驟，TSN-2 比 TSN-1 具有更佳的特徵調變頻譜正規化的效果，這也呼應了在第二章的圖三，原始 TSN 法無法有效降低各雜訊環境下，原始語音 MFCC 特徵時間序列之功率頻譜密度曲線的不匹配現象。

②ERTF、LSSF 與 MSI 法三種新方法在各種不同的雜訊環境下皆能明顯提升辨識率，對 Set A 環境而言，它們分別使辨識率提升了 12.99%、11.91%與 11.15%，對 Set B 環境而言，辨識率分別提升了 18.61%、17.90%與 17.05%，在 Set C 環境下，辨識率分別提升了 6.52%、5.90%與 5.46%。這三種方法中，又以 ERTF 法的表現最好，明顯優於 LSSF 法與 MSI 法，但它們所能達到的相對錯誤降低率都高達 40%以上，明顯優於 TSN-1 法與 TSN-2 法。另外，值得一提的是，TSN-2 法在 Set C 中的效果比 TSN-1 與基礎實驗差，但 ERTF、LSSF 與 MSI 法卻未有這樣的不良結果。

③ 兩種目前廣爲人用的特徵正規化技術：CMVN 法與 MVA 法，對辨識率的提升都十分明顯，CMVN 的效能與我們所提的三種新方法大致相同，但結合了 CMVN 與 ARMA 濾波處理的 MVA 法其效能又比 CMVN 法來的好，基於這樣的觀察，在下兩小節中，我們將試著把各種調變頻譜正規化法與 CMVN 法或 MVA 法加以整合，探討是否能帶來辨識率上更顯著的進步。

當我們使用 LSSF 法與 MSI 法時，我們會將原始爲 N 點的特徵序列轉換成 $2P$ 點之功率頻譜密度或離散頻譜，然而由於通常 $2P > N$ ，我們會以補零的方式先將原始的 N 點的特徵序列變長爲 $2P$ 點，意即多補了 $2P - N$ 個零點，這樣的作法容易產生非零值的點與零值的點之間訊號值不連續的情形，而引進了不必要的高頻成份，這效應類似於直接於一訊號加上矩形窗所造成頻譜遺漏(leakage)[10]的缺點，因此，我們這裡在 LSSF 與 MSI 法之補零的程序前，先將原始的 N 點的特徵序列乘上一漢寧窗(Hanning window)[10]，來降低上述可能的不良效應，觀察這樣的操作是否可進一步提升 LSSF 法與 MSI 法的效果，我們稱這樣修改結果分別爲修正式 LSSF 法(modified LSSF)與修正式 MSI 法(modified MSI)。

圖十爲原始與修正式 LSSF 與 MSI 作用於原始 MFCC 特徵之平均辨識率長條圖。由此圖中可以看出修正式 LSSF 法相較於原始 LSSF 法而言，平均辨識率有 0.67%的提升，而修正式 MSI 相較於原始 MSI 而言，在平均辨識率上有 0.92%的提升。由此我們驗證了，在修正法中所作的窗化處理確實能有效改進 LSSF 法與 MSI 法的效能。



圖十、原始和修正式 LSSF 與 MSI 作用於原始 MFCC 特徵之平均辨識率

（三）調變頻譜正規化法結合倒頻譜平均與變異數正規化法之實驗結果

前面提到，倒頻譜平均與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]對雜訊環境下的語音辨識率有明顯的改進，因此這裡我們嘗試將各種調變頻譜正規化法與 CMVN 法作結合，意即原始 MFCC 特徵先經過 CMVN 法處理後，再以各種調變頻譜正規化法分別作處理。以下我們測試這樣的結合是否有加成性的效果。在表二中，我們整理了 CMVN 法分別結合 TSN-1、TSN-2、ERTF、LSSF、MSI 及 ARMA 濾波法(MVA)[5]各方法所得的平均辨識率，其中 AR 與 RR 分別爲相較於單一 CMVN 結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。

表二、各調變頻譜處理法作用於 CMVN 處理後之 MFCC 特徵所得之辨識率(%)

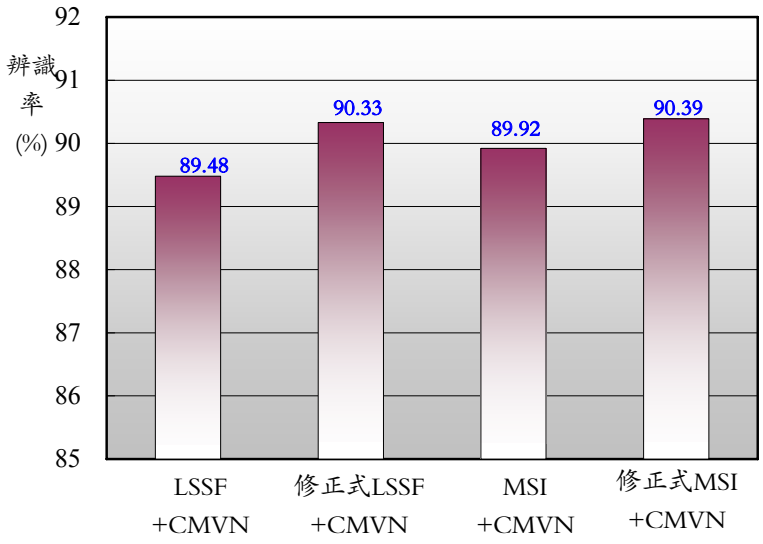
Method	Set A	Set B	Set C	average	AR	RR
CMVN	85.03	85.56	85.60	85.40	—	—
CMVN+TSN-1	89.42	90.03	89.03	89.49	4.10	28.05
CMVN+TSN-2	89.59	90.36	89.34	89.76	4.36	29.90
CMVN+ERTF	89.61	90.67	89.28	89.85	4.45	30.52
CMVN+LSSF	89.12	90.17	89.16	89.48	4.09	27.98
CMVN+MSI	89.59	90.56	89.60	89.92	4.52	30.95
CMVN+ARMA(MVA)	88.12	88.81	88.50	88.48	3.08	21.09

由表二的數據，我們可看出以下幾點現象：

① TSN-1 法作用於 CMVN 處理過的 MFCC 特徵，其改進辨識率的效能十分顯著，相較於單一 CMVN 法而言，在 Set A、Set B 與 Set C 環境下分別具有 4.39%、4.47%與 3.43%的辨識率改善，此結果十分吻合在 TSN 法的原始文獻[8]裡之結果，相較於表一所呈現之 TSN-1 並未明顯改善受雜訊影響之原始 MFCC 特徵的現象，在這裡，TSN-1 法能有明顯改進之效能的原因可能在於，CMVN 法已事先有效地降低原始 MFCC 特徵受雜訊影響所造成之調變頻譜上下偏移的失真，因此 TSN-1 能單純處理調變頻譜正規化的部份，而帶來辨識率的改善。另外，我們也發現到，TSN-1 和 TSN-2 所得結果之間的差距變得較小，但 TSN-2 的整體辨識率還是比 TSN-1 來的好，再一次驗證原 TSN 法中直

流增益正規化的步驟應該是不必要的。
②當我們所提出的新方法 ERTF、LSSF 與 MSI 法作用於 CMVN 處理後之 MFCC 特徵時，相較於單一 CMVN 所得的辨識率而言，皆帶來十分顯著的改善，例如在 Set A 環境下，這三種方法分別具有 4.58%、4.09%與 4.56%的辨識率提昇，此顯示了這三種新方法與 CMVN 有良好的加成性。而三個新方法中，ERTF 和 MSI 法表現的都比 TSN-1 或 TSN-2 法更好，雖然 LSSF 法表現稍不如預期，但是可能原因在於前一節所討論到的，原始 LSSF 法和 MSI 法可能會產生頻譜遺漏(leakage)現象之類的不良效應，因此在後面，我們將以修正式 LSSF 法與 MSI 法結合 CMVN 法來探討其可能的改進效果。
③在之前的表一數據顯示，當作用於原始 MFCC 特徵時，ERTF 表現比 LSSF 和 MSI 法更好。但是在這裡我們發現當這些方法和 CMVN 法結合時，其效果變得十分接近，這也意味著 CMVN 法確實已對原始 MFCC 特徵作了十分有效的強健性處理，而使後續的改進技術，其進步的空間相對變小。

如之前所提到的，原始 LSSF 法和 MSI 法可能有頻譜遺漏(leakage)的缺點，因此我們這裡使用之前所述的修正式 LSSF 與 MSI 法，作用於 CMVN 處理後的 MFCC 特徵，即在此兩方法補零的程序前先將原始 N 點的 CMVN 法處理後之 MFCC 特徵序列乘上一漢寧窗(Hanning window)，觀察這樣的操作是否可進一步提升原始 LSSF 法與 MSI 法結合 CMVN 法的效果。
圖十一為原始與修正式 LSSF 與 MSI 作用於 CMVN 法處理後 MFCC 特徵之平均辨識率長條圖。由此圖可以看出，在結合 CMVN 法後，修正式 LSSF 法相較於原始 LSSF 法而言，有 0.85%之平均辨識率的提升，同樣地，修正式 MSI 法相對於原始 MSI 法而言，有 0.47%之平均辨識率的提升，二者平均辨識率皆超過 90%。此外，當與表二的數據比較，我們看到這兩種修正式方法結合 CMVN 法後在總平均辨識率上皆明顯優於與 CMVN 法結合的 TSN-1 法(89.49%)與 TSN-2 法(89.76%)，以上結果都顯示了這樣的修正確實能有效改進原方法的缺點，而提升其效能。



圖十一、原始和修正式 LSSF 與 MSI 作用於 CMVN 法處理後 MFCC 特徵之平均辨識率

（四）調變頻譜正規化法結合倒頻譜平均與變異數正規化結合自動回歸動態平均濾波器法之實驗結果

前面提到，倒頻譜平均與變異數正規化結合自動回歸動態平均濾波器法(MVA)[5]

能夠對雜訊環境下的語音特徵有明顯的強健化效果，而帶來十分顯著的辨識率提升，且其效能優於 CMVN，因此在這裡，我們將各種調變頻譜正規化法與 MVA 法作結合，也就是把這些正規化法作用於經 MVA 法處理後之 MFCC 特徵上，以檢視這些正規化法與 MVA 法是否有加成性。實驗中我們設定 MVA 法中的 ARMA 濾波器階數爲 2(參照[5])。在下表三中，我們列出了 MVA 法分別結合 TSN-1、TSN-2、ERTF、LSSF 與 MSI 各方法所得的平均辨識率，其中 AR 與 RR 分別爲相較於單一 MVA 法之結果的絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。

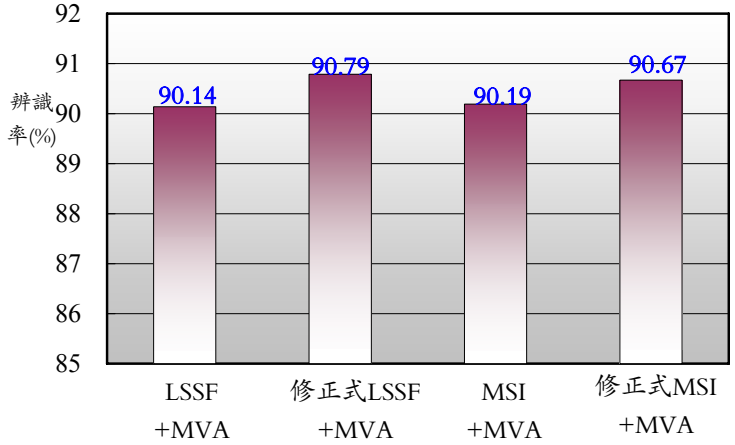
表三、各調變頻譜處理法作用於 MVA 處理後之 MFCC 特徵所得之辨識率(%)

Method	Set A	Set B	Set C	average	AR	RR
MVA	88.12	88.81	88.50	88.48	—	—
MVA+TSN-1	89.58	90.19	89.74	89.84	1.36	11.80
MVA+TSN-2	89.81	90.34	89.84	99.00	1.52	13.19
MVA+ERTF	89.75	90.81	89.64	90.07	1.59	13.80
MVA+LSSF	89.63	90.87	89.94	90.14	1.67	14.49
MVA+MSI	89.71	90.91	89.94	90.19	1.71	14.84

由表三可看出，TSN-1 在結合 MVA 後，其效能有明顯的提昇，而 TSN-1 和 TSN-2 之間的差異雖然不明顯，但是省略直流增益正規化步驟的 TSN-2 法仍然表現比較好，相對於單一 MVA 法的結果而言，結合了 MVA 法之 TSN-1 在總平均辨識率上提升 1.36%，而 TSN-2 提升了 1.52%。此外，結合 MVA 法之後，我們提出的 ERTF、LSSF 與 MSI 三個方法仍優於 TSN-1 與 TSN-2，而其中以 MSI 法最好，在辨識率上提升 1.71%，其次爲 LSSF 法，提升了 1.67%， ERTF 法則提升了 1.59%。儘管如此，我們可明顯看出，這些方法在結合 MVA 法後，所帶來的辨識率提升程度相對而言都已十分接近。

如同前節所描述之原始 LSSF 法與 MSI 法的可能缺點，在這裡，我們同樣地測試修正式 LSSF 法與 MSI 法結合 MVA 法的效果，即在原始 LSSF 法或 MSI 法之補零的程序前先將原始 N 點之 MVA 法處理後之 MFCC 特徵序列乘上一漢寧窗(Hanning window)，觀察這樣的操作能否帶來進步。

圖十二爲原始與修正式 LSSF 與 MSI 作用於 MVA 法處理後 MFCC 特徵之平均辨識率長條圖。由此圖可以看出，在結合 MVA 法的前提下，修正式 LSSF 法相較於原始 LSSF 法而言，有 0.65%平均辨識率的提升，而修正式 MSI 法相對於原始 MSI 法而言，有 0.48%平均辨識率的提升，因此，我們驗證了兩種修正式方法都能使原始方法進一步提升效能。



圖十二、原始和修正式 LSSF 與 MSI 法作用於 MVA 法處理後 MFCC 特徵之平均辨識率

五、結論

在作用於原始 MFCC 特徵時，我們發現，原始 TSN 法(TSN-1)的直流增益正規化步驟是造成其效果不彰的原因之一，挪去此步驟所得之 TSN-2 法即可有十分顯著的表現，而我們提出的三種新方法，相較於 TSN-1 與 TSN-2，都能有更佳的效果，而其中又以 ERTF 法之表現最好，由於 ERTF 與 TSN-2 只有在設計時間序列濾波器的程序上有差別，這表示我們 ERTF 設計出來的濾波器，比起 TSN-2 法的濾波器更精確地對特徵之調變頻譜作正規化。而當我們將這些方法作用於 CMVN 法或 MVA 法處理後的 MFCC 特徵時，發現它們相較於單一 CMVN 法或 MVA 法而言，能帶來更佳的辨識率，且我們所提出之三種新方法的表現幾乎仍然優於 TSN-1 法與 TSN-2 法。此外，我們探討 LSSF 法與 MSI 法可能存在之頻譜遺漏(leakage)的缺點，而提出相對應的修正方法，發現這些修正法能更進一步改善原始 LSSF 法與 MSI 法的效能。

若就三種新方法彼此作比較，ERTF 法與 LSSF 法運算複雜度較大，MSI 法則相對較小，雖然 ERTF 法對原始 MFCC 特徵而言，表現比 LSSF 法與 MSI 法來得好，但當它們作用於 CMVN 法或 MVA 法處理過後的 MFCC 特徵時，其效能的差異性已經很小，這意味著運算複雜度較小的 MSI 法相對於 ERTF 法與 LSSF 法而言，可能有更佳的應用性。

六、參考文獻

[1] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language,1995

[2] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on Acoustics, Speech and Signal Processing, 1981

[3] S. Tiberewala and H. Hermansky, "Multiband and Adaptation Approaches to Robust Speech Recognition", 1997 European Conference on Speech Communication and Technology (Eurospeech 1997)

[4] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Trans. on Speech and Audio Processing, 1994

[5] C-P. Chen and J-A. Bilmes, "MVA Processing of Speech Features", IEEE Trans. on Audio, Speech, and Language Processing, 2006

[6] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, "Cepstral Gain Normalization for Noise Robust Speech Recognition", 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)

[7] J-W. Hung and L-S. Lee, "Optimization of Temporal Filters for Constructing Robust Features in Speech Recognition", IEEE Trans. on Audio, Speech and Language Processing, 2006

[8] X. Xiao, E-S. Chng, and Haizhou Li, "Temporal Structure Normalization of Speech Feature for Robust Speech Recognition", IEEE Signal Processing Letters, vol. 14, 2007

[9] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", Proceedings of ISCA IIWR ASR2000, Paris, France, 2000

[10] S. K. Mitra, "Digital Signal Processing", The McGraw-Hill International, 3rd edition, 2006

[11] ITU recommendation G.712, "Transmission Performance Characteristics of Pulse Code Modulation Channels," Nov. 1996

[12] http://htk.eng.cam.ac.uk/

形音相近的易混淆漢字的搜尋與應用

劉昭麟　　黃志斌　　翁睿妤　　莊怡軒
國立政治大學 資訊科學系
{chaolin, g9614, s9403, s9436}@cs.nccu.edu.tw

摘要

在中文裡面，漢字包含因爲發音相近或者形體相似的易混淆字，這一些易混淆字對於電腦輔助教學和語言心理學的相關研究具有相當意義。我們運用倉頡碼的設計理念和電子詞典所提供的發音資訊，配合網際網路可以得到的文字資訊，設計一個不須仰賴影像處理技術，就可以找到形音相近漢字的方法。經過實驗證明，以提交五個甚至一個建議字爲限，我們的方法所建議的形音相近字集，能夠包含一般與專業受試者所提供的常見錯別字集。

關鍵詞：漢字研究、漢字搜尋、漢字構字資訊、電腦輔助語言教學、語文認知

1. 簡介

個別漢字是構成中文的基本單位，有自己的發音、筆畫構造與所攜帶的意涵；透過個別漢字所組成的單字詞、雙字詞等詞彙，依據漢語語法組成中文句子。因此，學習漢字雖然不是學習漢語會話的必要工作，但卻是進階中文學習者一個重要的功課。同時，語言使用者如何透過語言的聲音(pronunciation)和文字的形體(grapheme)來擷取語意，更是研究語言認知歷程的學者所專注的重要議題。因此，本論文探討如何利用軟體技術找尋因爲發音和形體近似而容易混淆的漢字，以供電腦輔助教學和認知語言學的研究之用。

中文句子「今天上午我們來試場買菜」包含一個典型的錯誤；試場雖然是一個存在的詞彙，代表考試的場所，但是除非情境特殊，否則在這一例句裡面的「試場」應改爲「市場」。「經理要我構買一部計算機」這個句子也有一個錯誤：「構買」應改爲「購買」。雖然在簡體字的環境中比較多的人會寫「構買」，但是在繁體中文的使用群中，也有人把「購買」寫成「構買」。

因爲形音近似而誤用詞彙並不是中文所特有的現象，英文也有類似的問題[4]。舉例來說，“John plays an important roll in this event.” 包含一個錯誤的字；“roll” 應改爲“role”。其他像下列這一些字組，都是易混淆字的範例，principle和principal、teen和team、there和their、leak和leek、wait和weight、knows和nose以及knit和nit等等。

形音近似的漢字常被用於國民小學國語科試題的「改錯字」試題[6]。教師把一句正確的中文句子其中一個字改成另一個具有相當吸引力的錯字，以這一句帶有錯字的中文當作試題，要求受測學生找出並且更正這一錯字。這一類的試題也可以變形爲中文的

克漏詞試題(cloze) [9, 13]，克漏詞試題雖然在中文試題中比較少出現，卻是國內外英文測驗，如托福、GRE和大學指考等，幾乎是必然採用的題型。

形音近似的漢字在語言心理學的研究上也相當有用。Taft、Zhu和Peng [15] 研究部首位置對於受試者的詞彙決策(lexical decisions)與命名反應(naming responses)。Tsai等學者[16]則研究相近漢字的字數的多寡(neighborhood size)對於詞彙決策與閱讀的影響。Yeh 和 Li [17] 研究近形字對於一個熟練的中文閱讀者所執行的詞彙決策的影響。

發音相近的字可能可以藉由電子詞典所記載的資訊來判斷；相對地，形體相近的字則尚未有簡易的方法來找尋。影像處理技術雖然可能有用，但是對於爲數眾多、且近似方式繁複的漢字來說，應用影像處理技術的時效恐怕不佳。本文從應用朱邦復所設計的倉頡碼出發[2]，改變倉頡碼的原始設計，參考原本爲了補足漢字字形缺字所創造的漢字構形資訊[1]，得到一套可以爲任何漢字找尋形體近似的漢字的方法。

結合所找到音形相近的漢字字集之後，我們利用谷歌(Google)的搜尋介面所提供的資訊來排序所找到的字集的候選字，藉此排序可以限制我們所提供的近似字的字數。實驗結果顯示，不管以真人受試者或者專家意見作爲評比的標準答案，我們的系統所提供的字集都能有效協助教師編輯高品質的「改錯字」試題。

我們在第 2 節討論如何利用倉頡與構形資訊來建構一個找尋近形字的子系統。在第 3 節討論找尋漢字同音、近音字的技術問題。在第 4 節討論如何利用谷歌搜尋所得的資訊，來評比形音相近的字當中哪一些字是比較具有吸引力的錯別字。我們在第 5 節提報和分析相關的測試的結果。第 6 節則是簡單的結語。

2. 搜尋形體近似的漢字

我們在第 1 小節介紹一些近形字，在第 2 小節簡述倉頡輸入法如何將中文字編碼，在第 3 小節說明我們如何改進現有倉頡碼的編碼方式，最後第 4 小節說明我們利用關於個別漢字的資訊來找尋近形字的方法。

2.1 近形字實例

圖一、圖二和圖三包含三大類容易搞混的中文字，我們用空白將相似的中文字做分群。圖一當中的近形字，差別只在於筆劃的層次。圖二第一行各群的近形字分享同一個部件(component)而非部首。圖二第二行各群近形字則是分享同一個部件同時也是部首。圖二各組的近形字都有不同的發音。圖三爲六組分享同一部件的同音異義字。發音與內部結構相近的近形字最能造成語文學習者學習上的困擾。

士土工千千 戌戍成 田由甲申
母毋 勿匆 人入 未末 采采 凹凸

圖一、主要差異在筆畫層次的漢字

頸勁 搆溝 陪倍 硯現 裸棵 搞篙
列刑 盆盎盂盅 因困囚 間閒閃開

圖二、形體相近的漢字

形刑型 踵種腫 購構搆 紀記計
園圓員 脛逕徑痙勁

圖三、形體與發音皆相近的漢字

要有效率地找到形體相近的漢字並不見得是一件簡單的事。藉由圖像比對方法找出

形體相似的漢字，雖然是一個可能的方法，但是卻有相當的困難。以「構」與「購」爲例，雖然以肉眼比較這兩個字的影像的時候，我們會覺得這兩個字的右側所共享的部件「冓」會重疊。實際上，經過我們測試，這樣的直覺是一個誤判。字形檔的建構，並不保證共享的部件的所有影像點(pixels)都必須能夠重疊，即便共享的部件確實有相當的影像點應該可以重疊在一起。

除了以上所描述的「非完美重疊部件」的問題之外，漢字之間的相似關係還有別的類別。以「員」和「圓」爲例，不管我們把這兩個字的影像如何平移，所得的最大交集的影像點的數量可能都不容易讓我們認定這兩個漢字的相似性。所謂「相似」，其實有其主觀的因素存在，雖然不一定每一個人都會認爲「員」和「圓」相似，但是大多數的人應該都會接受這樣的看法。在某一些可能是有一些極端的應用之中，我們或許還會希望我們的程式可以找到「貝」和「圓」的相似處，這時「貝」甚至只是「圓」的內部構件的一小部分。又請看圖三中第二行右手邊的字群，他們共同分享的部件出現在不同的位置。這時候影像處理技術雖非毫無用武之地，但是所須進行的計算量可能就不小，除了平移還須要考慮放大（或者縮小）的問題。不管是平移或者是放大，都須要決定平移量、平移方向和放大的比例，這一些決策都會使得計算變得相當地複雜。而即便引入其他更加複雜的演算法，例如紋路分析(texture analysis)，計算速度也是很難提供即時快速的服務。

上述的討論，還侷限在兩個漢字的直接比對上。如果考慮到漢字的數量龐大，計算的功夫就可能更加耗時費力。中文擁有超過 22000 個漢字[11]，所以直接用影像比對字的相似度須要很大的計算量；如果欠缺一些有效資訊支援，直接比較任意兩個漢字的話，就必須處理超過 4.8 億種組合。如果只有考慮我國教育部所提出的 5401 個中文常用字[3]，則大約會有 2900 萬種組合。

詞典編纂者利用中文字的部首(radicals)，將中文字在字典中有組織地進行分段，因此部首訊息是有用處的。在圖二中的第二行，我們舉了一些例子。這些字群中擁有的共同部件，皆爲這些中文字的部首，所以我們可以在中文字典中的某一段落，找到同屬這一個字群的中文字。然而光靠詞典編纂者定義的中文部首資訊是不夠的。在圖二中第一行的中文字群，有著共同的部件。然而這些部件並非中文字的部首，舉例說明：「頸」及「勁」在字典中分屬於兩個不同的部首。

2.2 倉頡原始碼

倉頡輸入法以 25 個字作爲基本單位，創造出一套分解漢字的方法；透過這 25 個字的組合，就能把漢字輸入到電腦中。倉頡輸入法分解漢字的方法，雖然不是非常完美，但是這一個分解個別漢字爲基本單位的出發點，跟我們尋找近形字的需求是相接近的。

表一分成三個主要部分，由左而右分別列出圖一到圖三部分漢字的倉頡碼。在一部有安裝倉頡輸入法的電腦上，可以用倉頡碼輸入中文字，例如輸入「一一一月金」的話，就可以得到「頸」（註：「一一一月金」是英文鍵盤上的 MMMBC）。在倉頡輸入法中，每個漢字都被分解成爲一個有序的元素；簡而言之，我們可以發現其中的子序列能組合成一個字的主要部件。很顯然地，透過計算個別漢字所分享的倉頡碼的數目，是一個可

以決定相似字的方式。舉例來說，我們可以說「搞」和「篙」是相似的，因爲他們的倉頡碼裡都有代表「高」這個部件的「卜口月」。我們也可以輕易發現，「踵」、「種」和「腫」分享了「重」這一個部件，因爲他們的倉頡碼都包含了「竹十土」這一個子序列。

表一、 一些漢字的倉頡（原始）碼

漢字	倉頡碼	漢字	倉頡碼	漢字	倉頡碼
土	十一	頸	一一一月金	踵	口一竹十土
土	土	勁	一一大尸	種	竹木竹十土
工	一中一	硯	一口月山山	腫	月竹十土
干	一十	現	一土月山山	購	月金廿廿月
勿	心竹竹	搞	手卜口月	構	木廿廿月
匆	竹田心	篙	竹卜口月	圓	田口月金
未	十木	列	一弓中弓	員	口月山金
末	木十	刑	一廿中弓	脛	月一女一
		因	田大	逕	卜一女一
		困	田木	徑	竹人一女一
		間	日弓日	痙	大一女一
		閒	日弓月		

然而，某些形狀有微妙變化的漢字，倉頡碼似乎無法提供出它們相似的證據；例如「土土工干」和其他列在表一最左邊欄位內的字。這些字是依據特殊的分解規則解構的，這種特殊的規則使得我們無法輕易利用倉頡碼的相似度來找尋近形字。

爲了維持輸入一個漢字不須要敲擊超過五個鍵的輸入效率，倉頡輸入法蓄意簡化某些部件較多或者較複雜的漢字的倉頡碼。例如，在「脛」和「徑」的倉頡碼裡，「一女一」代表了「巠」這個部件，但是在「頸」和「勁」的倉頡碼裡，「巠」這個部件卻被簡化成「一一」。而「員」的「口月山金」在「圓」的裡面只剩下「口月金」。

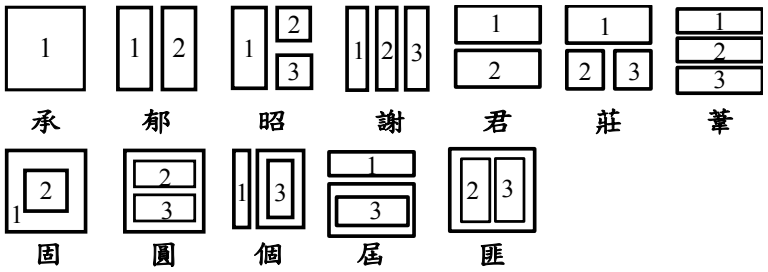
以輸入效率作爲設計要件，倉頡輸入法簡化用來代表個別漢字的內碼序列這一作法是可以理解的。然而，這樣的簡化程序使我們難以依照真實的倉頡碼來比對字的相似度。因此，我們並沒有直接採用倉頡碼做爲比較漢字相似度的基礎。爲了自己的需求，我們恢復了被倉頡輸入法所簡化的內碼，以鍵盤能夠用倉頡輸入法打出的最小構件做爲恢復原則。例如不管「巠」這個部件怎麼樣被簡化，如果我們用倉頡輸入法打「一女女一」能夠得到「巠」的話，那麼就將「巠」被簡化後的倉頡碼一律恢復爲「一女女一」。爲了在稱呼上有所區分，我們稱呼原始的倉頡碼爲***倉頡原始碼***，而我們所建構的倉頡碼爲***倉頡詳碼***。

2.3 倉頡詳碼與漢字構形資訊

雖然經過簡化所得的倉頡碼可以提升輸入效率，卻也造成我們在比對相似度的困難。因此，我們選擇使用完整的倉頡碼來建構我們的漢字資料庫。舉例來說，在我們的資料庫裡，「巠」、「脛」、「徑」、「頸」和「勁」的倉頡碼分別對應到 「一女女一」、「月一女女一」、「竹人一女女一」、「一女女一一月山金」和「一女女一大尸」。

除了不因輸入效率而簡化倉頡碼之外，我們還可以利用漢字的構形資訊來提高我們找尋近形字的效果。中研院早在將近十年前就開始了漢字構形的相關研究[1]，這一研究方向發展出了一套可以建構各種漢字字形的技術[11]。漢字的構造特殊，一般都認爲是方塊字，在這四四方方的空間裡面，我們又可以把一個漢字切割成幾個小部分，每一部分都是一個子結構；這一些子結構雖然可能正好是詞典所列的部首，有些則不然。目前我們採用Lee切割漢字的方式[12]。這一方式是以倉頡碼的「連體字」、「字首」、「字身」、「次字首」和「次字身」的觀念出發。儘管中研院所提出的切割方式並沒有完全侷

限在倉頡碼的切割觀點[1]，但是因爲漢字本身幾乎欲蓋彌彰的構造，使得我們目前所採用的切割方式與中研院的看法多所類似。



圖四、以倉頡碼爲基準的漢字基本構形

以圖二所列的漢字爲例，有些字可以從垂直方向分解成兩個部分，像是「盅」可分解成上面的「中」和下面的「皿」；有些字可以被水平分解，像是「現」是由「王」和「見」這兩個部件所組成的；而有些字則是一個包圍的結構，譬如說「囚」這個字是「人」部件被「口」部件包圍起來。因此，在決定漢字的相似度時，我們可以同時考慮構形資訊和所分享的相同部件這兩個因素。

雖然漢字有其看似自然的構形方式，但是並非每一個學者或者母語使用者都認同某一切割方式。圖四是Lee切割漢字的構形方式[12]，每個構形方式下方有一個漢字作爲範例。爲了記錄構形方式，我們賦予構形方式編號(以下稱爲***構形編號***)，在圖四裡面，由左到右、由上到下依序從 1 號開始編號。構形中的一個小方塊表示一個子結構，一個漢字最多可以有三個子結構，而且我們用數字替這些部件編號(以下稱爲***構件編號***)。依照倉頡區分子結構的方式，圖四上列最左邊的字是***連體字***，其他的字都有子結構，所以稱爲***分體字***。在分體字裡面，構件編號 1 號的子結構稱爲***字首***；如果一個分體字總共只有兩個子結構的話，構件編號 2 號的子結構就是***字身***。如果一個分體字有三個子結構的話，則構件編號 1、2 和 3 號的子結構分別是***字首***、***次字首***和***次字身***。倉頡輸入法的漢字切割方式，十分看重最左側的子結構和最上方的子結構，因此，構形編號 2、3、4 和 10 號的字的字首都是最左側的子結構；構形編號 5、6、7 和 11 號的字首都是上方的子結構。有外框的構形的字則都是以外框子結構當作是字首。

表二、一些漢字的倉頡詳碼

漢字	構形編號	構件 1	構件 2	構件 3	漢字	構形編號	構件 1	構件 2	構件 3
承	1	弓弓手人			頸	2	一女女一	一月山金	
郁	2	大月	弓中		徑	3	竹人	一女女一	
昭	3	日	尸竹	口	脛	3	月	一女女一	
謝	4	卜一一口	竹難竹	木戈	痙	7	大	一女女一	
君	5	尸大	口		員	5	口	月山金	
森	6	木	木	木	圓	9	田	口	月山金
葦	7	廿	木一	手	相	2	木	月山	
因	8	田	大		想	5	木月山	心	
國	9	田	戈	口一	箱	6	竹	木	月山
個	10	人	田	十口					
屆	11	尸	山	土					
匪	12	尸	中一	卜卜卜					

這一些構形資訊，除了影響倉頡碼簡化的規則之外，對於我們的系統提供了有利的資訊。如果把構形資訊也放進個別漢字的資料裡面的話，我們就可以知道兩個漢字的外型是否相似。也可以知道兩個漢字如果有相同的子結構，那這一相同的子結構分別出現在方塊字的那一個子空間中。這樣的資訊有助於我們更加精確地掌握兩個相似的程度。

當所有簡化的倉頡碼都回復成倉頡詳碼之後，我們可以把每個字都加上所屬的構形編號，並且將其倉頡碼依照構形中的構件編號分類。因此，我們所建構的字的紀錄包含構形標籤和一到三組倉頡碼序列。表二所列的紀錄包含圖四中的字的倉頡詳碼，並加上一些我們之後會討論的字的紀錄。在原本倉頡碼裡被省略的元素以方框框起來表示；例如，「徑」的倉頡原始碼是「竹人一女一」，同時「徑」的倉頡詳碼是「竹人一女女一」。

將原本倉頡碼中省略的元素重新加進去，以及把倉頡碼的子序列分成三部分都有助於我們辨別相似的漢字。現在我們可以輕易的分辨出由「一女女一」和「一月山金」所組成的「頸」，與「竹人」和「一女女一」所組成的「徑」是兩個相似的漢字。我們的系統也可以推測得知「員」及「圓」是相近的字。如果我們只採用倉頡原始碼的話，這將是一件不容易辦到的工作。

構形資訊則提供更加精確的近似度的訊息，「頸」、「徑」、「脛」和「痙」的倉頡詳碼都包含了「一女女一」這個子序列，但是因爲「徑」和「脛」屬於同一構形，而且所共享的子序列屬於同一子結構，所以「徑」和「脛」的相似度要高於「徑」和「頸」及「徑」和「痙」的相似度。實際上，我們還可以導入一個簡單的機制，在尋找某一特定構形編號的字的時候，特別偏好某一特定構形編號的字。例如，當我們在搜尋本身屬於構形編號 2 的字的近形字的時候，比較偏好構形編號是 2、3、4 和 10 的字，因爲這一些屬於這一些構形的字的輪廓線是類似的。

2.4 比對的方法和計算的效率

我們已經以人工建立了五千多個漢字的倉頡詳碼的資料庫[7]。從這五千多字之中，找尋一組形體相近的漢字，只須要用最簡單的字串比對即可。在最糟糕的情況下，比對兩個都擁有三個子結構的字的時候，我們必須比較他們各自的倉頡碼子序列九次。實施字串比對的步驟是很簡單的，因此計算量極小。

在計算近似度的時候，目前我們考慮幾個因素。首先是兩個漢字的倉頡詳碼的子結構是否有完全相同的倉頡碼。如果有的話，我們會記錄這一共享的子結構的倉頡碼的個數當作分數。如果共享的子結構是同一空間位置的子結構的話，則分數還會加倍。分別比對完最多九個子結構組合之後，會把所得的分數加總，然後得到一個初步的總分。我們以這一個總分把資料庫裡面的字排序(爲了計算效率，我們很早就把得到零分的字捨棄)；如果有同分的字，則再以構形編號和目前所查詢的漢字的構形編號相同者作爲最後分數較高者，然後才是構形相似的字。

以表二的「頸」、「徑」、「脛」和「痙」爲例。如果我們是要尋找「徑」的近形字時，因爲這四個字共享「一女女一」這一個構件，所以我們可以知道「頸」、「脛」和「痙」跟「徑」有基本的相似性；相對之下，其他的字，如「員」，跟「徑」沒有相同的構件，因此不是相近的字。除了知道「頸」、「脛」和「痙」跟「徑」相近之外，我們如何知道

這三個字之中哪一個字跟「徑」比較相像？在這一個例子裡面，我們發現「脛」和「徑」的共享構件都是出現在第二個構件。相對地，「頸」的「一女女一」是在第一個構件，而「痙」的構形編號是 7，並不是構形編號 2、3、4 和 10 的字，因此視覺上跟「徑」的相似度還是比不上「脛」。

在一部使用 2.24G 的 RAM 和 Windows XP 的 Pentium 4 2.8G 機器上，從五千多個字搜尋資料庫內的字的相似字，只須要花費不到一秒鐘的時間。如果是要把這一系統當作是電腦輔助出題系統的基礎子系統的話，這樣的速率，在實務上應該符合時效。

3. 發音相近的漢字

我們可以利用一部可以提供漢字發音訊息的電子化詞典，找出任意字的同音字或者近音字。如果能夠掌握一部可靠的詞典，則找尋同音字是一件簡單的事情；例如「試場」和「市場」。然而要找尋所謂的近音字，例如「光陰」和「光影」，則須要一些語音學知識的支援。

目前我們還沒有考慮實際生活中漢語發音的變調(Sandhi)[10]現象，所以在判斷個別漢字發音的近似度的時候，還沒有考慮語境的影響，因此並不是絕對準確。變調是許多語言都有的現象，最常聽到的漢語變調規則是兩個連續的三聲字的第一個三聲字要以二聲來發音；例如，在自然發音的情境下，「選舉」和「演講」的「選」和「演」都被當作二聲字來發音。如果有多個三聲字連續出現，則還有更加複雜的發音慣例。此外，習慣上我們會把「媽媽」的第二個「媽」以輕聲來發音，這也是一種受語境影響而改變發音的例子。目前我們還沒有完成處理這一類發音變化的程式。

除此之外，我們可以蒐集一些生活中一些常見的發音問題或者透過問卷調查，以獲取母語使用者對於所謂「容易混淆的音」的資料。舉例來說，在一般國民小學的教學經驗裡面，ㄓ、ㄔ和ㄕ這三個音分別容易跟ㄗ、ㄘ和ㄙ混淆。類似的經驗如，ㄣ跟ㄥ這兩個韻母相當接近，所以「金雞」和「京畿」聽起來很相像；而「ㄏㄨㄢ」 這一個發音跟「ㄈㄢ」聽起來相當接近，所有常有人把「冒險患難」寫成「冒險犯難」；或者有人把「翻書」唸成「歡書」。這一類的「近音字」如果有好的資訊來源，很容易由程式列出所有相關資料。目前我們採用了中研院語言所李佳穎研究員所提供的一些資料[5]，作爲尋找漢字近音字的依據。

4. 排比易混淆的漢字

我們可以用第 2 節和第 3 節所闡述的技術，來找尋單一漢字的近形字、近音字和同音字。我們可以利用這一些相近的字，爲某一個詞彙找出易混淆的錯誤寫法，例如，蓄意把「冒險患難」寫成「冒險犯難」。這類的錯誤詞彙對編寫國語科的「改錯字」試題和研究閱讀認知歷程都有特定的用處。

對近形字而言，不管是以中研院的漢字構形資訊[1]做爲構形編號的依據也好，或是以Lee切割漢字的構形方式[12]做爲依據也罷，本論文所提出的找近形字技術只是一種過濾機制，試圖取出「可能」會被誤用的易混淆字群而已；我們並不急著在近形字的部分就取出最常被誤用的易混淆字，因爲最常被誤用的易混淆字也有可能是同音字或是近音字。

以「勉強」這一個詞爲例，如果目的是把「勉」由一個錯別字來取代，則可以應用找尋近形字的技術找到「免」和「兔」，再用找尋近音字的技術找到「冕」、「娩」和「緬」等其他同音字。於是我們面臨了如何把這一些近形字、同音字和近音字排序，好讓我們系統的使用者盡快找到滿意的錯別字的需求。跟一般的注音輸入法類似，愈容易混淆的漢字最好是放在建議名單的前頭，以減少使用者的搜尋時間。這就是本節所要交代的「排比」問題。

實際經驗顯示，雖然我們可以依據倉頡詳碼找尋近形字，但是光是形體相近並不見得就是很有用的錯別字；因爲即使兩個漢字真的有一些相似的地方，也不一定會讓人們感到容易混淆。以經驗上的直覺來看，「改錯字」試題比較常用同音字或者近音字來取代正確的字。形體相近的漢字可能對於語言心理學中關於語文閱讀的研究有比較大的用途。儘管如此，我們仍將以編寫「改錯字」試題爲目的，探討如何排比我們所得到的候選字（包含近形字、同音字和近音字）。

延續「勉強」這一個例子，我們如何猜測「免」、「兔」、「冕」、「娩」和「緬」個別的適合程度？一位出題老師當然有他的主觀感覺，但是一個軟體程式如何能有這樣的「主觀」意識呢？我們借重谷歌(Google)所提供的搜尋服務來模擬這樣的直覺。如果我們以「免強」加上雙引號進行查詢，在所得的查詢結果中，觀看 “Results of 1-10 of about 220,000 …”（附註：如果是使用谷歌的中文介面，則會看到 “關於**免強**大約有 220,000 頁…”），可以知道大約有廿二萬筆網頁資料用到「免強」這一個詞。在查詢的時候加上雙引號，用意在於告知谷歌把查詢的詞彙當作一個連續字串、不可以分開查詢，因此會排除只包含類似「免，強」之類的網頁資料，所得的結果會比較符合我們的需求。我們可以撰寫一段簡單的程式，把所要查詢的條件傳送給谷歌，然後再從所得的回傳資料，透過簡單的資訊擷取機制得到所要的數字。這一程序所得的數字大略地反應了網路社群中使用這一個錯誤詞彙的頻率，可以詮釋爲生活中人們犯同樣錯誤的相對機會，如果數量愈大則表示人們採用那一個錯誤的詞彙的機會也相對地高。

我們以「免強」、「兔強」、「冕強」、「娩強」和「緬強」這五個詞，分別加上雙引號作爲給谷歌的查詢關鍵詞，會得到 222,000、4720、506、78 和 1510（附註：這一批數字得自於 2008 年 7 月 7 日的試驗）。因此，如果我們要從「免」、「兔」、「冕」、「娩」和「緬」之中，提交三個候選字給使用者時，我們依序提出「免」、「兔」和「緬」；如果是要提交五個候選字的話，則依序在後面加上「冕」和「娩」給使用者。

5. 實驗結果與分析

在這節裡，我們將以編寫「改錯字」試題爲目的，檢驗我們的系統是否能找出現實生活中易混淆的漢字。本節將分成四個部分來介紹我們所使用的實驗設計、分別使用一般受試者與專家意見所進行的實驗結果、最後討論使用倉頡詳碼判別近形字的缺點。

5.1 實驗設計

首先我們從*新編錯別字門診*[8]這一本書，找出 20 個包含有容易混淆的字的詞彙。表三所列的是我們所選定的詞，每一個詞彙都包含一個以底線標示的粗體字。爲了行文簡潔，以下我們以***易混淆字***來稱呼表三裡面這一些以底線標示的粗體字。這個易混淆字將會特意地被一些***錯別字***來取代。

我們以兩種資料來檢驗我們系統所提出的建議錯別字的品質。我們請真人受試者寫出他們認爲適合取代這一些易混淆字的錯別字。然後以所蒐集的這一些資料，來評比我們系統所建議的錯別字的效用和評量受試者之間的一致性。同時，我們也會利用*新編錯別字門診*這一本書所討論的常見錯字，來評比我們系統的建議字和受試者所寫的錯別字。我們以真人受試者所提供的資料來反應一般人對於這一些錯別字的選擇，而用書本所提供的錯別字來代表專家的意見。

首先，利用我們的系統將這 20 個易混淆字找出「近形」、「同音」、「近音」三個候選字表；再將這三個候選字表組合成「同音、近音」、「近形、同音、近音」兩個***建議字表***，並以建議字表裡的字逐一取替表三所列詞彙的易混淆字。以這樣程序所產生的詞彙，再利用第 4 節所描述的程序，取得個別詞彙被網頁資料採用的頻率，藉以將建議字表內的字排序，使得最前面的字爲被採用頻率最高的字。如果前一程序所得到的搜尋結果數量爲 0，再利用谷歌搜尋「包含全部的字詞」功能(即不加上雙引號直接進行查詢)所回傳的搜尋結果數量，由大到小排於前一個程序的排序之後。

本實驗希望探討我們的系統是否能找出現實生活中被用來取代易混淆字的錯別字，因此請了 21 位大學在學生（校名因匿名投稿之故，暫且不明列）擔任受試者來進行實驗。我們請這 21 位受試者針對前述 20 個詞裡的易混淆字寫下至多五個錯別字。所收集到的共 420 (=21×20)個題次的回覆中，一共包含 599 個字（包含重複的字），其中有 24 個其實不是漢字的錯字。這 24 個錯字答案分佈在 7 個真人受試者的答案裡面。如果不管字的對錯，平均每一題次，真人受試者每一題平均填寫了 1.426 個建議字；如果扣除錯誤的字的話，就只剩下 1.369 個建議字。

表三、測試詞彙的列表

編號	詞彙	編號	詞彙	編號	詞彙	編號	詞彙
1	一**剎**那	2	一**炷**香	3	眼花**撩**亂	4	相形見**絀**
5	作**踐**	6	剛**愎**自用	7	可見一**斑**	8	和**藹**可親
9	**彗**星	10	**委**靡不振	11	**穠**纖合度	12	待價而**沽**
13	獎**券**	14	意興闌**珊**	15	**罄**竹難書	16	**搔**首弄姿
17	根深**柢**固	18	**椿**萱並茂	19	煩**躁**	20	璀**璨**

我們採用資訊檢索相關研究中最常使

用的評估標準，即精確率(precision)與召回率(recall) [14]做爲評估我們系統的方式；精確率是系統所建議的字當中是標準答案的比例，召回率則是標準答案中被囊括到系統所建議的字的比例。在計算精確率和召回率的時候，我們把錯誤的答案也當作答案，實際上我們的系統是不能建議這一些根本不存在的漢字的。因此，我們目前並不會因爲真人受試者的回覆的品質不佳，而高估了我們的效用。這一些真人受試者的錯誤甚至還讓我們低估了我們系統的分數。

在目前的研究中，我們沒有採用 F 分數(F measure)。F 分數是利用召回率和精確率計算所得的單一分數，雖然可以提供基礎的比較。但是在我們的實驗跟漢字輸入法的評估相當類似，使用者所能接受的建議錯別字的數量可能極少，因此分別檢視精確率與召回率，比起只有提供 F 分數更能讓研究者看清問題的本質。

5.2 一般受試者的評估

我們以 21 位受試者爲表三的 20 個易混淆字所寫下的 20 組錯別字做爲標準答案來進行評估。表四爲「同音、近音」與「近形、同音、近音」兩個建議字表各取前五個字與前十個字的實驗結果，我們利用建議字表所提供的錯別字與另外 21 位受試者所寫下的 20 組錯別字一組一組地進行評估比對，分別得到 20 組各個易混淆字的精確率與召回率，接著把這二十組精確率與召回率做平均的計算，因此得到以這 21 位受試者的錯別字爲標準答案時建議字表的精確率與召回率。然後再計算這 21 組數據的平均，所得到的計算結果就是表四中的平均精確率與平均召回率。

表四所列的數據顯示，我們所提出的方法相當有效地捕捉到受試者的偏好。以「近形、同音、近音」資料所建構的建議表的實驗來說，平均精確率看起來雖然不高，在提交五個建議字和十個建議字的時候，平均的精確率分別略低於 0.2 和 0.1。不過這意味著不管是提交五個字或者十個字，我們的系統都能夠大約提供出受試者所寫的錯別字。依據前一小節的分析，平均而言，針對每一題次，受試者只有寫出 1.369 個實際上存在的漢字作爲錯別字。因此，我們的系統能夠捕捉到這一些特定的字並不是一件絕對簡單的任務。從這一個觀點看我們的系統，便能看出它的可用性。如果仔細比較一下，提交五個建議字的時候，平均有 0.88 個可用字；提交十個建議的話，平均就有 0.95 個可用字，命中率不可謂不高。

表四、兩組系統建議字表所達成的平均精確率與平均召回率

系統建議字表 / 效果評估	「同音、近音」		「近形、同音、近音」	
	取前五個字	取前十個字	取前五個字	取前十個字
平均精確率	0.166	0.094	0.176	0.095
平均召回率	0.618	0.672	0.649	0.680

這一組實驗，同時也讓我們看到近形字對於改錯字試題的編輯工作的貢獻度似乎不大。比較表四的左半側和右半側的實驗數據，我們發現加入近形字之後所建構的建議表雖然效果都有所提升，但是提升的幅度並不顯著。這一實驗結果暗示著中文錯別字跟發音（同音字或者近音字）的關係，可能比跟字的形體的關係要密切。這樣的觀察當然可能是跟表三裡面我們所選擇的詞彙有關。在表三裡面，常見的錯別字只有第四題是跟字形比較相關；第 15 題勉強也算跟字形相關。其他的試題則都是明顯的跟字的發音有比較高的關連。再以「不虛此行」爲例，除非是以同音字作爲搜尋的要件，否則很難找到

「不需此行」這一個誤用的形式。

表五、受試者之間的平均精確率與平均召回率

受試者代號	平均精確率	平均召回率	受試者代號	平均精確率	平均召回率
A	0.569	0.473	L	0.458	0.368
B	0.553	0.508	M	0.365	0.455
C	0.408	0.635	N	0.520	0.491
D	0.495	0.468	O	0.448	0.546
E	0.497	0.520	P	0.558	0.481
F	0.489	0.479	Q	0.370	0.513
G	0.580	0.462	R	0.379	0.559
H	0.408	0.304	S	0.441	0.444
I	0.628	0.509	T	0.435	0.543
J	0.539	0.431	U	0.451	0.491
K	0.531	0.443			

中文錯別字跟發音的關係是不是真的比跟字形的關係要來得密切呢？雖然我們是透過一個隨機的程序挑選出表三所列的 20 道題，因此也認爲我們的實驗數據應該是暗示了這一現象，但是須要一個更大規模檢驗才能更進一步地驗證這一直覺。

爲了能夠更客觀地評估系統的效果，我們又做了另一組實驗。從這 21 位受試者當中輪流取出一人所寫下的 20 組錯別字當做系統建議字表所提供的錯別字，並以其他 20 位受試者的錯別字當作標準答案，來評估這一個暫時被挑出來的受試者答案的品質。我們用英文字母 A 到 U 來代表這 21 位受試者，表五列出這一些受試者輪流被當作被評估對象時所得的分數。

在這一組新的實驗中，我們取出每位受試者所寫下的 20 組錯別字作爲假想的建議字表，再以假想之建議字表所提供的錯別字與另外 20 位受試者個別寫下的 20 組錯別字一組一組地進行評估。評估的過程中我們會計算 20 組各個錯別字的精確率與召回率，接著計算這 20 組精確率與召回率的平均，得到以個別受試者所提供的錯別字爲標準答案時的精確率與召回率。接著再計算這 20 組數據的平均，最後所得到的計算結果就是表五中的平均精確率與平均召回率。

以整個表五的數據來評估所有受試者的平均表現，分別把所有的精確率加總，然後除以人數，平均的精確率和平均召回率分別是 0.48200 和 0.48205，兩者幾乎相等。受試者之間的共識度雖然表面上看起來不高，但是這一些受試者在接受我們測試之前並未事先相互交換意見，同時是獨立回答問卷，所以這一個平均的精確率和召回率應該算是相當高。如果拿這一個總平均數跟表四的平均數相比的話，我們發現我們系統的精確率雖然比不上人類受試者，但是召回率卻能高於人類受試者間的召回率。這一項比較顯示我們的系統可以提供有用的服務，在容許系統提供五個建議字或者十個建議字的情形下，我們系統比人類受試者更能提供確實有用的建議字。

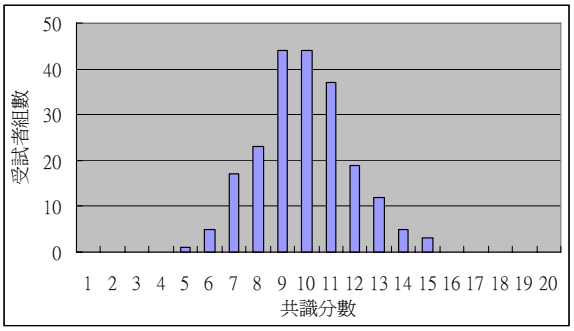
另外，我們對於這 21 位學生所認定的易混淆字是否有清楚的共識也有很大的興趣，於是我們取出每個學生所寫下的 20 組易混淆字當中的第一個字，並以此 20 個字爲該學生的***第一印象字***，故共有 21 組第一印象字，每組 20 字。接著以組爲單位，兩兩比較在相同位置上的字，若是一樣的話，則給予一分的權重；因此若是一模一樣的兩組來比較的話，則可以獲得滿分 20 分。我們將此認知共識的分析結果以方陣表示，並由以

表六、受試者之間的共識的分數方陣

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	甲	乙	丙
A	20	13	10	13	9	9	9	8	12	10	12	9	8	13	8	12	7	7	9	10	8	9	9	11
B	13	20	9	10	9	9	8	8	11	9	12	8	8	9	8	12	6	7	9	9	8	7	7	13
C	10	9	20	7	11	8	7	5	8	8	10	6	7	7	8	11	7	8	9	9	9	4	4	7
D	13	10	7	20	9	10	11	6	11	10	10	8	8	14	8	9	6	7	8	9	9	11	11	12
E	9	9	11	9	20	10	9	6	10	10	11	8	7	9	6	9	8	10	11	9	8	7	7	10
F	9	9	8	10	10	20	14	9	11	10	9	9	9	10	7	12	7	9	8	10	10	8	8	12
G	9	8	7	11	9	14	20	8	12	10	10	10	10	14	8	11	8	9	8	10	13	10	10	14
H	8	8	5	6	6	9	8	20	8	8	7	5	7	5	5	10	6	6	6	9	4	4	5	7
I	12	11	8	11	10	11	12	8	20	12	12	11	8	13	7	10	8	10	10	11	12	7	8	13
J	10	9	8	10	10	10	10	8	12	20	12	10	9	10	6	11	8	8	10	10	9	8	9	12
K	12	12	10	10	11	9	10	7	12	12	20	9	9	10	6	10	8	8	10	9	11	7	7	11
L	9	8	6	8	8	9	10	5	11	10	9	20	7	11	5	7	6	7	7	9	9	8	8	11
M	8	8	7	8	7	9	10	7	8	9	9	7	20	8	6	11	6	9	6	9	8	5	5	8
N	13	9	7	14	9	10	14	5	13	10	10	11	8	20	8	8	6	8	8	10	11	11	11	14
O	8	8	8	8	6	7	8	5	7	6	6	5	6	8	20	9	7	8	6	7	7	7	7	9
P	12	12	11	9	9	12	11	10	10	11	10	7	11	8	9	20	9	8	10	11	11	6	7	10
Q	7	6	7	6	8	7	8	6	8	8	8	6	6	6	7	9	20	11	10	9	7	7	8	8
R	7	7	8	7	10	9	9	6	10	8	8	7	9	8	8	8	11	20	12	9	8	5	5	9
S	9	9	9	8	11	8	8	6	10	10	10	7	6	8	6	10	10	12	20	10	9	6	7	11
T	10	9	9	9	9	10	10	9	11	10	9	9	9	10	7	11	9	9	10	20	10	7	8	12
U	8	8	9	9	8	10	13	4	12	9	11	9	8	11	7	11	7	8	9	10	20	7	7	10
甲	9	7	4	11	7	8	10	4	7	8	7	8	5	11	7	6	7	5	6	7	7	20	16	12
乙	9	7	4	11	7	8	10	5	8	9	7	8	5	11	7	7	8	5	7	8	7	16	20	13
丙	11	13	7	12	10	12	14	7	13	12	11	11	8	14	9	10	8	9	11	12	10	12	13	20

上的論述推導可知，此方陣必是對角線爲 20 之對稱方陣，如表六所示（表六中的「甲」、「乙」和「丙」爲標頭的三個欄和三個列的資料，將在下一小節說明）。

表六這一組原始數據可以讓我們看到個別受試者之間共識程度的變化，但是不容易看到一般的趨勢。我們可以分析以左上到右下的對角線爲準的右上三角形的數字，在不考慮對角線上的完美分數 20 分的情形下，計算有多少組受試者的共識是 19 分、18 分、…、0 分，把得到這一些分數的組數記錄下來，然後繪製一個受試者共識的趨勢圖。圖五是這一程序的產物，橫軸是分數，縱軸是得到橫軸所列分數的受試者組數。從圖形可以看得出來，受試者的共識分數像是常態分佈。經過簡單的計算，由 21 位受試者所形成的 210 個分數的平均數是 8.905。如果以受試之間的共識分數當作分子，完美的共識分數 20 當作分母計算，則這一平均分數的百分分數只有 44.5%。



圖五、21 位受試者的共識分數的分佈

5.3 專業意見的評估

*新編錯別字門診*的作者除了表列包含有易混淆字的詞彙之外，也提供了所列詞彙最常見的錯誤寫法。我們可以把這一些錯別字當作是專業意見，以這一些專業意見來評估我們的建議字表的效用和前面一般受試者所寫的錯別字的品質。表七的內容是把表三的易混淆字改成專家所提供的錯別字，第 7 和第 13 題有兩個可能的錯別字，第二順位的錯別

字放在括號裡面。

表七、專業意見管道所列的錯別字[8]

編號	詞彙	編號	詞彙	編號	詞彙	編號	詞彙
1	一霎那	2	一柱香	3	眼花瞭亂	4	相形見拙
5	作賤	6	剛復自用	7	可見一般(斑)	8	和靄可親
9	慧星	10	萎靡不振	11	濃纖合度	12	待價而估
13	獎卷(券)	14	意興闌跚	15	磬竹難書	16	騷首弄姿
17	根深底固	18	樁萱並茂	19	煩燥	20	璀璨

我們以這一專家意見來評估我們系統的成效，重複前一小節建立表四的同一實驗，得到表八的數據。結果顯示，當我們以專家意見作爲標準答案的時候，我們系統所得的平均精確率和平均召回率，都還比用一般受試者的意見爲標準答案時要高。其中一部份的原因，應該是跟專家意見裡面不會有不存在的漢字，所以召回率明顯提高有關。

表八、依據專家意見爲標準所計算的平均精確率與平均召回率

系統建議字表 / 效果評估	「同音、近音」		「近形、同音、近音」	
	取前五個字	取前十個字	取前五個字	取前十個字
平均精確率	0.170	0.085	0.190	0.095
平均召回率	0.775	0.775	0.875	0.875

表九、以專家意見爲標準答案，21 位受試者的答案的品質

受試者代號	平均精確率	平均召回率	受試者代號	平均精確率	平均召回率
A	0.550	0.550	L	0.550	0.525
B	0.650	0.725	M	0.317	0.500
C	0.371	0.675	N	0.667	0.725
D	0.575	0.625	O	0.533	0.700
E	0.504	0.625	P	0.550	0.550
F	0.600	0.650	Q	0.329	0.550
G	0.750	0.700	R	0.327	0.600
H	0.400	0.375	S	0.458	0.525
I	0.675	0.650	T	0.467	0.675
J	0.575	0.575	U	0.458	0.575
K	0.525	0.500			

表九是以專家意見的錯別字爲標準答案的時候，21 位一般受試者所提供的錯別字的品質。計算表九 21 組數據的平均，可以得到一般受試者的平均精確率和平均召回率分別是 0.51576 和 0.59881。比起表五的平均值 0.48200 和 0.48205 要明顯高很多。這一項結果可以有兩種詮釋方式；直覺上的看法是：一般受試者的意見與專家意見有比較高的一致性；而另一個詮釋則是，書本的作者確實掌握到了一般讀者能夠想到的錯別字。

重複上一小節中第三個實驗時（只允許我們系統建議一個候選字），我們可以加入兩個建議表和專家意見。表六裡面的甲欄和甲列是「同音、近音」系統建議字表，乙欄和乙列是「近形、同音、近音」系統建議字表，丙欄和丙列是專家意見所得的分數。因爲表六是一個對稱方陣，所以同一標頭的欄與列的資料都會是一樣的。

我們計算甲欄和乙欄裡面，由上而下從 A 列到 U 列的一致性分數的平均，分別得到 7.19 和 7.52 分。也就是，這 21 位一般受試者跟我們兩種建議表的一致性大約落在七分的位置。如果計算丙欄，同樣這 21 個數字的平均的話，我們得到 10.66。專家意見跟一般受試者意見的一致性超過一半的測試題目，專家的意見還是比我們系統的建議更能捕捉到一般受試者的想法。

丙欄的甲列和乙列的分數，代表專家意見與我們兩個建議表的一致性，分別是 12 和 13 分，平均爲 12.50 分。如果拿 12.50 和 10.66 直接比較的話，專家意見給我們系統的分數還要高於給予 21 位一般受試者的平均分數。

5.4 使用倉頡碼的缺點

使用倉頡碼作爲基礎來比較字的相似度會存在一些潛在的問題。

倉頡碼在一些漢字的分類上，尤其是比較簡單的字，會採用一些模糊的規則，這讓我們在比對字的相似度時發生困難。舉例來說，「分」是採用圖四裡面的第五種構形，但是「兌」卻是採用第一種構形。此外，表一最左邊欄位所列舉出的字都很不容易被鑑定爲近形字。這一類筆畫非常簡單的字爲數不多，處理的方式可能是以人工建立資料庫，比起回頭運用影像處理技術來找近形字經濟得多。

一個字的倉頡碼的左半邊或者上半邊的字首最多只能有一個子結構。以表二的「相」、「想」和「箱」爲例，「相」這個字單獨存在時是使用構形編號 2 號；而在「箱」裡面，「相」這個子結構則是被分解成左右兩個部份。但是，在「想」這個字裡，「相」卻被當作是一個單獨的子結構，因爲它的位置處於整個字的上半邊，被當作一個字首。類似這樣的問題常常發生，例如「森」和「焚」及「恩」和「 困」。另外還有一些比較特殊的例子，像是「品」這個字使用第六種構形，但「闆」卻是使用第五種。

這一種問題的處理，暗示了我們須要重新檢視倉頡碼觀點的構形原則。我們或許應該建立自己的構形方式，這樣可以讓我們系統所找出的近形字的精確度更加提高。

6. 結語

本篇論文報告了我們如何利用倉頡碼拆解漢字的觀念來搜尋漢字的近形字。配合適當的電子詞典資料，我們的系統能夠從發音和形體兩個不同角度，找出所欲查詢的漢字的候選錯別字。我們進一步利用谷歌的查詢功能替所查到的候選錯別字排序，實驗顯示排序之後所得的建議字表確實能夠掌握一般受試者和專業意見所提供的錯別字。我們系統所產出的形音相近的字表，除了可以應用於電腦輔助試題編輯系統中的「改錯字」試題之外，也可以用於心理語言學實驗中檢驗中文使用者的閱讀行爲。

儘管現在有相當不錯的成果，但是我們也發現了以倉頡碼作爲系統設計的核心所引起的問題，我們也尚未完成對於漢語變調規則的處理程式，這一些都是正在進行的改進項目。

致謝

本研究承蒙國科會研究計畫 NSC-95-2221-E-004-013-MY2 的部分補助謹此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導，雖然我們已經在從事相關的部分研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] 中央研究院。中央研究院漢字構形資料庫。網址：http://www.sinica.edu.tw/~cdp/cdphanzi/。Last visited on 6 July 2008。

[2] 朱邦復。*第五代倉頡碼輸入法手冊*。網址：http://www.cbflabs.com/book/ocj5/ocj5/index.html。Last visited on 6 July 2008。

[3] 行政院主計處電子資料處理中心。CNS11643 中文標準交換碼。網址：http://www.cns11643.gov.tw/web/word/big5/index.html。Last visited on 6 July 2008。

[4] 任紹曾主編，張少伯譯。*同音異義詞 Homophones*，商務印書館，2000。

[5] 李佳穎研究員，中央研究員語言學研究所，私人通訊，2008。

[6] 林仁祥及劉昭麟。國小國語科測驗卷出題輔助系統，*2007 年台灣網際網路研討會論文集*，論文光碟，2007。

[7] 教育部字頻總表。網址：http://www.edu.tw/files/site_content/M0001/pin/biau1.htm?open。Last visited on 9 July 2008。

[8] 蔡有秩及蔡仲慶。*新編錯別字門診*，螢火蟲出版社，2003。

[9] J. Burstein and C. Leacock. Editors. *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, ACL, 2005.

[10] M. Y. Chen. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press, 2000.

[11] D. Juang, J.-H. Wang, C.-Y. Lai, C.-C. Hsieh, L.-F. Chien, and J.-M. Ho. Resolving the unencoded character problem for Chinese digital libraries, *Proceedings of the Fifth ACM/IEEE Joint Conference on Digital Libraries*, 311–319, 2005.

[12] H. Lee. *Cangjie Input Methods in 30 Days*, http://input.foruto.com/cjdict/Search_1.php, Foruto Company, Hong Kong. Last visited on 8 July 2008.

[13] C.-L. Liu, C.-H. Wang, and Z.-M. Gao. Using lexical constraints for enhancing computer-generated multiple-choice cloze items, *International Journal of Computational Linguistics and Chinese Language Processing*, **10**(3), 303−328. 2005.

[14] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[15] M. Taft, X. Zhu, and D. Peng. Positional specificity of radicals in Chinese character recognition, *Journal of Memory and Language*, **40**, 498–519, 1999.

[16] J.-L. Tsai, C.-Y. Lee, Y.-C. Lin, O. J. L. Tzeng, and D. L. Hung. Neighborhood size effects of Chinese words in lexical decision and reading, *Language and Linguistics*, **7**(3), 659−675, 2006.

[17] S.-L. Yeh and J.-L. Li. Role of structure and component in judgments of visual similarity of Chinese characters, *Journal of Experimental Psychology: Human Perception and Performance*, **28**(4), 933–947, 2002.

A Realistic and Robust Model for Chinese Word Segmentation

Huang Chu-Ren
Institute of Linguistics
Academia Sinica
churenhuang@gmail.com

Yo Ting-Shuo
TIGP CLCLP
Academia Sinica
tingshuo.yo@gmail.com

Petr Šimon
TIGP CLCLP
Academia Sinica
petr.simon@gmail.com

Hsieh Shu-Kai
Department of English
National Taiwan Normal University
shukai@gmail.com

Abstract

A realistic Chinese word segmentation tool must adapt to textual variations with minimal training input and yet robust enough to yield reliable segmentation result for all variants. Various lexicon-driven approaches to Chinese segmentation, e.g. [1,16], achieve high f-scores yet require massive training for any variation. Text-driven approach, e.g. [12], can be easily adapted for domain and genre changes yet has difficulty matching the high f-scores of the lexicon-driven approaches. In this paper, we refine and implement an innovative text-driven word boundary decision (WBD) segmentation model proposed in [15]. The WBD model treats word segmentation simply and efficiently as a binary decision on whether to realize the natural textual break between two adjacent characters as a word boundary. The WBD model allows simple and quick training data preparation converting characters as contextual vectors for learning the word boundary decision. Machine learning experiments with four different classifiers show that training with 1,000 vectors and 1 million vectors achieve comparable and reliable results. In addition, when applied to SigHAN Bakeoff 3 competition data, the WBD model produces OOV recall rates that are higher than all published results. Unlike all

previous work, our OOV recall rate is comparable to our own F-score. Both experiments support the claim that the WBD model is a realistic model for Chinese word segmentation as it can be easily adapted for new variants with robust result. In conclusion, we will discuss linguistic ramifications as well as future implications for the WBD approach.

Keywords: segmentation.

1. Background and Motivation

The paper deals with the fundamental issue why Chinese word segmentation remains a research topic and not a language technology application after more than twenty years of intensive study. Chinese text is typically presented as a continuous string of characters without conventionalized demarcation of word boundaries. Hence tokenization of words, commonly called word segmentation in literature, is a pre-requisite first step for Chinese language processing. Recent advances in Chinese word segmentation (CWS) include popular standardized competitions run by ACL SigHAN and typically high F-scores around 0.95 from leading teams [8]. However, these results are achieved at the cost of high computational demands, including massive resources and long machine learning time. In fact, all leading systems are expected to under-perform substantially without prior substantial training. It is also important to note that SigHAN competitions are conducted under the assumption that a segmentation program must be tuned separately for different source texts and will perform differently. This is a bow to the fact that different communities may conventionalize the concept of word differently; but also an implicit concession that it is hard for existing segmentation programs to deal with textual variations robustly.

[15] proposed an innovative model for Chinese word segmentation which formulates it as simple two class classification task without having to refer to massive lexical knowledge base. We refine and implement this Word Boundary Decision (WBD) model and show that it is indeed realistic and robust. With drastically smaller demand on computational resources, we achieved comparable F-score with leading Bakeoff3 teams and outperform all on OOV recall, the most reliable criterion to show that our system deals with new events effectively.

In what follows, we will discuss modeling issues and survey previous work in the first section. The WBD model will be introduced in the second section. This is followed by a description of the machine learning model is trained in Section 4. Results of applying this implementation to SigHAN Bakeoff3 data is presented in Section 5. We conclude with

discussion of theoretical ramifications and implications in Section 6.

2. How to model Chinese word segmentation

The performance of CWS systems is directly influenced by their design criteria and how Chinese word segmentation task is modeled. These modeling issues did not receive in-depth discussion in previous literature:

Modeling Segmentation. The input to Chinese word segmentation is a string of characters. However, the task of segmentation can be modeled differently. All previous work share the assumption that the task of segmentation is to find our all segments of the string that are words. This can be done intuitively by dictionary lookup, or be looking at strength of collocation within a string, e.g. [12]. Recent studies, e.g. [14, 16, 5, 17], reduce the complexity of this model and avoided the thorny issue of the elusive concept of word at the same time by modeling segmentation as learning the likelihood of characters being the edges of these word strings. These studies showed that, with sufficient features, machine learning algorithms can learn from training corpus and use their inherent model to tokenize Chinese text satisfactorily. The antagonistic null hypothesis of treating segmentation as simply identifying inherent textual breaks between two adjacent characters was never pursued.

Out-of-Vocabulary Words. Identification of Out-of Vocabulary words (OOV, sometimes conveniently referred to as new words) has been a challenge to all systems due to data sparseness problem, as well as for dealing with true neologisms which cannot be learned from training data per se. This requirement means that CWS system design must incorporate explicit or implicit morphology knowledge to assure appropriate sensitivity to context in which potential words occur as previously unseen character sequences.

Language Variations. Especially among different Chinese speaking communities. Note that different Chinese speaking communities in PRC, Taiwan, Hong Kong Singapore etc. developed different textual conventions as well as lexical items. This is compounded by the usual text type, domain, and genre contrasts. A robust CWS system must be able to adapt to these variations without requiring massive retraining. A production environment with it's time restrictions possesses great demands on the segmentation system to be able to quickly accommodate even to mixture of text types, since such a mixture would introduce confusing contexts and confuse system that would rely too heavily on text type, i.e. particular lexicon choice and specific morphology, and too large a context.

Space and time demands. Current CWS systems cannot avoid long training times and large memory demands. This is a consequence of the segmentation model employed. This is acceptable when CWS systems are used for offline tasks such as corpora preprocessing, where time and space can be easily provided and when needed. However, for any typically web-based practical language engineering applications, such high demand on computing time is not acceptable.

2.1 Previous works: a critical review

Two contrasting approaches to Chinese word segmentation summarize the dilemma of segmentation system design. A priori, one can argue that segmentation is the essential tool for building a (mental) lexicon hence segmentation cannot presuppose lexical knowledge. On the other hand, as a practical language technology issue, one can also argue that segmentation is simply matching all possible words from a (hypothetical) universal lexicon and can be simplified as mapping to a large yet incomplete lexicon. Hence we can largely divide previous approaches to Chinese word segmentation as lexicon-driven or text-driven.

Text-Driven. Text-driven approach to segmentation relies on contextual information to identify words and do not assume any prior lexical knowledge. Researches in this approach typically emphasize the need for an empirical approach to define the concept of a word in a language. [12] work based on mutual information (MI) is the best-known and most comprehensive in this approach. The advantage of this approach it can be applied to all different variations of language and yet be highly adaptive. However, the basic implementation of MI applies bi-syllabic words only. In addition, it cannot differentiate between highly collocative bigrams (such as 就不 *jiubu* "…then not…") and words. Hence it typically has lower recall and precision rate than current methods. Even though text-driven approaches are no longer popular, they are still widely used to deal with OOV with a lexicon-driven approach.

Tokenization. The classical lexicon-driven segmentation model, described in [1] and is still adopted in many recent works. Segmentation is typically divided into two stages: dictionary look up and OOV word identification. This approach requires comparing and matching tens of thousands of dictionary entries in addition to guessing a good number of OOV words. In other words, it has a 10^4 x 10^4 scale mapping problem with unavoidable data sparseness. This model also has the unavoidable problem of overlapping ambiguity where e.g. a string [C_{i-1}, C_i, C_{i+1}] contains multiple sub-strings, such as [C_{i-1}, C_i] and [C_i, C_{i+1}], which are entries in the

dictionary. The degree of such ambiguities is estimated to fall between 5% to 20% [2, 6].

Character classification. Character classification or tagging, first proposed in [14], became a very popular approach recently since it is proved to be very effective in addressing problems of scalability and data sparseness [14, 4, 16, 17]. Since it tries to model the possible position of a character in a word as character-strings, it is still lexicon-driven. This approach has been also successfully applied by to name entity resolution, e.g. [17]. This approach is closely related to the adoption of the machine learning algorithm of conditional random field (CRF), [7]. CRF has been shown [11] to be optimal algorithm for sequence classification. The major disadvantages are big memory and computational time requirement.

3. Model

Our approach is based on a simplified idea of Chinese text, which we have introduced earlier in [15]. Chinese text can be formalized as a sequence of characters and intervals as illustrated in Figure 1.

$$c_1, I_1, c_2, I_2 \ldots, c_{n-1}, I_{n-1}, c_n$$

Figure 1: Chinese text formalization

There is no indication of word boundaries in Chinese text, only string of characters c_i. Characters in this string can be conceived as being separated by interval I_i. To obtain a segmented text, i.e. a text where individual words are delimited by some graphical mark such as space, we need to identify which of these intervals are to be replaced by such word delimiter.

We can introduce a utility notion of imaginary intervals between characters, which we formally classify into two types:

Type 0: a character boundary (CB) is an imaginary boundary between two characters

Type 1: a word boundary (WB), an interval separating two words.

With such a formulation, segmentation task can be easily defined as a classification task and machine learning algorithms can be employed to solve it. For conventional machine learning algorithms, classifications are made based on a set of features, which identify certain properties of the target to be classified.

In a segmented text, all the intervals between characters are labeled as a word boundary or as

a character boundary, however, characters are not considered as being part of any particular word. Their sole function is to act as a contextual aid for identification of the most probable interval label. Since the intervals between characters (be it a word boundary or a character boundary) don't carry any information at all, we need to rely on the information provided by group of characters surrounding them.

Now we can collect n-grams that will provide data for construction of features that will provide learning basis for machine learning algorithm. A sequence, such the one illustrated in Figure 1, can be obtained from segmented corpus, and hence the probability of word boundary with specified relation to each n-gram may be derived. The resulting table which consists of each distinct n-gram entry observed in the corpus and the probability of a word

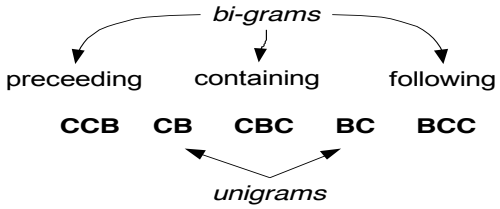
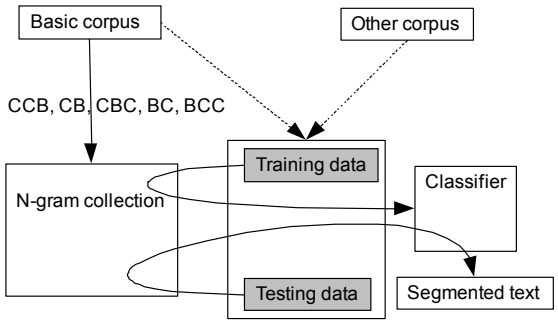


Figure 2: The feature vectors used in this study. While C denotes a character in the sequence, B indicates the imaginary boundary. Thus CBC denotes a bi-gram containing the interval.

boundary defines our n-gram collection.

Figure 2 shows the format of the feature vectors, or interval vectors, used in this study. We build the n-gram model up to n = 2.

To allow for a more fine-grained statistical information we have decomposed an interval



surrounding context into two unigrams and three bi-grams. For convenience, we can define each interval by the two characters that surround it. Then, for each interval <*b,c*> in a 4-character window *abcd* we collect two unigrams *b* and *c* and three bi-grams *ab, bc, cd* and compute probability of that interval being a word boundary. These five n-grams are stored in a vector, which is labeled as Type 0 (character boundary) or Type 1 (word boundary): <*ab, b, bc, c, cb*, 0> or <*ab, b, bc, c, cb*, 1>. An example of an encoding of a sample from the beginning of Bakeoff 3 AS training corpus: "時間：三月十日" (shijian:sanyueshiri), which

AB	B	BC	C	CD	Typ.	Inter.
0.500	0.595	0.003	0.173	0.021	0	時間
0.983	0.958	1.000	0.998	1.000	1	間：
1.000	0.998	1.000	0.713	0.994	1	：三
0.301	0.539	0.010	0.318	0.054	0	三月
0.964	0.852	1.000	0.426	0.468	1	月十
0.002	0.245	0.065	0.490	0.010	0	十日

Table 1: Example of encoding and labeling of interval vectors in a 4-character window ABCD

would be correctly segmented as "時間 ： 三月十日" (shijian : sanyue shiri) can be seen in Table 1.

Set of such interval vectors provides a training corpus on which we apply machine learning algorithm, in our case logarithmic regression. Unsegmented text is prepared in the same fashion and the interval vectors are subsequently labeled by a classifier.

4. Training the Machine Learning Model

It is our goal to develop a segmentation system that would be able handle different types of

text. A large uniform training corpus is desirable for high precision of segmentation, but that would cause a specialization of the classifier to types of texts contained in the corpus and system's generality would be compromised.

Furthermore, using a training data set converted from an independent corpus may give supplementary information and provide certain adaptation mechanism for the classifier during training, but leave the basic n-gram collection untouched. However, a smaller set of

No of vectors	LogReg	LDA	NNet	SVM
17,577,301	0.9857	0.9784	0.9865	0.9862
1,000,000	0.9862	0.9796	0.9881	0.9876
100,000	0.9856	0.9796	0.9844	0.9867
10,000	0.9872	0.9811	0.9892	0.9879
1,000	0.9910	0.9820	0.9940	0.9920
100	1.0000	0.9700	1.0000	0.9900

Table 2: Performance during training

training data may give similar performance but with much lower cost.

If the features in the n-gram collection are properly defined, the final results from different machine learning algorithms may not differ too much. On the contrary, if the available n-gram collection does not provide efficient information, classifiers with ability to adjust the feature space may be necessary.

In our preliminary tests, during which we wanted to decide which machine learning algorithm would be most appropriate, the Academia Sinica Balance Corpus (ASBC) is used for the derivation of the n-gram collection and training data. The CityU corpus from the SigHAN Bakeoff2 collection is used for testing.

In order to verify the effect of the size of the training data, the full ASBC (~17 million intervals) and a subset of it (1 million randomly selected intervals) are used for training separately.

No of vectors	LogReg	LDA	NNet	SVM
17,577,301	0.9386	0.9326	0.9373	0.9362
1,000,000	0.9386	0.9325	0.9360	0.9359
100,000	0.9389	0.9326	0.9331	0.9369
10,000	0.9393	0.9326	0.9338	0.9364
1,000	0.9373	0.9330	0.9334	0.9366
100	0.9106	0.9355	0.9198	0.9386

Table 3: Performance during testing

Furthermore, four different classifiers, i.e., logistic regression (LogReg) [9], linear discriminative analysis (LDA)[13], multi-layer perceptron (NNET)[13], and support vector machine (SVM)[3], were tested.

The segmentation results are compared with the "gold standard" provided by the SigHAN Bakeoff2.

No of vectors	LogReg	LDA	NNet	SVM
17,577,301	0.9386	0.9326	0.9373	0.9362
1,000,000	0.9386	0.9325	0.9360	0.9359
100,000	0.9389	0.9326	0.9331	0.9369
10,000	0.9393	0.9326	0.9338	0.9364
1,000	0.9373	0.9330	0.9334	0.9366
100	0.9106	0.9355	0.9198	0.9386

Table 4: Performance during training: new corpus

Tables 2 and 3 show the training and testing accuracies of various classifiers trained with the ASBC. All classifiers tested perform as expected, with their training errors increase with the size of the training data, and the testing errors decrease with it. Table 2 clearly shows that the training data size has little effect on the testing error while it is above 1000. This proves that once a sufficient n-gram collection is provided for preparation of the interval vectors, classifier can be trained with little input.

It is also shown in Table 2 that four classifiers give similar performance when the training data size is above 1000. However, while the training sample size drops to 100, the SVM and LDA algorithms show their strength by giving similar performance to the experiments trained with larger training data sets.

No of vectors	LogReg	LDA	NNet	SVM
1,000,000	0.9424	0.9390	0.9423	0.9443
100,000	0.9425	0.9387	0.9417	0.9441
10,000	0.9421	0.9410	0.9409	0.9430
1,000	0.9419	0.9418	0.9332	0.9400
100	0.8857	0.9350	0.8812	0.9299

Table 5: Performance during testing: new corpus

To further explore the effectiveness of our approach, we have modified the experiment to show the performance in model adaptation. In the modified experiments the training and testing data sets are both taken from a foreign corpus (CityU), while our n-gram collection is still from ASBC. The relation between the derived features and the true segmentation may be different from the ASBC, and hence is learned by the classifiers. The results of the modified experiments are shown in Tables 4 and 5.

5. Results

In our test to compare our performance objectively with other approaches, we adopt logarithmic regression as our learning algorithm as it yielded best results during our test. We apply the segmentation system to two traditional Chinese corpora, CKIP and CityU, provided for SigHAN Bakeoff 3. In the first set of tests, we used training corpora provided by SigHAN Bakeoff3 for n-gram collection, training and testing. Results of these tests are presented in Table 6.

	cityu	ckip
F-measure	0.933	0.919
OOV Rate	0.179	0.204
OOV Recall Rate	0.888	0.871
IV Recall Rate	0.941	0.943

Table 7: Results (Bakeoff 3 dataset): traditional Chinese

	cityu	ckip
F-measure	0.920	0.925
OOV Rate	0.167	0.187
OOV Recall Rate	0.920	0.893
IV Recall Rate	0.920	0.930

Table 6: Combined results (Bakeoff 3 dataset): traditional Chinese

In addition, to underline the adaptability of this approach, we also tried combining both corpora and then ran training on random sample of vectors. This set of tests is designed to exclude the possibility of over-fitting and to underline the robustness of the WBD model. Note that such tests are not performed in SigHAN Bakeoffs as many of the best performances are likely over-fitted. Results of this test are shown in Table 7.

Table 6 and 7 show that our OOV recall is comparable with our overall F-score, especially when our system is trained on selected vectors from combined corpus. This is in direct contrast with all existing systems, which typically has a much lower OOV recall than IV recall. In other words, our approach applies robustly to all textual variations with reliably good results. Table 8 shows that indeed our OOV recall rate shows over 16% improvement

over the best Bakeoff3 result for CityU, and over 27% improvement over best result for CKIP data.

	ckip	cityu
Microsoft Research Asia	0.702	0.792
IASL	0.656	0.792
Respective corpus	**0.888**	**0.871**
Combined corpora	**0.893**	**0.920**

Table 8: Our OOV recall results compared to

6. Discussion

We refined and implemented the WBD model for Chinese word segmentation and show that it is a robust and realistic model for Chinese language technology. Most crucially, we show that the WBD model is able to reconcile the two competitive goals of the lexicon-driven and text-driven approaches. The WBD model maintains comparable F-score level with the most recent CRF character-classification based results, yet improves substantially on the OOV recall.

We showed that our system is robust and not over-fitted to a particular corpus, as it yields comparable and reliable results for both OOV and IV words. In addition, we show that same level of consistently high results can be achieved across different text sources. Our results show that Chinese word segmentation system can be quite efficient even when using very simple model and simple set of features.

Our current system, which has not been optimized for speed, is able to segment text in less then 50 seconds. Time measurement includes preparation of testing data, but also training phase. We believe that with optimized and linked computing power, it will be easy to implement a real time application system based on our model. In the training stage, we have shown that sampling of around 1,000 vectors is enough to yield one of the best results. Again, this is a promise fact for the WBD model of segmentation to be robust. It is notable, that in case of training on combined corpora (CKIP and CityU) the results are even better than test in respective data sets, i.e. CKIP training corpus for segmenting CKIP testing text, or CityU respectively. This is undoubtedly the result of our strategy of granulation of the context around each interval. Since four characters that we use for representation of the interval context are broken up into two unigrams and three bi-grams, we let the system to get more

refined insight into the segmented area.

Consequently, the system is learning morphology of Chinese with greater generality and this results in higher OOV scores. It can be argued that in our combined corpora test, the OOV recall is even higher, because the input contains two different variants of Chinese language, Taiwanese variant contained in CKIP corpus and Hong Kong variant contained in CityU corpus.

Text preparation and post-processing also add to overall processing time. In our current results, apart from context vector preparation there was no other preprocessing employed and neither any post-processing. This fact also shows that our system is able to handle any type of input without the need to define special rules to pre- or post-process the text. Early results applying our model to simplified Chinese corpora are also promising.

In sum, our WBD model for Chinese word segmentation yields one of the truly robust and realistic segmentation program for language technology applications. If these experiments are treated as simulation, our results also support the linguistic hypothesis that word can be reliably discovered without a built-in/innate lexicon. We will look into developing a more complete model to allow for more explanatory account for domain specific shifts as well as for effective bootstrapping with some lexical seeds.

References

[1] K.J Chen and S.H. Liu, "Word Identification for Mandarin Chinese sentences", in *Proceedings of the 14th conference on Computational Linguistics,* pp.101-107, 1992.

[2] T.-H. Chiang, J.-S. Chang, M.-Y. Lin and K.-Y. Su, "Statistical Word Segmentation", in C.-R. Huang, K.-J. Chen and B.K. T'sou (eds.): *Journal of Chinese Linguistics*, Monograph Series, Number 9, Readings in Chinese Natural Language Processing, pp. 147-173, 1996.

[3] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer and A.Weingessel "e1071: Misc. Functions of the Department of Statistics (e1071)", TU Wien R package version 1.5-17., 2007.

[4] J. Gao, A. Wu, M. Li, C.-N. Huang, H. Li, X. Xia and H.Qin, "Adaptive Chinese Word Segmentation", in *Proceedings of ACL-2004*, 2004.

[5] C.-N. Huang and H. Zhao, “Which Is Essential for Chinese Word Segmentation: Character versus Word”, *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20)*, pp.1-12, 2006.

[6] H. Meng and C. W. Ip, “An Analytical Study of Transformational Tagging for Chinese Text”, in *Proceedings of ROCLING XII*, pp. 101-122, 1999.

[7] J.D. Lafferty, A. McCallum and F.C.N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, in *ICML 2001*, pp 282–289, 2001.

[8] G.-A. Levow, “The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition”, in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Association for Computational Linguistics, pp.108–117, 2006.

[9] R Development Core Team, “R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing”, Vienna, Austria, 2008.

[10] M. Redington, N. Chater, C. Huang, L. Chang and K. Chen, “The Universality of Simple Distributional Methods: Identifying Syntactic Categories in Mandarin Chinese”, in *Proceedings of the International Conference on Cognitive Science and Natural Language Processing*, 1995.

[11] B. Rosenfeld, R. Feldman and M. Fresko, “A systematic cross-comparison of sequence classifiers”, in *SDM 2006*, 2006.

[12] R. Sproat, C. Shih, W. Gale and N. Chang, “A Stochastic Finite-State Word-Segmentation Algorithm for Chinese”, *Computational Linguistics*, 22(3) pp. 377-404, 1997.

[13] W. N. Venables and B. D. Ripley, “Modern Applied Statistics with S”, Fourth Edition, Springer New York, ISBN 0-387-95457-0, 2002.

[14] N. Xue, “Chinese Word Segmentation as Character Tagging”, *Computational Linguistics and Chinese Language Processing*. 8(1), pp. 29-48, 2003.

[15] C.-R. Huang, P. Šimon, S.-K. Hsieh and L. Prévot, “Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification”, *Proceedings of the 45th Annual Meeting of the Association for Computational*

Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics,pp. 69—72, Prague, Czech Republic

[16] H. Zhao, C.-N. Huang and M. Li, "An improved Chinese word segmentation system with conditional random field", in *SIGHAN-5*, pp 162–165, 2006.

[17] H. Zhao and C. Kit, "Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition", *The Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, pp.106-111, 2008.

國台語無聲調拼音輸入法實作

An Implementation of Toneless Input for Mandarin and Taiwanese

余明興 Ming-Shing Yu
國立中興大學資訊科學與工程學系
Dept. of Computer Science and Engineering
National Chung-Hsing University
Taichung, Taiwan, 40227
msyu@nchu.edu.tw

蔡承融 Cheng-Rong Tsai
國立中興大學資訊科學與工程學系
Dept. of Computer Science and Engineering
National Chung-Hsing University
Taichung, Taiwan, 40227
s9556010@cs.nchu.edu.tw

摘要

目前的市面上很少有支援輸入台語、客語、原住民語等語言的輸入法。在拼音方案百家爭鳴的環境下，以及減少使用者因聲調的關係而造成聲調輸入錯誤，我們的目標是要提供一個支援多種拼音方案且無聲調拼音的國語及台語的輸入法。本文使用連續三個詞長詞優先法，在國語無聲調音轉字得到 89.590%的正確率，混合國語多種拼音方案得到 88.854%的正確率。

ABSTRACT

There are very few input systems supporting Taiwanese and Hakka on the market at present. Our purpose is to provide an input system supporting Mandarin and Taiwanese which is toneless and complies with multiple types of phonetic symbols, we hope that such an approach can resolve the problems resulting from tone and phonetic symbols. In this paper,

we use an algorithm based on three continuous longest word first whose precision is 89.59% in one type of phonetic symbols, and 88.54% in the combination of many types of phonetic symbols.

關鍵詞：無聲調輸入法，音轉字，連續三個詞長詞優先法

Keywords: toneless input system, phoneme to character, three continuous longest word first.

一、前言

由於電腦及網際網路的普及化，使得愈來愈多人透過網站的方式來取得最新的知識；使用網路即時通訊軟體、電子信箱聯絡親朋好友，而這些都必須經由文字的輸入與電腦溝通，因此輸入法是一個重要的議題。 然而占台灣 73.3%人口的閩南人和 12%的客家人及 1.7%的原住民，在目前的市面上卻很少有支援其輸入台語、客語、原住民語等母語的輸入法，因此本文將針對人口比例最高的閩南族群開發出一個輸入法。

目前的拼音輸入法主要分爲注音符號及羅馬拼音兩種。 我們採用羅馬拼音，主因是利用二十六個英文字母，就可以正確的標示出國語、台語、客語及英語裡的所有字，另外全世界的電腦都可以輸入英文字母。

由於國語及台語的拼音方案百家爭鳴，沒有統一的版本，因此我們打算製作出多種拼音系統共存的輸入法。另外在國語及台語會因爲語意的關係而發生變調「如:老李 買好酒 和 老李 買好 酒」聲調分別爲「23323」及「23223」，台語也會因各地發音腔調的差異而產生聲調的不同，例如:「台南」一詞，北部調爲「dai3 lam5」，南部調爲「dai7 lam5」。爲了減少使用者的負擔，因此我們系統設計爲不用輸入聲調。

二、拼音方案概述

(一) 國語拼音方案簡介

「國語」是台灣地區以北京話爲基礎制定的中華民國國語，由於台灣已獨立發展了數十年，與北京話出現了差異，因此又稱爲「台灣國語」。以下將簡介台灣常見的國語羅馬字拼寫方案及說明其優缺點。

1.國語注音符號第二式

「國語注音符號第二式」簡稱「注音二式」，是台灣教育部爲了因應外籍人士學習中文的需求，在民國七十三年成立專案研究小組所制定。民國八十五年行政院經建會決議以國語注意符號第二式翻譯地名。注音二式有符合英語發音習慣的優點，但其缺乏國際性。

2.漢語拼音

漢語拼音是中國在西元 1958 年正式通過的漢語普通話拉丁轉寫統一規範，也是國際承認的標準 ISO 7098。漢語拼音中沒有台語和客語的設計，其中有些不合台語使用，例如「iu」用來代表ㄧㄡ。比較好的選擇是用「iu」來代表「優」（台語），而用來「iou」代表「優」（國語）。在這種情況下，「niu」（娘，台語）和「niou」（牛，國語）可以併存而且英語直覺性較佳。

3.通用拼音

通用拼音是民國八十七年台灣中研院副研究員余伯泉以 KK 音標爲基礎而發展出的拼音系統。它盡量和漢語拼音相容，而且能同時標注國語和台語。去除漢語拼音中不符合英語讀的習慣(x、q)。台北市政府研發出國語、台語及客語三種通用的通用拼音。民國八十九年台灣教育部宣佈使用，取代國語注音符號第二式及威妥瑪拼音。另外自民國九十一年起台灣行政院全面推行以通用拼音爲基礎的中文譯音政策。

4.美式拼音

美式拼音是中興大學語音與語言實驗室[9]所提出的拼音方案，它是依照通用拼音來修改的，修改之處如下。「chi」取代通用拼音的「ci」來標示ㄑ。「tz」取代通用拼音的「z」來標示ㄗ，用「ts」取代通用拼音的「c」來標示ㄘ，這些取代都是注音二式早已經發展出來，且是我們認爲比較符合美式英語的發音方式。用「chi」取代「ci」是因爲「ci」在美語中較常讀成「si」。用「tz」取代「z」之後，「z」可以在台語中使用，例如「zin」可拼「人」的音。表一是注音二式、漢語拼音、通用拼音及美式拼音的子音和母音對照表。

表一、國語拼音對照表。

注音符號	注音二式	漢語拼音	通用拼音	美式拼音	注音符號	注音二式	漢語拼音	通用拼音	美式拼音
ㄅ	b	b	b	b	ㄗ	tz	z	z	tz
ㄆ	p	p	p	p	ㄘ	ts	c	c	ts
ㄇ	m	m	m	m	ㄙ	s	s	s	s
ㄈ	f	f	f	f	ㄚ	a	a	a	a
ㄉ	d	d	d	d	ㄛ	o	o	o	o
ㄊ	t	t	t	t	ㄜ	e	e	e	e
ㄋ	n	n	n	n	ㄝ	ie	ie	ie	ie
ㄌ	l	l	l	l	ㄞ	ai	ai	ai	ai
ㄍ	g	g	g	g	ㄟ	ei	ei	ei	ei
ㄎ	k	k	k	k	ㄠ	au	ao	ao	ao
ㄏ	h	h	h	h	ㄡ	ou	ou	ou	ou
ㄐ	j(i)	j	j(i / y-)	j	ㄢ	an	an	an	an
ㄑ	ch(i)	q	c(i / y-)	ch(i)	ㄣ	en	en	en	en
ㄒ	sh(i)	x	s(i / y-)	s	ㄤ	ang	ang	ang	ang
ㄓ	j	zh	jh	jh	ㄥ	eng	eng	eng	eng
ㄔ	ch	ch	ch	ch	ㄦ	er	er	er	er
ㄕ	sh	sh	sh	sh	ㄧ	i,yi	i,yi	i,yi	i,yi
ㄖ	r	r	r	r	ㄨ	u,wu	u,wu	u,wu	u,wu
空韻	r,z	i	ih	ii	ㄩ	iu,yu	ü,yu	yu	yu

(二) 台語拼音方案簡介

「台語」爲臺灣地區所使用的閩南語，又可稱爲「台灣話」、「福佬話」、「河洛話」、「台灣閩南語」，是目前在台灣使用最多的語言之一。其由閩南人在明朝末年開始渡台，歷經了荷蘭、日本的統治，發展出獨特的閩南語。在所有的閩南語中，台灣閩南語與廈門話最爲接近，主要的差異是在詞彙方面，據王育德、鄭良偉等人指出約有百分之十的不同。在本章節將簡介目前常用的台語拼寫方案。

1.教會羅馬字

「教會羅馬字」簡稱「教羅」，是 19 世紀時由基督教長老教會所創造，是以拉丁字母書寫的閩南語正字法，又稱可爲「白話字」。

2.臺灣閩南語羅馬字拼音

臺灣閩南語羅馬字拼音簡稱「台羅拼音」，是台灣官方的閩南語拼音方案，它是整合原有的台灣閩南語音標(TLPA)及教會羅馬字而成的閩南語羅馬字拼音。

3.通用拼音

通用拼音和台羅拼音最大的不同有兩點。第一點是通用拼音有一併考慮到國語和客語，而台羅拼音只考慮到台語。第二點是關於ㄅ、ㄉ、ㄍ的拼音選擇。通用拼音使用 b、d、g 來標註，較合美式英語；而台羅拼音使用 p、t、k 來標註，較像國際音標。

4.美式拼音

除了前面的描述之外，美式拼音將國語ㄓ、ㄔ、ㄕ、ㄖ、ㄗ、ㄘ、ㄙ的空韻由加上「ih」改成加上「ii」。原因是「h」結尾的音爲台語的入聲音，可能會有衝突發生。例如 cih(ㄘ)和 sih(ㄙ)也是台語的入聲音。台羅、通用和美式拼音的子音對照表如附錄一所示。

三、音轉字處理

音轉字的輸入方式可分爲二種，一種是透過鍵盤輸入，另一種採用語音的輸入方式，而它們的主要問題是如何從一音多字裡選擇正確的字。在國語中字的音節有四百多個，而國語的字卻有一萬個以上，台語字的音節有七百多個，台語字也有八千個以上，平均一個音節會對應到一、二十個字。以下我們將探討此問題。

在處理中文音轉字的問題中，目前常見的方法有規則法及統計法。

- 規則法：從語言學中訂出規則，依據所訂的規則判斷出合理的結果，其缺點是需要大量的專業人士參與。

◆ 統計法：其中 N-Gram 語言模型[3]是目前最常用的方法，使用語料庫來訓練語言模型得到字、詞或詞性間的關係。

以下將介紹我們處理音轉字的方式。

(一) 國語音轉字

由於我們是要實作一個輸入法，所以不希望系統修改太久以前輸入的資訊照成使用者的困惱，因此我們的組字視窗限制在十二個字內。系統只會修改組字視窗內的字，當組字視窗超過十二個字時，系統則會自動輸出第一個辭彙。在組字視窗內，我們依連續三個詞的長詞優先演算法[6][15]找出合理的結果。當連續三個詞的長詞優先演算法找出的結果有二組以上時，我們實驗二種方式來音轉字的計算分數，分別爲公式 1 及公式 2。W_m 是候選詞組的詞彙，$P(W_m)$是詞的機率。

$$T_1 = \prod_{m=1}^{n} P(W_m) \qquad (1)$$

$$T_2 = \prod_{m=1}^{n} P(W_m) * \prod_{m=2}^{n} P(W_m \mid W_{m-1}) \qquad (2)$$

(二) 台語音轉字

我們的台語音轉字採用上節所提的連續三個詞的長詞優先法，台語輸入採用長詞優先法的好處是較不需要訓練語料。台語輸入有一個要注意的情況是台語音 $S = S_1,S_2,\ldots,S_N$ (N 是總音節數，S_i 是第 i 個音節)對應至台語文 $X = X_1,X_2,\ldots,X_T$ (T 是總字數，X_i 是第 i 個台語字)時，N 與 T 的長度不一定是相同。在考量各種情況下，我們系統優先輸出音節數與辭彙字數相同的辭彙，使用者如果想要字數與音節數不相同的詞彙可以在修改模式中去選取。

(三) 智慧型處理

由於國語及台語的拼音方案眾多，因此對於拼音系統不熟悉的初學者往往會混合著不同的拼音方案輸入；所以我們希望系統能給與多種拼音方案相容的方式，以減少初學者拼音的錯誤率。我們採取的方法爲當使用者輸入一個音串，系統會去評估使用者是輸入那一種拼音方案的那個音節，例如:使用者輸入「cyuan」音串，系統就評估它是輸入

通用拼音方案的「ㄑㄩㄢ」音節。

然而有些音串對應到的音節不只一種，例如:「niu」音串，可以對應到漢語拼音的「ㄋㄧㄡ」音節及注音二式的「ㄋㄩ」音節，表二爲國語拼音系統互相衝突的地方。對於此狀況，我們將可能的音節都送入至系統中，然後再依上下文估算合理的結果，例如:「liu ju mei guo yang ji duei wang jian min jhu tou liu ju wu shih fen , liu ju jin ji guan jyun jhan de si wang」，這一句中「liu」音串可以對應到漢語拼音的「ㄌㄧㄡ」音節及注音二式的「ㄌㄩ」音節，「ju」音串可以對應到漢語拼音的「ㄐㄩ」音節及注音二式的「ㄓㄨ」音節。我們依上下文推斷合理的結果爲「旅居美國洋基隊王建民主投六局無失分,留住晉級冠軍戰的希望」。

表二、國語拼音系統衝突表

音串	漢語拼音	注音二式
niu	ㄋㄧㄡ	ㄋㄩ
liu	ㄌㄧㄡ	ㄌㄩ
jiu	ㄐㄧㄡ	ㄐㄩ
ju	ㄐㄩ	ㄓㄨ
juan	ㄐㄩㄢ	ㄓㄨㄢ
chi	ㄔ	ㄑㄧ
shi	ㄕ	ㄒㄧ

四、實驗與討論

在本章節中將說明我們訓練、測試語料庫和辭典的來源，及實驗結果的數據。

(一)語料庫

我們的國語語料庫來源有兩個。第一個是中央研究院平衡語料庫[17]，這是個約有五百萬詞的語料。另外一個是我們實驗室從各新聞網站上蒐集而來的新聞語料庫，這是一個含有約一千六百萬個詞(二千八百萬個字)的語料庫。

其中我們利用中央研究院的中文斷詞器[16]來訓練詞的 Bigram 模型，以及擷取約百分之十的語料庫，再透過[11]中文字轉音程式得到測試語料庫。

(二)辭典

在音轉字的過程中，辭典扮演著一個非常重要的角色，一個不錯的辭典可以提高正確率。在國語部份，我們的辭典來源有兩個。辭典 1 是論文[13]中使用的辭典，其辭典是以中研院八萬目詞爲基礎，另外再從中研院平衡語料庫[17]抽取出未在八萬目詞內的詞，一共約 13 萬詞的辭典。辭典 2 是[11]從文章中擷取出約 44 萬個中文的常用字串，這些字串有對應的音，但沒有詞性可用。

在台語辭典部份，我們使用[10]所整理的辭典爲基礎，其辭典約有 6 萬詞。

(三)實驗結果

1.實驗一

首先我們將實驗加入中文常用字串[11]對國語音轉字的影響，音轉字的演算法如上節所提的連續三個詞長詞優先詞。 由實驗結果表三得知單獨使用[11]中文常用字串的辭典 2 就可以得到 87%的正確率，另外合併這二個辭典正確率可以提升至 88%。因此我們使用辭典 3 作爲我們國語音轉字的辭典，後面的實驗也皆使用辭典 3。

表三、實驗一的結果

辭典	正確率
辭典 1	77.622%
辭典 2	87.500%
辭典 3	88.459%

註: 辭典 1 :[13]論文中使用的 13 萬詞辭典。

辭典 2 :[11]從文章中擷取出約 44 萬個中文的常用字串。

辭典 3 :合併辭典 1 與辭典 2。

2.實驗二

實驗 3.1 節所提的二種方法在國語音轉字的正確率，由實驗結果表四得知有加入 Bigram 的資訊的方法 2 所得到的正確率略高於只用到 Uigram 資訊的方法 1。

表四、實驗二的結果

方法	正確率
方法 1	88.459%
方法 2	89.590%

註: 方法 1 ： 候選詞組不只一組時，使用公式 1。

方法 2 ： 候選詞組不只一組時，使用公式 2。

3.實驗三

最後我們將實驗在混合通用、漢語、注音二式及美式四種拼音方案時，國語音轉字的正確率。由實驗結果表五得知有加入 Bigram 的資訊的方法 2 所得到的正確率略高於只用到 Uigram 資訊的方法 1。另外這二種方法音轉字的正確率只略低一點點於無混合拼音方案的情況。

表五、實驗三的結果

方法	正確率
方法 1	87.879%
方法 2	88.854%

註: 方法 1 ： 候選詞組不只一組時，使用公式 1。

方法 2 ： 候選詞組不只一組時，使用公式 2。

(四)實驗討論

根據實驗 1 的結果，得知合併辭典 1 及辭典 2 的辭典 3 可以得到 88%國語音轉字的正確率，遠高於辭典 1 的 77%正確率，推測原因可能是大部份的句子中都包含了常用的字串。因此我們的系統使用辭典 3 作爲國語音轉字的辭典。

在實驗 2 的結果，發現加入 Bigram 資訊的方法 2 正確率有 89.590%，比方法 1 提昇 1%的正確率。在實驗 3 的結果得知我們的方法在混合通用、漢語、注音二式及美式四種拼音方案的情況下，國語音轉字的正確率皆有 87%以上。由於所訓練的 Bigram 資訊有約 120 萬筆，檔案龐大且正確率只提昇一點點，因此我們的音轉字採用方法 1。

五、系統設計考量

由於目前台語並沒有像中文一樣有正式且統一的文字，因此使用者對於系統的台語辭彙可能會不滿意，所以我們提供台語音對應至國語辭彙的功能。例如：若是使用者不滿意台語音「bhin-a-am」對應至台語詞「明仔暗」，可透過修改的方式，將台語音「bhin-a-am」對應至國語詞「明天晚上」或「明晚」。

當使用者選擇將台語音對應至國語辭彙時，我們的系統將會改變使用者原始輸入的音串，把使用者輸入的台語音改成國語辭彙可能的台語音(音節數與辭彙字數相同)。例如：使用者選擇將台語音「bhin-a-am」對應至國語詞「明晚」，我們將會把使用者輸入的台語音「bhin-a-am」轉成國語詞「明晚」最有可能的台語音「bhin- bhuan」。

六、結論與未來改進方向

雖然本論文完成了多種拼音方案相容的國台語無聲調拼音輸入法，但仍有些部份可以改進。在台語音轉字的部份由於欠缺語料庫，因此我們打算收集使用者輸入的語料，以改善台語音轉字的正確率。在國語音轉字雖然達到近九成的正確率，但未來我們可以加入構詞器及收集使用者輸入的語料，來提昇音轉字的正確率。另外我們希望未來能提供國語、台語及英語混合輸入的功能，讓使用者不必透過切換，就可以同時輸入國語、台語及英語這三種語言。

參考文獻

[1] Amelia-Fong Lochovsky and Hon-Kit Cheung, “N-gram Estimates in Probabilistic Models for Pinyin to Hanzi Transcription”, IEEE International Conference on Intelligent Processing Systems, Beijing, 1997, pp. 1798-1803.

[2] Bing-Quan Liu and Xiao-Long Wang, "An Approach to Machine Learning of Chinese Pinyin-to-Character Conversion for Small-Memory Application", Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 2002, pp. 1287-1291.

[3] Frederick Jelinek, “Statistical Methods for Speech Recognition”, The MIT Press, Cambridge Massachusetts, 1997.

[4] Shuanfan Huang, “Language, Society, and Ethnic Identity (語言、社會、與族群意識)，Taipei Crane，1993。

[5] OpenVanilla，http://openvanilla.org/index-zh_TW.php。

[6] T. H. Ho, K. C. Wang, J. S. Lin, and L. S. Lee, ”Integrating Long-Distance Language Modeling to Phoneme-to-Character Conversion, ”Proceeding of ROCLING X, pp. 287-292, 1997.

[7] Xiaolong Wang, Qingcai Chen, and Daniel S. Yeung, “Mining Pinyin-to-Character Conversion Rules from Large-Scale Corpus: A Rough Set Approach”, IEEE Transactions on System, Man, and Cybernetics, 2004, pp. 834-844.

[8] Xuan Wang, Lu Li, Lin Yao, and Waqas Anwar, “A Maximum Entropy Approach to Chinese Pinyin-to-Character Conversion”, IEEE International Conference on Systems, Man, and Cybernetics, Taipei, 2006, pp. 2956-2959.

[9] 余明興，“台灣共通語言”，第十九屆自然語言與語音處理研討會，2007，pp. 319-333.

[10] 蔡宗謀，“中文文句轉台語語音系統初步研究”，中興大學資訊科學與工程研究所碩士論文，2008。

[11] 林義証，“中文常用字串-一個優於傳統語言模型的新觀念”，中興大學應用數學系

博士論文，2002。

[12]林嘉信，“與多種拼音方法相容的國語輸入系統”，中興大學應用數學研究所資訊組碩士論文，2002。

[13]張唐瑜，“以大量詞彙作爲合成單元的中文文轉音系統”，中興大學資訊科學研究所碩士論文，2005。

[14]許聞廉與陳克健，“自然智慧型輸入系統的語意分析脈絡會意法”，台灣中央研究院資訊所，1993。

[15]羅火嵐，“中文無聲調拼音輸入法及其實作”，中興大學資訊科學研究所碩士論文，2006。

[16]中央研究院中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw。

[17]中央研究院平衡語料庫，http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh。

[18]世界台灣語通用協會，中研研究院， http://abc.iis.sinica.edu.tw/。

[19]國語注音符號第二式，教育部，
http://www.edu.tw/files/site_content/M0001/er/cmain.htm。

[20]漢語拼音方案，中國教育和科研計算機網，
http://www.edu.cn/20011114/3009777.shtml。

附錄一. 國台語拼音的子音表。台語羅馬字只列出教育部2006年公佈的部分。

注音符號	美式拼音	通用乙式	漢語拼音	注音二式	台語羅馬字
ㄅ	b	b	b	b	p
台帽	bh	bh			b
ㄆ	p	p	p	p	ph
ㄇ	m	m	m	m	m
ㄈ	f	f	f	f	
ㄉ	d	d	d	d	t
ㄊ	t	t	t	t	th
ㄋ	n	n	n	n	n
ㄌ	l	l	l	l	l
ㄍ	g	g	g	g	k
台鵝	gh	gh			g
ㄎ	k	k	k	k	kh
ㄏ	h	h	h	h	h
ㄐ	ji	ji	j	ji	
ㄑ	chi	ci	q	chi	
ㄒ	si	si	x	shi	
ㄓ	jh	jh	zh	j	
ㄔ	ch	ch	ch	ch	
ㄕ	sh	sh	sh	sh	
ㄖ	r	r	r	r	
ㄗ	tz	z	z	tz	ts
ㄘ	ts	c	c	ts	tsh
ㄙ	s	s	s	s	s
台字人如	z	zz			j
零韻	-ii	-ih	-i	-ih	
台姆(台□)	m(mh)	m(mh)			
台秧(台嗯)	ng(ngh)	ng(ngh)			ng

Propositional Term Extraction over Short Text using Word Cohesiveness and Conditional Random Fields with Multi-Level Features

張如瑩　Ru-Yng Chang
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
ruyng@csie.ncku.edu.tw

吳宗憲　Chung-Hsien Wu
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
chwu@csie.ncku.edu.tw

Abstract

Propositional terms in a research abstract (RA) generally convey the most important information for readers to quickly glean the contribution of a research article. This paper considers propositional term extraction from RAs as a sequence labeling task using the IOB (Inside, Outside, Beginning) encoding scheme. In this study, conditional random fields (CRFs) are used to initially detect the propositional terms, and the combined association measure (CAM) is applied to further adjust the term boundaries. This method can extract beyond simply NP-based propositional terms by combining multi-level features and inner lexical cohesion. Experimental results show that CRFs can significantly increase the recall rate of imperfect boundary term extraction and the CAM can further effectively improve the term boundaries.

摘要

命題術語(Propositional Term)表達文章中重要概念且引導讀者文章脈絡之發展。這篇論文以學術論文摘要爲實驗對象進行命題術語擷取，研究中整合條件隨機域(Conditional Random Fields, CRFs) 以及結合聯繫測量(Combined Association Measure, CAM) 兩種方法，考量詞彙內部凝聚力和文脈兩大類訊息，截取出的命題術語不再侷限於名詞片語型態，且可由單詞或多詞所構成。在命題術語擷取的過程中，將其視爲一種序列資料標籤的任務，並利用 IOB 編碼方式識別命題述語的邊界，CRF 考量多層次構成命題述語的特徵，負責初步命題術語偵測，再利用 CAM 計算詞彙凝聚力，藉以加強確認命題術語詞彙的邊界。實驗結果顯示 ，本研究所提出的方法比以往述語偵測方法在效能上有明顯增進，其中，CRF 明顯增進非完美術語詞彙邊界辨識(Imperfect hits)的召回率，而 CAM 則有效修正術語詞彙邊界。

Keywords: Propositional Term Extraction, Conditional Random Fields, Combined Association Measure, Multi-Level Feature

關鍵詞：命題述語擷取，條件隨機域，結合聯繫測量，多層次特徵

1. Introduction

Researchers generally review Research Abstracts (RAs) to quickly track recent research trends. However, many non-native speakers experience difficulties in writing and reading RAs [1]. The author-defined keywords and categories of the research articles currently utilized to provide researchers with access to content guiding information are cursory and general. Therefore, developing a propositional term extraction system is an attempt to exploit the linguistic evidence and other characteristics of RAs to achieve efficient paper comprehension. Other applications of the proposed method contain sentence extension, text generation, and content summarization.

A term is a linguistic representation of a concept with a specific meaning in a particular field. It may be composed of a single word (called a simple term), or several words (a multiword term) [2]. A propositional term is a term that refers to the basic meaning of a sentence (the proposition) and helps to extend or control the development of ideas in a text. The main difference between a term and a propositional term is that a propositional term, which can guide the reader through the flow of the content, is determined by not only syntax or morphology but semantic information. Take RAs to illustrate the difference between a term and a propositional term. Cheng [3] indicted that a science RA is composed of background, manner, attribute, comparison and evaluation concepts. In Figure 1, the terms underlined are the propositional terms which convey the important information of the RA. In the clause "*we present one of the first robust LVCSR systems that use a syllable-level acoustic unit for LVCSR,*" the terms "*LVCSR systems*", "*syllable-level acoustic unit*" and "*LVCSR*" respectively represent the background, manner and background concepts of the research topic, and can thus be regarded as propositional terms in this RA. The background concepts can be identified by clues from the linguistic context, such as the phrases "*most…LVCSR systems*" and "*in the past decade*", which indicate the aspects of previous research on LVCSR. For the manner concept, contextual indicators such as the phrases "*present one of…*", "*that use*" and "*for LVCSR*" express the aspects of the methodology used in the research. Propositional terms may be composed of a variety of word forms and syntactic structures and thus may not only be NP-based, and therefore cannot be extracted by previous NP-based term extraction approaches.

Most large vocabulary continuous speech recognition (LVCSR) systems in the past decade have used a context-dependent (CD) phone as the fundamental acoustic unit. In this paper, we present one of the first robust LVCSR systems that use a syllable-level acoustic unit for LVCSR on telephone-bandwidth speech. This effort is motivated by the inherent limitations in phone-based approaches-namely the lack of an easy and efficient way for modeling long-term temporal dependencies. A syllable unit spans a longer time frame, typically three phones, thereby offering a more parsimonious framework for modeling pronunciation variation in spontaneous speech. We present encouraging results which show that a syllable-based system exceeds the performance of a comparable triphone system both in terms of word error rate (WER) and complexity. The WER of the best syllable system reported here is 49.1% on a standard SWITCHBOARD evaluation, a small improvement over the triphone system. We also report results on a much smaller recognition task, OGI Alphadigits, which was used to validate some of the benefits syllables offer over triphones. The syllable-based system exceeds the performance of the triphone system by nearly 20%, an impressive accomplishment since the alphadigits application consists mostly of phone-level minimal pair distinctions.

Figure1. A Manually-Tagged Example of Propositional Terms in an RA

In the past, there were three main approaches to term extraction: linguistic [4], statistical [5, 6], and C/NC-value based [7,8] hybrid approaches. Most previous approaches can only achieve a good performance on a test article composed of a relatively large amount of words. Without the use of large amount of words, this study proposes a method for extracting and

weighting single- and multi-word propositional terms of varying syntactic structures.

2. System Design and Development

This research extracts the propositional terms beyond simply the NP-based propositional terms from the abstract of technical papers and then regards propositional term extraction as a sequence labeling task. To this end, this approach employs an IOB (Inside, Outside, Beginning) encoding scheme [9] to specify the propositional term boundaries, and conditional random fields (CRFs) [10] to combine arbitrary observation features to find the globally optimal term boundaries. The combined association measure (CAM) [11] is further adopted to modify the propositional term boundaries. In other words, this research not only considers the multi-level contextual information of an RA (such as word statistics, tense, morphology, syntax, semantics, sentence structure, and cue words) but also computes the lexical cohesion of word sequences to determine whether or not a propositional term is formed, since contextual information and lexical cohesion are two major factors for propositional term generation.

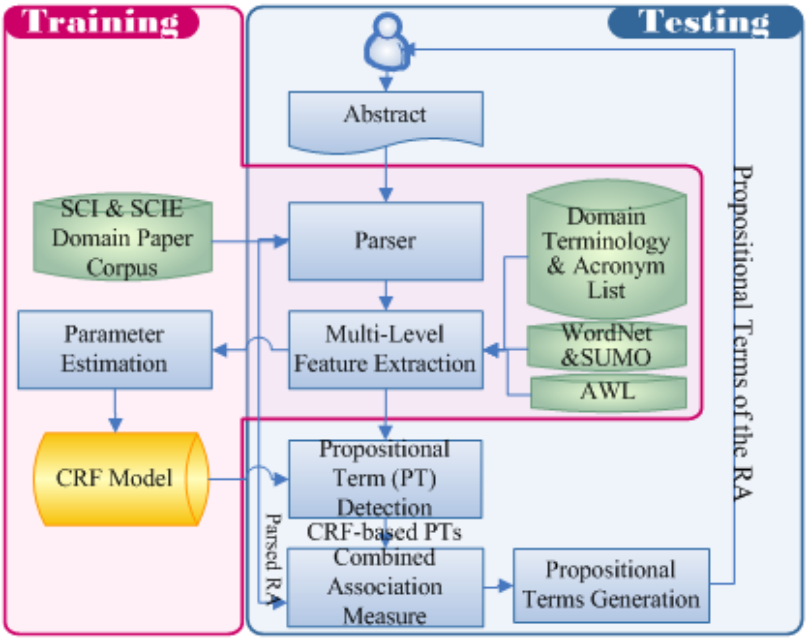


Figure 2. The System Framework of Propositional Term Extraction

The system framework essentially consists of a training phase and a test phase. In the training phase, the multi-level features were extracted from specific domain papers which were gathered from the SCI (Science Citation Index)-indexed and SCIE (Science Citation Index Expanded)-indexed databases. The specific domain papers are annotated by experts and then parsed. The feature extraction module collects statistical, syntactic, semantic and morphological level global and local features, and the parameter estimation module calculates conditional probabilities and optimal weights. The propositional term detection CRF model was built with feature extraction module and the parameter estimation module. During the test phase users can input an RA and obtain system feedback, i.e. the propositional terms of the RA. When the CRF model produces the preliminary candidate propositional terms, the propositional term generation module utilizes the combined association measure (CAM) to adjust the propositional term boundaries. The system framework proposed in this paper for RA propositional term extraction is shown in Figure 2. A more detailed discussion is presented in the following subsections.

2.1. Assisted Resource

In order to produce different levels of information and further assist feature extraction in the training and test phases, several resources were employed. This study chooses the ACM Computing Classification System (ACM CSS) [12] to serve as the domain terminology list for propositional term extraction from computer science RAs. The ACM CSS provides important subject descriptors for computer science, and was developed by the Association for Computing Machinery. The ACM CSS also provides a list of Implicit Subject Descriptors, which includes names of languages, people, and products in the field of computing. A mapping database, derived from WordNet (http://wordnet.princeton.edu/) and SUMO (Suggested Upper Merged Ontology) (http://ontology.teknowledge.com/) [13], supplies the semantic concept information of each word and the hierarchical concept information from the ontology. The AWL (Academic Words List) (http://www.vuw.ac.nz/lals/research/awl/) [14] is an academic word list containing 570 word families whose words are selected from different subjects. The syntactic level information of the RAs was obtained using Charniak's parser [15], which is a "maximum-entropy inspired" probabilistic generative model parser for English.

2.2. Conditional Random Fields (CRFs)

For this research goal, given a word sequence $W = \{w_1, w_2, ..., w_n\}$, the most likely propositional term label sequence $S = \{s_1, s_2, ..., s_n\}$ in the CRF framework with the set of weights Ψ can be obtained from the following equation.

$$\hat{S} = \arg\max_S P_\Psi (S \mid W) \tag{1}$$

A CRF is a conditional probability sequence as well as an undirected graphical model which defines a conditional distribution over the entire label sequence given the observation sequence. Unlike Maximum Entropy Markov Models (MEMMs), CRFs use an exponential model for the joint probability of the whole label sequence given the observation to solve the label bias problem. CRFs also have a conditional nature and model the real-world data depending on non-independent and interacting features of the observation sequence. A CRF allows the combination of overlapping, arbitrary and agglomerative observation features from both the past and future. The propositional terms extracted by CRFs are not restricted by syntactic variations or multiword forms and the global optimum is generated from different global and local contributor types.

The CRF consists of the observed input word sequence $W = \{w_1, w_2, ..., w_n\}$ and label state sequence $S = \{s_1, s_2, ..., s_n\}$ such that the expansion joint probability of a state label sequence given an observation word sequence can be written as

$$P(S \mid W) = \frac{1}{Z_0} \exp\left(\sum_t \sum_k \lambda_k f_k (s_{t-1}, s_t, W) + \sum_t \sum_k \mu_k g_k (s_t, W) \right) \tag{2}$$

where $f_k(s_{t-1}, s_t, W)$ are the transition features of the global observation sequence and the states at positions t and t-1 in the corresponding state sequence, and $g_k(s_t, W)$ is a state feature function of the label at position t and the observation sequence. Let λ_k be the weight of each f_k, μ_k be the weight of g_k and $\frac{1}{Z_0}$ be a normalization factor over all state sequences,

where $Z_0 = \sum_S \exp\left(\sum_t \sum_k \lambda_k f_k \left(s_{t-1}, s_t, W \right) + \sum_t \sum_k \mu_k g_k \left(s_t, W \right) \right)$.

The set of weights in a CRF model, $\Psi = (\lambda_k, \mu_k)$, is usually estimated by maximizing the conditional log-likelihood of the labeled sequences in the training data $D = \left\{ S^{(i)}, W^{(i)} \right\}_{i=1}^{n}$. (Equation (3)) For fast training, parameter estimation was based on L-BFGS (the limited-memory BFGS) algorithm, a quasi-Newton algorithm for large scale numerical optimization problems [16]. The L-BFGS had proved [17] that converges significantly faster than Improved Iterative Scaling (IIS) and General Iterative Scaling (GIS).

$$L_\Psi = \sum_{i=1\ldots N} \log\left(P_\Psi \left(S^{(i)} \mid W^{(i)} \right) \right) \tag{3}$$

After the CRF model is trained to maximize the conditional log-likelihood of a given training set P(S|W), the test phase finds the most likely sequence using the combination of forward Viterbi and backward A* search [18]. The forward Viterbi search makes the labeling task more efficient and the backward A* search finds the n-best probable labels.

2.3. Multi-Level Features

According to the properties of propositional term generation and the characteristics of the CRF feature function, this paper adopted local and global features which consider statistical, syntactic, semantic, morphological, and structural level information. In the CRF model, the features used were binary and were formed by instantiating templates, and the maximum entropy principle was provided for choosing the potential functions. Equation (4) shows an example of a feature function, which was set to 1 when the word was found in the rare words list (RW).

$$g_{s, w_1, w_2, \ldots, w_n} \left(s_t, w_1^n \right) = \begin{cases} 1, & \text{if } \mathrm{s_t} = s \cap isRW\left(\mathrm{W_t} \right) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

2.3.1. Local Feature

(1). Morphological Level:

Scientific terminology often ends with similar words, e.g. "*algorithm*" or "*model*", or is represented by connected words (CW) expressed with hyphenation, quotation marks or brackets. ACMCSS represents entries in the ACM Computing Classification System (ACM CSS). The last word of every entry in the ACM CSS (ACMCSSAff) satisfies the condition that it is a commonly occurring last word in scientific terminology. The existing propositional terms of the training data were the seeds of multiword terms (MTSeed).

Words identified as acronyms were stored as useful features, consisting of IsNenadic, IsISD, and IsUC. IsNenadic was defined using the methodology of Nenadić, Spasić and Ananiadou [19] to acquire possible acronyms of a word sequence that was extracted by the C/NC value method. IsISD refers to the list of Implicit Subject Descriptors in the ACM CCS and IsUC signifies that all characters of the word were uppercase

(2). Semantic Level:

MeasureConcept infers that the word was found under SUMO's

"UNITS-OF-MEASURE" concept subclass and SeedConcept denotes that the concept of the word corresponded to the concept of a propositional term in the training data.

(3). Frequency Level:

A high frequency word list (HF) was generated from the top 5 percent of words in the training data. A special words list (SW) consists of the out-of-vocabulary and rare words. Out-of-vocabulary words are those words that do not exist in WordNet. Rare words are words not appearing in the AWL or which appear in less than 5 different abstracts.

(4). Syntactic Level:

This feature was set to 1 if the syntactic pattern of the word sequence matched the regular expression "*(NP)*(preposition)?(NP)**" (SynPattern), or matched the terms in the training data (SeedSynPattern). SyntaxCon means that concordances of ACMCSSAff or ACMCSSAffSyn (ACMCSSAff synonyms) used the keyword in context to find the syntactic frame in the training data. If the part-of-speech (POS) of the word was a cardinal number, then this feature CDPOS was set to 1.

(5). Statistical and Syntactic Level:

This research used the CRF model to filter terms extracted by the C/NC value approach with no frequency threshold

2.3.2. Global Feature

(1). Cue word:

KeyWord infers that the word sequence matched one of the user's keywords or one word of the user's title. IsTransW and IsCV represent that a word was found in an NP after TransW or CV respectively. TransW indicates summative and enumerative transitional words, such as "*in summary*", "*to conclude*", "*then*", "*moreover*", and "*therefore*", and CV refers to words under SUMO's "*communication*" concepts, such as "*propose*", "*argue*", "*attempt*" and so on.

(2). Tense:

If the first sentence of the RA is in the past tense and contains an NP, then the word sequence of that NP was used as a useful feature PastNP. This is because the first sentence often impresses upon the reader the shortest possible relevant characterization of the paper, and the use of past tense emphasizes the importance of the statement.

(3). Sentence structure:

Phrases in a parallel structure sentence refers to the phrases appearing in a sentence structure such as Phrase, Phrase, or (and) Phrase, and implies that the same pattern of words represents the same concept. ParallelStruct indicates that the word was part of a phrase in a parallel structure.

2.4. Word Cohesiveness Measure

By calculating the cohesiveness of words, the combined association measure (CAM) can assist in further enhancing and editing the CRF-based propositional term boundaries for achieving a perfect boundary of propositional terms. CAM extracts the most relevant word sequence by combining endogenous linguistic statistical information, including word form sequence and its POS sequence. CAM is a variant of normalized expectation (NE) and

mutual expectation (ME) methods.

To characterize the degree of cohesiveness of a sequence of textual units, NE evaluates the average cost of loss for a component in a potential word sequence. NE is defined in Equation (5) where the function c(·) means the count of any potential word sequence. An example of NE is shown in Equation (6).

$$NE\left(\left[w_1 \ldots w_i \ldots w_n\right]\right) = \frac{C\left(\left[w_1 \ldots w_i \ldots w_n\right]\right)}{\frac{1}{n}\left(C\left(\left[w_1 \ldots w_i \ldots w_n\right]\right) + \sum_{i=2}^{n} C\left(\left[w_1 \ldots \hat{w}_i \ldots w_n\right]\right)\right)} \tag{5}$$

$$\begin{aligned} &NE\left(\left[\text{large vocabulary continuous speech recognition}\right]\right) \\ &= \frac{C\left(\left[\text{large vocabulary continuous speech recognition}\right]\right)}{\frac{1}{5}\left(\begin{array}{l} C\left(\left[\text{large vocabulary continuous speech recognition}\right]\right) \\ +C\left(\left[\text{large continuous speech recognition}\right]\right) \\ +C\left(\left[\text{large vocabulary speech recognition}\right]\right) \\ +C\left(\left[\text{large vocabulary continuous recognition}\right]\right) \\ +C\left(\left[\text{large vocabulary continuous speech}\right]\right) \end{array}\right)} \end{aligned} \tag{6}$$

Based on NE and relative frequency, the ME of any potential word sequence is defined as Equation (7), where function P(·) represents the relative frequency.

$$ME\left(\left[w_1 \ldots w_i \ldots w_n\right]\right) = P\left(\left[w_1 \ldots w_i \ldots w_n\right]\right) \times NE\left(\left[w_1 \ldots w_i \ldots w_n\right]\right) \tag{7}$$

CAM considers that the global degree of cohesiveness of any word sequence is evaluated by integrating the strength in a word sequence and the interdependence of its POS. Thus CAM evaluates the cohesiveness of a word sequence by the combination of its own ME and the ME of its associated POS sequence. In Equation (8), CAM integrates the ME of word form sequence $\left[w_1 \ldots w_i \ldots w_n\right]$ and its POS $\left[p_1 \ldots p_i \ldots p_n\right]$. Let α be a weight between 0 and 1, which determines the degree of the effect of POS or word sequence in the word cohesiveness measure.

$$CAM\left(\left[w_1 \ldots w_i \ldots w_n\right]\right) = ME\left(\left[w_1 \ldots w_i \ldots w_n\right]\right)^{\alpha} \times ME\left(\left[p_1 \ldots p_i \ldots p_n\right]\right)^{1-\alpha} \tag{8}$$

This paper uses a sliding window moving in a frame and compares the CAM value of neighboring word sequences to determine the optimal propositional term boundary. Most lexical relations associate words distributed by the five neighboring words [20]. Therefore this paper only calculates the CAM value of the three words to the right and the three words to the left of the CRF-based terms. Figure 3 represents an illustration for the CAM computation that was fixed in the [(2*3) + length(CRF-Based term)] frame size with a sliding window. When the window starts a forward or backward move in the frame, the three marginal words of a term are the natural components of the window. As the word number of the CRF term is less than three words, the initial sliding windows size is equal to the word number of the term.

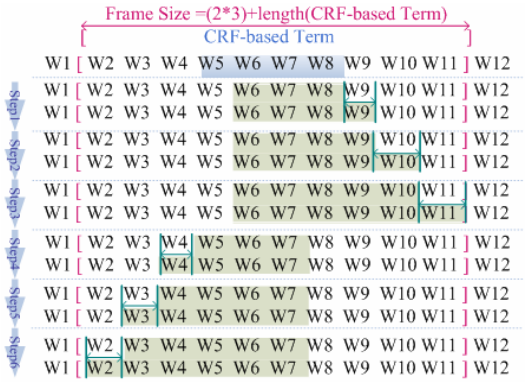


Figure 3. An Illustration for the CAM Computation Steps

To find the optimal propositional term boundary, this study calculates the local maximum CAM value by using the Modified CamLocalMax Algorithm. The principle of the original algorithm [21] is to infer the word sequence as a multiword unit if the CAM value is higher than or equal to the CAM value of all its sub-group of (n-1) words and if the CAM value is higher than the CAM value of all its super-group of (n+1) words. In the Modified CamLocalMax Algorithm, when the CAM value of the combination of CRF-based single word propositional terms and its immediate neighbor word is higher than the average of the CAM value of bi-gram propositional terms in the training data, the components of the CRF-based single word propositional terms are turned into a bi-gram propositional term. The complete Modified CamLocalMax Algorithm is shown in the following, where *cam* means the combined association measure, size(·) returns the number of words of a possible propositional term, *M* represents a possible propositional term, Ω_{n+1} denotes the set of all the possible (n+1)grams containing *M*, Ω_{n-1} denotes the set of all the possible (n-1)grams contained in *M*, and bi-term typifies bi-gram propositional terms in the training data.

Input: *M, a possible propositional term,* $\forall y \in \Omega_{n+1}$, *the set of all the possible (n+1)grams containing M,* $\forall x \in \Omega_{n-1}$, the set of all the possible (n-1)grams contained in M

Output: *CT={ct₁,ct₂,…ctₙ}, a CRF+CAM-based propositional term set*

If (size(M)=2 and cam(M) > cam(y))
or (size(M)>2 and cam(M) ≧ cam(x) and cam(M) >cam(y))
or (size(M)=1 and cam(bi-gram) ≦ cam(M))
End if
Return ct

2.5. Propositional Term Generation Algorithm

The Propositional Term Generation algorithm utilizes the CRF model to generate a CRF-based propositional term set $T=\{t_1,t_2,\ldots t_n\}$ and calculates the CAM value to produce a CRF+CAM-based propositional term set $CT=\{ct_1,ct_2,\ldots ct_n\}$. The detailed processes of the Propositional Term Generation algorithm are as follows

t_n^k *: the word form sequence from the first word 1 to last word k of CRF-based propositional term* t_n

Input: *Word sequence* W_1^n

Output: *T={t₁,t₂,…tₙ}, a CRF-based propositional term set and, CT={ct₁,ct₂,…ctₙ}, a CRF+CAM-based propositional term set*

Input W_1^n *to generate T={t₁,t₂,…tₙ} by CRF*

For all $t_j \in T$
For a=0 to a =2 Step 1

ct_j=*Modified_CamLocalMax*(t_j^{k+a}, t_j^{k+a-1}, t_j^{k+a+1})
$CT \leftarrow CT \cup ct$
End for
If $t_j \notin CT$ *Then*
For a=0 to a =-2 Step -1
ct_j=*Modified_CamLocalMax*(t_j^{1+a}, t_j^{1+a-1}, t_j^{1+a+1})
$CT \leftarrow CT \cup ct_j$
End for
End if
End for
Return T, CT

2.6. Encoding Schema

The IOB encoding scheme was adopted to label the words, where I represents words Inside the propositional term, O marks words Outside the propositional term, and B denotes the Beginning of a propositional term. It should be noted that here the B tag differs slightly from Ramshaw and Marcus's definition, which marks the left-most component of a baseNP for discriminating recursive NPs. Figure 4 shows an example of the IOB encoding scheme that specifies the B, I, and O labels for the sentence fragment "*The syllable-based system exceeds the performance of the triphone system by…*". An advantage of this encoding scheme is that it can avoid the problem of ambiguous propositional term boundaries, since IOB tags can identify the boundaries of immediate neighbor propositional terms, whereas binary-based encoding schemes cannot. In Figure 4, "*syllable-based system*", and "*exceeds*" are individual and immediate neighbor propositional terms distinguished by B tags.

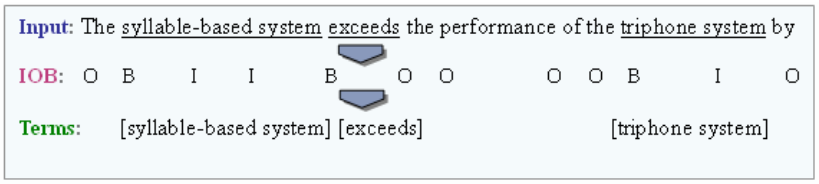


Figure 4. An Example of the IOB Encoding Scheme

3. Evaluation

3.1. Experimental Setup

To facilitate the development and evaluation of the propositional term extraction method, experts manually annotated 260 research abstracts, including speech, language, and multimedia information processing journal papers from SCI and SCIE-indexed databases. In all, there were 109, 72, and 79 annotated research abstracts in the fields of speech, language, and multimedia information processing, respectively. At run time, 90% of the RAs were allocated as the training data and the remaining 10% were reserved as the test data for all evaluation.

In system implementation, the CRF++: Yet Another CRF toolkit 0.44 [22] was adopted. The training parameters were chosen using ten-fold cross-validation on each experiment.

The proposed system was compared with three baseline systems. The first was the C/NC-value algorithm with no frequency threshold, because the C/NC-value algorithm is a hybrid methodology and its historical result is better than the linguistic and statistical approaches. The second baseline system proposed by Nenadić et al. [8] is a variant of the

C/NC-value algorithm enriched by morphological and structural variants. The final baseline system is a linguistic approach proposed by Ananiadou [4]. That study, however, made no comparisons with statistical approaches which are suitable for a document containing a large amount of words.

To evaluate the performance in this study, two hit types for propositional term extraction: perfect and imperfect [23] are employed. A perfect hit means that the boundaries of a term's maximal term form conform to the boundaries assigned by the automatic propositional term extraction. An imperfect hit means that the boundaries assigned by the automatic propositional term extraction do not conform to the boundaries of a term's maximal term form but include at least one word belonging to a term's maximal term form. Taking the word sequence "*large vocabulary continuous speech recognition*" as an example, when the system detects that "*vocabulary continuous speech recognition*" is a propositional term, it then becomes an imperfect hit. There is only one perfect hit condition where "*large vocabulary continuous speech recognition*" is recognized. The metrics of recall and precision were also used to measure the perfect and imperfect hits. The definition of recall and precision of perfect hits and imperfect hits are shown in Equation (9) and Equation (10). Thus, our system is evaluated with respect to the accuracies of propositional term detection and propositional term boundary detection. That is, our motivation for propositional term extraction was to provide CRF and CRF+CAM for accurate detection of propositional terms and the improvement of the detected propositional term boundaries.

$$\text{Recall} = \text{Hits Perfect (or Imperfect)} \,/\, \text{Target Termforms} \tag{9}$$

$$\text{Precision} = \text{Hits Perfect (or Imperfect)} \,/\, \text{Extracted Termforms} \tag{10}$$

3.2. Experimental Results

This study evaluated empirically two aspects of our research for different purposes. First, the performance of propositional term extraction for CRF-based and CRF+CAM-based propositional term sets on different data was measured. Second, the impact of different level features for propositional term extraction using CRF was evaluated.

Evaluation of Different Methods

Table 1. The Performance of Imperfect Hits on Different Data

Method	R	P	F	R	P	F
	All Data			Language Data		
CRF Inside Testing	93.2	94.5	93.9	96.7	98.1	97.4
CRF +CAM Inside Testing	96.6	96.0	96.3	98.4	99.6	99.0
CRF Outside Testing	77.1	74.1	75.6	78.6	76.3	77.4
CRF +CAM Outside Testing	82.6	82.5	82.6	85.8	88.8	87.2
C/NC Value	53.4	65.3	58.8	48.1	53.3	50.6
Ananiadou	51.3	70.0	59.2	52.4	68.4	59.3
Nenadić et al.	58.0	72.3	64.4	60.1	69.0	64.3
	Speech Data			Multimedia Data		
CRF Inside Testing	96.6	99.0	98.2	98.0	99.2	98.6
CRF +CAM Inside Testing	97.5	99.0	99.4	98.6	99.3	99.0
CRF Outside Testing	74.9	76.1	74.3	61.2	65.0	63.1
CRF +CAM Outside Testing	82.6	83.9	84.2	65.4	71.2	68.2
C/NC Value	53.5	79.0	62.7	67.7	53.2	59.6
Ananiadou	53.1	68.4	59.8	65.4	60.0	62.6

Nenadić et al.	59.6	72.2	65.3	68.9	55.2	61.3

Table 1 lists the recall rate, the precision rate and F-score of propositional term extraction for imperfect hits of different domain data. In each case, the recall and precision of imperfect hits using CRF inside testing was greater than 93%. The CRF outside test achieved approximately 73% average recall and 73% average precision for imperfect hits, and the CAM approach improved the original performance of recall and precision for imperfect hits. The C/NC-value approach achieved approximately 56% average recall and 63% average precision for imperfect hits. The performance of Ananiadou's approach was about 56% average recall and 67% average precision for imperfect hits. Another baseline, the approach of Nenadić, Ananiadou and McNaught, obtained approximately 62% average recall and 67% average precision for imperfect hits.

Table 2. The Performance of Perfect Hits on Different Data

Method	R	P	F	R	P	F
	All Data			Language Data		
CRF Inside Testing	66.5	66.2	66.3	66.4	67.5	67.0
CRF +CAM Inside Testing	69.0	68.6	68.8	69.4	69.9	69.6
CRF Outside Testing	39.8	42.2	41.9	43.2	37.3	40.0
CRF +CAM Outside Testing	43.5	49.2	46.2	45.3	45.4	45.3
C/NC Value	27.6	37.8	31.9	28.9	29.1	29.0
Ananiadou	26.3	37.9	31.1	31.3	37.7	34.2
Nenadić et al.	30.2	41.0	34.8	31.2	40.9	35.4
	Speech Data			Multimedia Data		
CRF Inside Testing	62.3	61.0	61.7	70.9	70.3	70.6
CRF +CAM Inside Testing	69.6	67.9	68.7	73.1	70.3	71.6
CRF Outside Testing	36.9	41.6	39.1	42.1	42.5	42.3
CRF +CAM Outside Testing	42.8	48.9	45.6	45.6	45.0	44.3
C/NC Value	29.0	40.0	33.6	34.6	29.9	32.1
Ananiadou	27.4	37.7	31.7	29.3	38.0	33.1
Nenadić et al.	30.0	38.6	33.7	35.3	37.6	35.3

Table 2 summarizes the recall rates, precision rates and F-score of propositional term extraction for perfect hits of data from different domains. The CRF inside test achieved approximately 67% average recall and 66% average precision on perfect hits, but the CRF outside test did not perform as well. However, the CAM approach still achieved an increase of 1%-7% for perfect hits. The C/NC-value approach obtained approximately 30% average recall and 34% average precision for perfect hits. Ananiadou's approach achieved approximately 29% average recall and 38% average precision for perfect hits. The performance of Nenadić, Ananiadou and McNaught's approach was about 32% average recall and 40% average precision for perfect hits.

The results show that the C/NC-value does not demonstrate a significant change over different fields, except for the multimedia field, which had slightly better recall rate. The main reasons for errors produced by C/NC-value were propositional terms that were single words or acronyms, propositional terms that were not NP-based, or propositional terms that consisted of more than four words.

Ananiadou's approach was based on a morphological analyzer and combination rules for the different levels of word forms. Experimental results showed that this approach is still unable to deal with single words or acronyms, and propositional terms that are not NP-based.

Nenadić et al.'s approach considered local morphological and syntactical variants using C value to determine the propositional terms. This approach had slightly better performance than the C/NC value methodology. Acronyms were included in the propositional term

candidates but were filtered by frequency, as they often appear only a few times. This approach also ignored single words, and propositional terms that were not NP-based. Furthermore, none of these three baseline systems are suitable for handling special symbols.

For CRF inside testing, both the precision and recall rates were significantly better for imperfect hits, but the precision and recall rates were reduced by about 30% for perfect hits in most RAs. Due to insufficient training data, CRF no longer achieved outstanding results. In particular, the large variability and abstract description of the multimedia field RAs led to huge differences between measures. For example, in the sentence "*For surfaces with varying material properties, a full segmentation into different material types is also computed*", "*full segmentation into different material types*" is a propositional term that it isn't concretely specified as a method. CRF achieved a better result in recall rate, but failed on propositional term boundary detection, unlike the C/NC-value approach.

The CAM approach effectively enhanced propositional term boundary detection by calculating word cohesiveness, except in the case of multimedia data. The CAM approach couldn't achieve similar performance for the multimedia data as a result of the longer word count of terms that differ from the data of other fields. However, the CAM approach performed best with α equal to 0.4, which demonstrates that the POS provided a little more contribution for multiword term construction. The CAM approach not only considered the POS sequence but also the word sequence, therefore the results are a little better for speech data, which is the biggest part of the training data (SCI and SCIE-indexed databases).

The above results show that the CRF approach exhibited impressive improvements in propositional term detection. The major reason for false positives was that the amount of the data was not enough to construct the optimal model. Experimental results revealed that the CAM is sufficiently efficient for propositional term boundary enhancement but the longer word count of propositional terms were excluded.

Evaluation of Different Level Features

In order to assess the impact of different level features on the extraction method, this paper also carried out an evaluation on the performance when different level features were omitted. Table 3 presents the performance of CRF when omitting different level features for imperfect hits and the symbol "-" denoted the test without a level feature. For all data, the recall rate was reduced by approximately 1%- 5% and the precision rate was reduced by approximately 2%- 6% in inside testing result. In all data outside testing, the recall rate was reduced by 2%-10% and the precision rate was reduced by 1%-5%. The recall and precision for speech data retained similar results from semantic level features, but showed little impact from other local features. For language data, without morphological, syntactic, frequency, and syntactic & statistical level features the performance was slightly worse than the original result and without semantic level features the original performance was preserved. The performance for multimedia data was affected greatly by semantic level features. A slight improvement without morphological, and syntactic & statistical level features and similar results were obtained when frequency and syntactic level features were omitted.

Table 3. The Performance of CRF Excepting Different Level Features for Imperfect Hits

Data Type / Testing Type	All		Speech		Language		Multimedia	
	R	P	R	P	R	P	R	P
Inside -Frequency Features	92	92	94	97	95	97	98	98
Inside -Morphological Features	88	90	92	96	93	96	97	97
Inside -Syntactic Features	90	89	94	96	95	97	97	98
Inside -Semantic Features	92	92	96	98	97	98	95	97
Inside -Syntactic & Statistical Features	90	93	93	95	95	96	96	98
Inside Testing	93	95	97	99	97	98	98	99
Outside -Frequency Features	74	73	71	73	76	74	60	65

Outside -Morphological Features	71	71	59	69	70	68	58	65
Outside -Syntactic Features	67	69	60	71	71	71	59	64
Outside -Semantic Features	75	75	75	76	78	76	41	60
Outside -Syntactic &Statistical Features	71	73	67	71	70	70	55	65
Outside Testing	77	74	75	76	79	76	61	65

In Table 4, it can be noticed that the omission of any single level features results in a deterioration in the performance of perfect hits. Removing the syntactic level features had the most pronounced effect on performance for all, speech and language data, while removing the semantic level features had the least effect on performance for all, speech and language data. According to the experimental results, the use of the frequency features did not result in any significant performance improvement for the multimedia data, and the use of the syntactic and syntactic & statistical level features did not result in any performance improvement for the multimedia data. Removing the semantic level features had the greatest effect on the performance for the multimedia data.

Table 4. The Performance of CRF without Different Level Features for Perfect Hits

Data Type / Testing Type	All		Speech		Language		Multimedia	
	R	P	R	P	R	P	R	P
Inside -Frequency Features	63	60	56	55	61	64	60	60
Inside -Morphological Features	61	61	57	54	61	64	70	68
Inside -Syntactic Features	60	60	55	57	63	65	68	67
Inside -Semantic Features	65	62	59	60	66	69	62	62
Inside -Syntactic &Statistical Features	62	61	57	52	62	64	71	68
Inside Testing	67	66	62	61	66	68	71	70
Outside -Frequency Features	36	38	34	35	37	34	40	40
Outside -Morphological Features	33	35	32	36	35	34	40	39
Outside -Syntactic Features	35	36	32	38	37	32	39	40
Outside -Semantic Features	38	40	36	40	41	36	29	31
Outside -Syntactic &Statistical Features	38	39	32	37	35	33	40	40
Outside Testing	40	42	37	42	42	37	42	42

Overall the five different level features were all somewhat effective for propositional term extraction. This suggests that propositional terms are determined by different level feature information which can be effectively used for propositional term extraction. The frequency level features contributed little for propositional term extraction in all and speech data. This may be due to the fact that speech data comprised the main portion of the training data. In the multimedia case, the semantic level features were useful. Although semantic level features may include some useful information, it was still a problem to correctly utilize such information in the different domain data for propositional term extraction. Syntactic and morphological level features obtained the best performance for all, speech and language data. This may be due to the amount of training data in each domain and the various word forms of propositional terms in the multimedia data. The syntactic and statistical level features improved or retained the same performance, which indicates the combined effectiveness of syntactic and statistical information.

3.3. Error Analysis

Table 5 shows the distribution of error types on propositional term extraction for each domain data using outside testing. This study adopts the measure used in [24] to evaluate the error type, where M indicates the condition when the boundary of the system and that of the standard match, O denotes the condition when the boundary of the system is outside that of the standard and I denotes the condition when the boundary of the system is inside that of the standard. Therefore, the MI, IM, II, MO, OM, IO, OI and OO error types were used to

evaluate error distribution. The relative error rate (RER) and the absolute error rate (AER) were computed in error analysis, the relative error rate was compared with all error types, and the absolute error rate was compared with the standard. In the overall error distribution, the main error type was "*IM*" and "*MI*" and the CRF+CAM can significantly reduce those two error types.

Table 5. Distribution of Error Types on Propositional Term Extraction

Error Type	CRF		CRF+CAM		CRF		CRF+CAM	
	RER	AER	RER	AER	RER	AER	RER	AER
	All Data				Speech Data			
MI	24.62	6.11	18.00	2.90	24.90	6.41	20.30	3.03
IM	36.48	8.72	28.50	4.88	38.22	8.06	32.50	4.08
II	18.67	4.96	23.40	3.88	12.37	2.88	14.80	2.05
MO, OM, IO, OI	7.49	3.08	12.50	1.07	10.50	2.46	12.85	1.85
OO	12.74	2.91	17.60	2.08	14.01	4.55	19.55	2.53
	Language Data				Multimedia Data			
MI	23.11	4.03	18.50	2.67	19.18	6.58	17.25	4.64
IM	31.25	9.08	28.50	3.56	25.72	9.00	19.10	4.05
II	26.48	7.50	31.00	4.07	36.34	10.63	34.34	8.30
MO,OM,IO,OI	8.12	1.03	12.45	1.89	6.42	5.00	10.09	1.53
OO	11.04	2.06	9.55	1.20	12.34	4.85	19.22	3.85

4. Conclusion

This study has presented a conditional random field model and a combined association measure approach to propositional term extraction from research abstracts. Unlike previous approaches using POS patterns and statistics to extract NP-based multiword terms, this research considers lexical cohesion and context information, integrating CRFs and CAM to extract single or multiword propositional terms. Experiments demonstrated that in each corpus, both CRF inside and outside tests showed an improved performance for imperfect hits. The proposed approach further effectively enhanced the propositional term boundaries by the combined association measure approach which calculates the cohesiveness of words. The conditional random field model initially detects propositional terms based on their local and global features, which includes statistical, syntactic, semantic, morphological, and structural level information. Experimental results also showed that different multi-level features played a key role in CRF propositional term detection model for different domain data.

References

[1] U. M. Connor, *Contrastive Rhetoric: Cross-Cultural Aspects of Second Language Writing* U.K.: Cambridge Applied Linguistics, 1996.

[2] C. Jacquemin and D. Bourigault, "Term Extraction and Automatic Indexing," in *Oxford Handbook of Computational Linguistics*, M. Ruslan, Ed. Oxford: Oxford University Press, 2003, pp. 599-615.

[3] C.-K. Cheng, *How to Write a Scientific Paper?* Taipei: Hwa Kong Press, 2003.

[4] S. Ananiadou, "A Methodology for Automatic Term Recognition," in *15th Conference on Computational Linguistics - Volume 2*, Kyoto, Japan, 1994, pp. 1034-1038.

[5] F. J. Damerau, "Generating and Evaluating Domain-Oriented Multi-word Terms From Texts," *Inf. Process. Manage.*, vol. 29, pp. 433-447, 1993.

[6] C. Enguehard and L. Pantera, "Automatic Natural Acquisition of a Terminology," *Journal of*

Quantitative Linguistics, vol. 2, pp. 27-32, 1995.

[7] K. T. Frantzi, S. Ananiadou, and H. Mima, "Automatic Recognition of Multi-word Terms: the C-value/NC-Value Method," *Int. J. on Digital Libraries*, vol. 3, pp. 115-130, 2000.

[8] G. Nenadić, S. Ananiadou, and J. McNaught, "Enhancing Automatic Term Recognition through Recognition of Variation," in *20th international conference on Computational Linguistics* Geneva, Switzerland: Association for Computational Linguistics, 2004.

[9] L. A. Ramshaw and M. P. Marcus, "Text Chunking Using Transformation-Based Learning," in *Third Workshop on Very Large Corpora*, 1995, pp. 82-94.

[10] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282-289.

[11] G. Dias, "Multiword Unit Hybrid Extraction," in *ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, 2003, pp. 41-48.

[12] Association for Computing Machinery, Inc., *The ACM Computing Classification System [1998 Version],* New York: ACM. Available: http://www.acm.org/class/1998/. [Accessed: June 17, 2006]

[13] I. Niles and A. Pease, *Suggested Upper Merged Ontology (SUMO) Mapping to WordNet*, Piscataway NJ: IEEE. Available: http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/. [Accessed: 2004]

[14] The School of Linguistics and Applied Language Studies at Victoria University of Wellington, *Academic Words List*, Wellington: Victoria University of Wellington. Available: http://www.vuw.ac.nz/lals/research/awl/. [Accessed: June 17, 2006]

[15] E. Charniak, *Eugene Charniak's Parser*, Providence: Brown University. Available: http://cs.brown.edu/~ec/. [Accessed: June 1, 2006]

[16] J. Nocedal, "Updating quasi-Newton Matrices with Limited Storage," *Mathematics of Computation*, vol. 35, pp. 773-782, 1980.

[17] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields," in *2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03)*, Edmonton, Canada, 2003, pp. 213-220.

[18] S. C. Lee, "Probabilistic Segmentation for Segment-Based Speech Recognition." M. S. thesis, Massachusetts Institute of Technology, MA, U.S.A., 1998.

[19] G. Nenadić, I. Spasić, and S. Ananiadou, "Automatic Acronym Acquisition and Term Variation Management within Domain-specific Texts," in *Third International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Canary Islands, Spain, 2002, pp. 2155-2162.

[20] S. Jones and J. Sinclair, "English Lexical Collocations: A Study in Computational Linguistics," *Cahiers de Lexicologie*, vol. 23, pp. 15-61, 1974.

[21] G. Dias, "Extraction Automatique d'Associations Lexicales àpartir de Corpora." Ph. D dissertation, DI/FCT New University of Lisbon, Lisbon, Portugal, and LIFO University, Orléans, France , 2002.

[22] K. Taku, *CRF++: Yet Another CRF toolkit 0.44*. Available: http://crfpp.sourceforge.net/. [Accessed: Oct 1, 2006]

[23] A. Lauriston, "Criteria for Measuring Term Recognition," in *Seventh Conference on European Chapter of the Association for Computational Linguistics*, Dublin, Ireland, 1995, pp. 17-22.

[24] K.-M. Park, S.-H. Kim, H.-C. Rim, and Y.-S. Hwang, "ME-based Biomedical Named Entity Recognition Using Lexical Knowledge," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, pp. 4-21, 2006.

利用統計方法及中文訓練資料處理台語文詞性標記

Modeling Taiwanese POS tagging with statistical methods and Mandarin training data

楊允言 [1] 戴嘉宏 [2] 劉杰岳 [3] 陳克健 [2] 高成炎 [1]
[1] 國立台灣大學資訊工程系
{d93001,cykao}@csie.ntu.edu.tw
[2] 中央研究院資訊科學研究所
{glaxy,kchen}@iis.sinica.edu.tw
[3] 獨立研究學者
kiatgak@gmail.com

摘要

本文提出利用有六萬多詞條的台華辭典以及千萬詞的中文訓練資料來做台語文詞性標記的方法。台語文語料爲包括全羅馬字及漢羅合用兩種書寫文本的文學資料，文類涵蓋散文、小說、劇本等，詞類集採用中央研究院詞庫小組所訂定的中文詞類集。

我們開發語詞對齊檢查程式，將兩種文本的語料逐詞對齊，透過台華辭典查詢每個語詞相對應的中文候選詞，接著利用中文訓練資料，以 HMM 機率模型挑選出最適當的中文對譯詞，再以 MEMM 分類器標記詞性。

實驗結果顯示，以此方法做台語文詞性標記，我們得到 91.49%的正確率，並針對標記錯誤分析其原因。以此基礎，我們也得到了初步的台語文訓練語料。

Abstract

In this paper, we propose a POS tagging method using more than 60 thousand entries of Taiwanese-Mandarin translation dictionary and 10 million words of Mandarin training data to tag Taiwanese. The literary written Taiwanese corpora have both Romanization script and Han-Romanization mixed script, the genre includes prose, fiction and drama. We follow tagset drawn up by CKIP.

We develop word alignment checker to help the two scripts word alignment work, and then lookup Taiwanese-Mandarin translation dictionary to find the corresponding Mandarin

candidate words, select the most suitable Mandarin word using HMM probabilistic model from the Mandarin training data, and finally tag the word using MEMM classifier.

We achieve an accuracy rate of 91.49% on Taiwanese POS tagging work, and analysis the errors. We also get the preliminary Taiwanese training data.

關鍵詞：詞性標記，台語文，中文

Keywords: POS tagging, written Taiwanese, Mandarin

一、前言

雖然一直沒有受到足夠的重視，閩南語在全世界語言人口數有四千六百多萬，是排名第 21 位的語言，主要分佈在台灣、新加坡、馬來西亞、汶萊、中國、泰國、菲律賓及印尼等地[1]。台灣閩南語的語言使用人口，在台灣約佔 70%以上，是最主要的台灣本土語言[2]。而這個語言的名稱，至今也還是很分歧，至少有 17 種稱呼，包括台語、福建話、閩南話、福佬話、…等等[3]。本文將使用一般大眾對其的稱呼：「台語」，不打算涉入名稱的討論。

台語有漢字及羅馬字書寫兩種文字傳統，漢字書寫可追溯自 16 世紀，包括南管戲文或是天主教書籍等[4]；羅馬字書寫則可追溯自 1832 年起，包括辭典、宗教作品、報紙、啓蒙讀物、教科書、…等[5]。兩種文字各有其優缺點，漢字書寫的，較難確認其實際發音，又有訓讀字、本字、借音字、本土字等不同類別的字，專家考證的本字又各有不同[6]。但是在台灣，因爲中文教育的普及，大家看到漢字書寫的台語文普遍比較不會排斥。羅馬字標注出實際的發音，語詞間以空格隔開，語詞內以連字符（hyphen）隔開每個音節，以資訊處理的角度來看，可能好用多了。

爲了建立台語計算語言學的基礎，過去幾年，我們陸續建立了台語華語對譯辭典[7]、台語文未加工語料庫等資源，以未加工語料庫爲基礎建立台語文語詞檢索系統[8]、以規則方法處理台語的變調處理問題[9]等。我們希望進一步將語料庫做更完整的標注。對於語料庫的標注，最基本且重要的，應當是詞性標記。

目前我們要做台語文語料的詞性標注，首先馬上面臨一個難題：台語的詞類集爲何？至今並沒有一套標準。在此情形下，我們暫時採用中央研究院詞庫小組所訂定的中文詞類

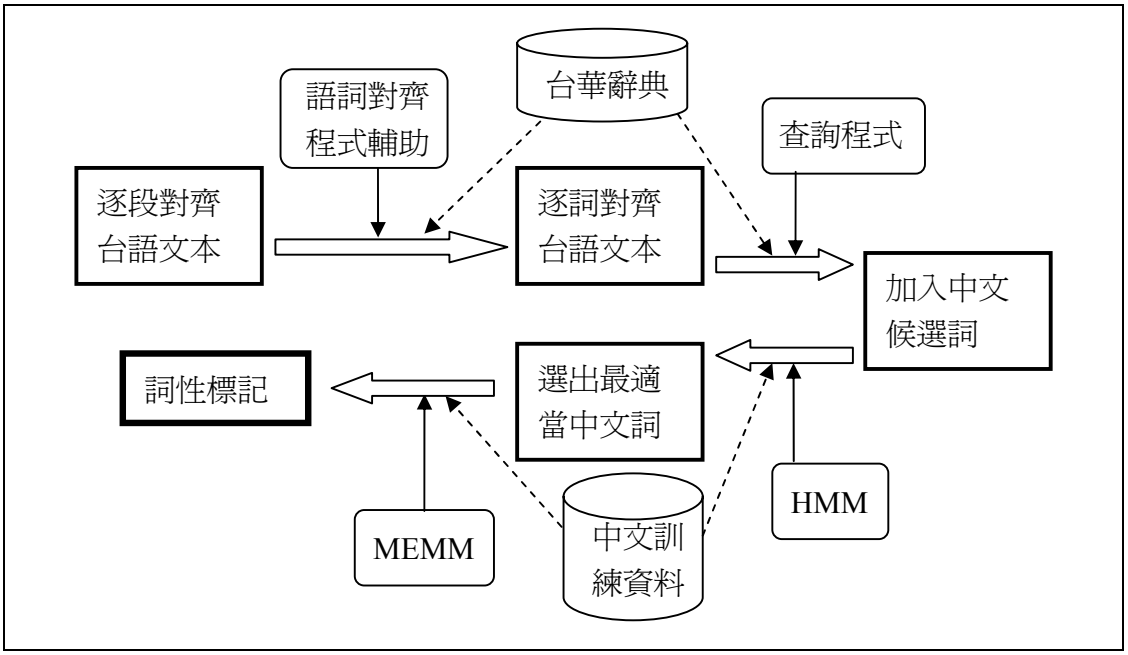
集[10]。這樣仍然會有問題，因爲我們並沒有標記中文詞類集的台語辭典，現有的台語辭典，只有基本詞類如名詞、動詞、形容詞等的訊息。

另外一個問題是人力的缺乏，我們並沒有充足的人力來進行台語文語料的詞性標記。

在此情形下，本文提出利用現有的中文資源，以及台語華語對譯辭典，加上統計方法來進行台語詞性自動標記工作。

二、實驗方法

圖一顯示系統架構圖。



圖一、台語詞性標記系統架構圖

（一）語料來源

我們所使用的語料爲國家台灣文學館「台語文數位典藏資料庫（第二階段）」的計畫成果，有 258 萬多音節全羅馬字（全羅）及漢字羅馬字合用（漢羅）段落對齊的兩種文本，包括小說、散文、劇本、新詩等文類[11]。

（二）逐語詞對齊

首先，我們開發語詞對齊程式輔助人工，逐步將段落對齊的兩種文本逐語詞對齊。這支程式除了核對兩種文本的音節數之外，還將羅馬字、漢羅合用的語詞與台華辭典中的內容做比對，如果這兩個語詞沒有出現在同一詞條內，程式將會標明出來以提醒使用者，有可能是未知詞，也可能是漢字用字不一致，或是此語詞打字錯誤等原因。

台華辭典的原始資料由鄭良偉提供，楊允言在 2000 年提供線上查詢系統，詞條內容經過多次增補，目前有六萬多詞條，包括台語羅馬字、台語漢羅、華語對譯詞、英文等欄位，台語部分並附上發音功能，平均每天有 2400 多次的查詢（過去一年平均）。其中英文欄位於 2007 年新增，資料尚不完整[12]。

（三）尋找對應中文候選詞

接著，我們繼續利用台華辭典，將全羅/漢羅配對的語詞，找出其對應的中文候選詞。這是一對多的對應，亦即一個台語全羅/漢羅配對的語詞，可能有一個以上的華語對譯詞。除此外，有些語詞因爲台華辭典沒有收錄而查不到，有些語詞則因漢羅的寫法不同而查不到（例如文本中出現「較贏[khah-iâⁿ]」，而辭典裡爲「khah 贏[khah-iâⁿ]」）。對於此問題，我們的解決方法是：如果全羅/漢羅配對語詞查不到，暫時把漢羅拿掉，用全羅找出對應的中文候選詞，若漢羅的寫法是全漢字，也將漢羅視爲中文候選詞之一（假設其爲台華共通詞）；這麼做有可能讓中文候選詞的詞數增加，尤其是單音節詞（例如文本中出現「轉[chōan]」，辭典找不到此詞條，但是羅馬字爲「chōan」的有兩個詞條，其中文對譯詞分別爲「扭」和「上」，再加上「轉」，詞義皆不同）。

如果還是查不到，則把漢羅語詞直接當做中文候選詞（例如文本中出現「有形[iú-hêng]」，辭典中沒有這個詞條，用羅馬字「iú-hêng」查，也查不到，就直接把「有形」視爲中文候選詞）[13]。

（四）挑選最適當的對應中文詞

我們採用 Markov bigram model，利用中央研究院詞庫小組千萬詞平衡語料庫的 bigram

語詞訓練資料，從中文候選詞中挑選最適當的中文詞。

假設某一句子有 m 個語詞，第一個語詞 w_1 是從 $w_{11}, w_{12}, \ldots, w_{1n_1}$ 候選詞中挑出來的中文詞，第二個語詞 w_2 是從 $w_{21}, w_{22}, \ldots, w_{2n_2}$ 候選詞中挑出來的中文詞，第 m 個語詞 w_m 從 $w_{m1}, w_{m2}, \ldots, w_{mn_m}$ 候選詞中挑出來的中文詞。我們要從中挑選出去可能的串列 $\hat{S} = w_1 w_2 \cdots w_m$，使得 $\prod_{i=1}^{m} P(w_i \mid w_{i-1})$ 爲最大，亦即使 $\sum_{i=1}^{m} \log P(w_i \mid w_{i-1})$ 爲最大。要說明的是，這個串列 $\hat{S}$，可能不是合法的中文句子[14]。

$$w_i = \begin{cases} \arg\max \quad \log_e P(w_{ij}) & i = 1 \quad or \quad \forall j \neg \exists w_{i-1} w_{ij} \\ \arg\max \quad \log_e P(w_{ij} \mid w_{i-1}) & otherwise \end{cases}$$

在訓練資料中，若雙連詞不存在，則取最高頻的單一語詞（unigram）。

（五）根據中文詞挑選最適當的詞性

我們採用 Maximal Entropy Markov Model (MEMM)來挑選詞性。

MEMM 包括一組包含語詞和詞性的歷程集合 H，和詞性集合 T，$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$，其中，$h \in H, t \in T$，$\pi$ 是常數，$\{\mu, \alpha_1, \ldots, \alpha_k\}$ 是大於 0 的參數，$\{f_1, \ldots, f_k\}$ 是特徵 features，$f_j(h,t) \in \{0,1\}$，參數 α_j 對應特徵 f_j。

對於目標語詞 w_i 的詞性 t_i，我們選取 10 個特徵，包括：

1. 語詞：有 $w_i, w_{i-1}, w_{i-2}w_{i-1}, w_{i+1}, w_{i+1}w_{i+2}$ 五個特徵；
2. 詞性：有 $t_{i-1}, t_{i-2}t_{i-1}$ 兩個特徵；
3. 構詞：m_1, m_2, m_n 三個特徵

m_1, m_2, m_n 是針對未知詞，如果 w_i 是未知詞，我們就對 w_i 採對大匹配來斷詞，$w_i = m_1 m_2 \cdots m_n$，在某些情況下，$m_2 = m_3 = \cdots = m_n$。如果 w_i 不是未知詞，則構詞的三個特徵值爲設爲 *null*。此外，若 w_i 爲句首或句尾，某些特徵值也是 *null*，例如 *i=1* 時，$w_{i-1}, w_{i-2}w_{i-1}, t_{i-1}, t_{i-2}t_{i-1}$ 等特徵值皆爲 *null*。

訓練資料爲詞庫小組一千萬詞平衡語料，並以 Viterbi 演算法實作[14-18]。

三、實驗結果

我們利用上述方法執行台語文詞性標記的工作，不過因爲沒有標準答案可以檢查正確率，我們只好抽取部分資料，以人工檢查結果，人工檢查時，主要參考中央研究院詞庫小組的中文斷詞系統。我們選取七篇文章，時間涵蓋清國、日治及終戰後三個不同的時代，文類包括散文（三篇）、劇本（一篇）及小說（三篇），每篇文章挑選第一段，若第一段文字太少，則挑選第二段。表一爲測試資料，爲挑選出來做人工檢視的，並列出每個段落的音節數、語詞數、挑錯語詞數、詞性標記錯誤數，以及詞性標記正確率。

表一、測試資料及其詞性標記正確率

id	年	文類	作者	題目	音節數	語詞數	語詞錯誤	標記錯誤	正確率
1	1885	散文	葉牧師	P̍eh-ōe-jī ê lī-ek(白話字的利益)	162	109	9	6	94.50%
2	1919	散文	H S K	Phín-hēng ê ûi-thôan(品行的遺傳)	180	119	6	8	93.28%
3	1990	散文	陳義仁	Lāu-lâng ê kè-t̍at(老人的價值)	75	49	7	7	85.71%
4	1950	劇本	陳清忠譯	Venice ê Seng-lí-lâng(威尼斯的生意人)	92	58	3	4	93.10%
5	1890	小說	佚名	An-l̍ok-ke(安樂街)	101	77	7	9	88.31%
6	1924	小說	賴仁聲	Án-niá ê B̍ak-sái(母親的眼淚)	133	93	7	9	90.32%
7	1990	小說	楊允言譯	Hái-phīⁿ Sin-niû(岬角上的新娘)	94	59	7	5	91.53%
					837	564	46	48	91.49%
說明：id 4 原著者爲莎士比亞，id 7 原著者爲宋澤萊									

$$\text{正確率} = \left(1 - \frac{\text{標記錯誤數}}{\text{語詞數}}\right) \times 100\%$$

所選取的資料，總共有 564 個語詞（837 個音節），經過人工檢查，有 46 個語詞挑選錯誤、48 個語詞的詞性標記錯誤，詞性標記的平均正確率爲 91.49%。要說明的是，有時所挑選出來的中文詞雖然是錯誤的，但是詞性標記結果可能仍然是正確的，另一方面，中文詞選對，未必詞性標記的結果是正確的。

此外，有時一個台語詞，對應到兩個中文詞，如台語的「壁頂」對應中文的「牆壁 上」；

也有兩個台語詞對應一個中文詞的，如台語的「中國 字」對應到中文的「中國字」。前者會當成未知詞來處理，而後者，因爲拆開的兩部分也都是詞，所以就當成兩個詞處理。這個部分，如果詞性標記是正確的，我們仍將結果視爲正確。如果這部分視爲錯誤，則平均的正確率會下降 2%左右。

以下爲實際的詞性標記結果舉例，此爲 id 7 的一部份。其中，第一個欄位是漢羅台語文，第二欄爲台語羅馬字（以中括弧刮起來），第三欄爲中文候選詞（以大括弧刮起來，前面加'@'符號者，表示第一欄和第二欄的台語文並未出現在辭典的同一詞條中，直接以漢羅台語文當成中文候選詞），第四欄爲挑選出的中文詞（以角括弧刮起來），最後一欄爲所挑選的詞性。我們將挑選的中文詞錯誤或是挑選的詞性標記錯誤者加上底線，之前並加上'**'，錯誤的詞性之後，加上正確的詞性標記（粗體，以小括弧刮起來）。

我[góa]{我}<我>(Nh)
將[chiong]{將}<將>(D)
草帽仔[chháu-bō-á]{@草帽仔}<草帽仔>(Na)
掛[kòa]{帶;掛;戴}<帶>(VC)
tī [tī]{在}<在>(P)
壁頂[piah-teng2]{牆壁上}<牆壁上>(Nc)
，[,]<,>(COMMACATEGORY)
行李[hêng-lí]{行李}<行李>(Na)
khêng[khêng]{收拾;盤點}<收拾>(VC)
khêng[khêng]{收拾;盤點}<收拾>(VC)
leh[leh]{咧}<咧>(T)
，[,]<,>(COMMACATEGORY)
坐[chē]{坐}<坐>(VA)
tòa[tòa]{住}<住>(VCL) (**P**)
小店[sió-tiàm]{@小店}<小店>(Na)
ê[ê]{的}<的>(DE)
tha-thá-mì[tha-tha-mì]{塌塌米}<塌塌米>(Na)
頂 kôan[téng-kôan]{上面}<上面>(Ncd)
，[,]<,>(COMMACATEGORY)
看[khòaⁿ]{看}<看>(VC)
窗外[thang-gōa]{@窗外}<窗外>(Nc)
ê[ê]{的}<的>(DE)
光景[kong-kéng]{風光;氣象;光景;風景;景氣}<**景氣>(Na)
，[,]<,>(COMMACATEGORY)

看[khòaⁿ]{看}<看>(VC)
起起[khí-khí]{@起起}<起起>(**Nb)(**VA**)
落落[lo̍h-lo̍h]{@落落}<落落>(VA)
ê[ê]{的}<的>(DE)
海湧[hái-éng]{海浪;海潮}<海浪>(Na)
，[,]<,>(COMMACATEGORY)
因爲[in-ūi]{由於;因爲}<因爲>(Cbb)
等待[tán-thāi]{留待;等待}<等待>(VK)
朋友[pêng-iú]{友人;朋友}<朋友>(Na)
，[,]<,>(COMMACATEGORY)
心適[sim-sek]{好玩;好玩兒;有趣;風趣;愉快;稀奇;鬧著玩}<有趣>(VH)
心適[sim-sek]{好玩;好玩兒;有趣;風趣;愉快;稀奇;鬧著玩}<有趣>(VH)
，[,]<,>(COMMACATEGORY)
輕輕仔[khin-khin-á]{輕輕的}<輕輕的>(**Nb)(**D**)
來[lâi]{來}<來>(D)
點[tiám]{燃點;檢點;點;點子}<點>(VC)
一支[chi̍t-ki]{@一支}<一支>(Na)
涼涼[liâng-liâng]{冷冷;涼絲絲}<**冷冷>(VH)
ê[ê]{的}<的>(DE)
芎蕉[kin-chio]{香蕉}<香蕉>(Na)
薰[hun]{香菸;香煙;薰}<香煙>(Na)
。[.]<.>(PERIODCATEGORY)

四、分析

我們針對選錯中文詞或詞性標記錯誤之處，做更詳細的檢視，發現其中有 13 處是因爲選錯中文詞導致詞性標記錯誤。表二列出選錯的中文詞及其標記的詞性。

表二、系統選錯的中文詞

台語詞	所選的中文詞及詞性	較適當的中文對譯及詞性	說明
押/ah	強制(D)	押(VC)	
無/bô	不(D)	沒有(VJ)	2 次
這號/chit-hō	這樣(VH)	這種(N?)	2 次
轉/chōan	上(Ncd)	轉(Vac)	2 次
夭壽/iáu-siū	非常(Dfa)	早夭(VH)	
價值/kè-ta̍t	值得(VH)	價值(Na)	
活/o̍ah	生活(Na)	活(VH)	
破相/phòa-siùⁿ	破(VHC)	殘廢(Na)	
相借問/sio-chioh-mn̄g	招呼(VC)	打招呼(VB)	
著/tio̍h	就(P)	得(D)	

有兩處選錯中文詞是因爲台華辭典中沒有正確中文詞的選項，也導致詞性標記錯誤。這表示台華辭典還需要繼續增補。表三列出這兩個語詞。

表三、台華辭典缺中文對譯導致選錯的詞

台語詞	系統所選的中文詞	較正確的中文詞
tiāⁿ-tiāⁿ / tiāⁿ-tiāⁿ	常常(D)	而已(T)
轉 / tńg	調解(VC)	轉(VAC)

另外，還有八處錯誤是由於未知詞的詞性標記錯誤。這些未知詞大部分是兩個中文詞。表四列出此八個未知詞。

表四、中文未知詞

台語詞	中文	系統所選的詞性	正確的詞性
bē 會/bē-ē	不會	Nb	D
食老/chia̍h-lāu	*食老	Na	V?
轉了/chōan-liáu	*轉了	VH	V?
法律上/hoat-lu̍t-siōng	法律上	VC	N?
非爲/hui-ûi	非爲	A	N?
窮志/kiông-chì	窮志	Na	V?
輕輕仔/khin-khin-á	輕輕的	Nb	D?
生子/seⁿ-kiáⁿ	生子	Na	V?

還有四個詞性標記錯誤的地方，可能是因爲之前一個詞性標記錯誤而受到影響的，屬於傳播錯誤(propagation error)，包括一個未知詞。

人名的部分，「天賜 ah/Thian-sù ah」的「天賜」（不是未知詞）被標記爲「A」，後綴的「ah」被標記爲「T」或「Di」（共出現兩次，一次選「啊」另外一次選「了」）。

另外有一個台語詞「對/tùi」，在大部分語境下，其中文對譯爲「從」。這個語詞在測試資料中共出現九次，不過系統有七次挑選出的中文詞是「對」，只有兩次挑選「從」。但是因爲詞性標記都是「P」，對詞性標記正確率沒有影響。

其它的錯誤共有 18 處，暫時沒有辦法明確分析出其詞性標記錯誤的原因。

總結我們所分析的詞性標記錯誤的原因及其比例，列於表五。

表五、詞性標記錯誤分析

錯誤原因	次數	比例	說明
選錯中文詞	13	27.08%	
沒有正確的中文詞可選	2	4.17%	
未知詞	8	16.67%	
人名	4	8.33%	
傳播錯誤	4	8.33%	包括一未知詞
總計	30	62.50%	扣除重複算的

最理想的情形，如果上述錯誤都得到解決，以此方法做台語詞性標記，將可以達到96.81%的正確率，但是顯然有極大的困難。

台語的詞序與中文的詞序畢竟有差異，選錯中文詞導致詞性標記錯誤是比例最高的。而沒有正確中文詞可選的問題，可以透過增補台華對譯辭典的詞條獲得解決，但是正確率僅提升不到 5%。

未知詞導致的詞性標記錯誤，佔第二高的比例，以中文的角度看，這些不是真正的未知詞，大都是因爲兩個不同語言的語詞對譯未必是一對一的緣故；另外一個重要的原因是，台語羅馬字在連字符(hyphen)的使用上也尚未標準化，漢語系各語言，可能是使用漢字的關係，語詞的界線相對不明確，台語使用羅馬字書寫，利用連字符，一方面斷開一語詞的各音節，讓一個音節仍可連結一個漢字，另一方面，則又負擔的分詞的功能，有連字符連接的音節，代表同一語詞，語詞間有空白分隔開；但是分詞的部分，原來的漢字書寫無可與之對應。

一個音節可以再細分爲聲母、韻母、聲調三部分，由這三部分所組成的音節，台語約有3,000 個音節，華語有 1,200 個左右，在此前提下，台語有較多的單音節詞。但是，一個單音節可能對應到好幾個不同的漢字，雙音節或以上的語詞則解決大部分的問題。例如台語的「這個」，若寫成「chit ê」（沒有連字符），看到「chit」，可對應的漢字包括「這、職、質、織、...」等，看到「ê」，可對應的漢字包括「的、個、鞋、...」等，如果寫成「chit-ê」（有連字符）時，通常閱讀者可直接對應到「這個」，因此在台語的羅馬字書寫，書寫者可能會傾向把單音節詞和另一個單音節詞用連字符連起來，如果這兩個單音節詞能形成複合詞或詞組。現實的情形是，在語料中，「這個」有加連字符，有的沒有加，存在著不一致的現象。

因爲連字符導致一個台語詞對應兩個中文詞的問題，如果不修改文本，當中文對譯詞是未知詞時，也許可考慮再回頭拆掉連字符試試看。這樣做，也許可以降低因爲未知詞導致詞性標記錯誤的機會。

至於不同時代及不同文類的文本，是否在詞性標記的正確率有所差異？根據表一的資料，表六列出三種不同文類文本的詞性標記正確率，表七列出三個不同時代文本的詞性標記正確率。由表六看來，小說類文本的詞性標記正確率較低，表七則顯示，不同時代文本的詞性標記正確率，並沒有顯著的差異。不過，因爲資料量少，此分析結果還需進一步驗證。

表六、不同文類文本詞性標記正確率比較

文類	語詞數	標記錯誤	正確率
散文	277	21	92.42%
劇本	58	4	93.10%
小說	229	23	89.96%

表七、不同時代文本詞性標記正確率比較

年代	語詞數	標記錯誤	正確率
清國	186	15	91.94%
日治	212	17	91.98%
戰後	166	16	90.36%

五、未來方向

在缺乏台語文訓練資料的情形下，我們繞了一圈，利用台語華語對譯以及中文的訓練資料，讓台語文詞性標記達到 91.49%的正確率。這兩百多萬音節的台語文語料的詞性標記結果雖然沒有完全正確，但是應該足以提供做爲台語文語詞及詞性的訓練資料，可供進一步研究使用。

如果這份台語文的訓練資料是可用的，未來我們希望能透過比較中文和台語文的雙連語詞或雙連詞性，進一步分析中文和台語文的異同處。

我們並開發台語文斷詞、詞性標示系統系統[19]供大眾使用，不過一般人要同時準備台語文全羅馬字及漢羅合用兩種文本有點困難，因此我們還提供只用台語羅馬字或只用漢羅合用（包括完全使用漢字）的台語詞性標記，做法上與上述的相同，只是在查閱台華

辭典時少核對一個欄位，這會造成中文候選詞增加，有可能造成詞性標記錯誤的機會。其結果如何，有待進一步分析。

另外，漢羅合用台語文是比較容易取得的文本，需先經過斷詞才能進行後續的詞性標記，因此我們也將斷詞系統整合在此線上系統中。

致謝

本研究得到國科會計畫「台語文語法結構樹建置(1/3)」NSC 95-2221-E-122 -006 經費補助，特此致謝。此外，也感謝三位匿名審查者所提供的建設性意見，讓本文得以更爲周延。

參考文獻

[1] Gordon, Raymond G. Jr., Ed., *Ethnologue : Languages of the world*. 15th ed. SIL International, 2005. [Online] Available: http://www.ethnologue.com/ [Accessed: Jun. 30, 2008].

[2] 黃宣範, *語言，社會與族群意識*, 台北 : 文鶴, 1993.

[3] 李勤岸 洪惟仁, "沒有名字的語言？「台灣話」、「閩南話」還是 Hó͘h-ló 話？", *台灣文學館通訊* 15 期, 台南 : 國家台灣文學館, pp36-41, May. 2007.

[4] 吳守禮, *閩台方言研究集 1*. 台北 : 南天 , 1995.

[5] 張裕宏, *白話字基本論：白話文對應&相關的議題淺說*, 台北 : 文鶴, 2001.

[6] H.K. Tiunn, "Writing in Two Scripts：A Case Study of Digraphia in Taiwanese," *Written Language and Literacy*, vol. 1, no. 2, pp. 223-231, 1998.

[7] 楊允言, "台文華文線上辭典建置技術及使用情形探討", in *2003 第三屆全球華文網路教育國際學術研討會*, 2003, pp. 132-141.

[8] 楊允言 and 劉杰岳, "台語文線頂辭典 kap 語料庫簡介", in *語言、社會與文化系列叢書之二 語言政策的多元文化思考*, 鄭錦全等 Ed. 台北 : 中央研究院語言學研

究所, 2007, pp. 132-141.

[9] U.G. Iunn, K.G. Lau, H.G. Tan-Tenn, S.A. Lee and C.Y. Kao, “Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods,” *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 12, no. 4, Dec. 2007, pp. 349-370.

[10] 詞庫小組, “中文詞類分析 ,” 詞庫小組, 台北, 台灣, Tech. Rep. no.93-05 , 1993.

[11] 楊允言, “台語白話文學ê全新表現——台語文數位典藏資料庫簡介,” *台灣文學館通訊* 15 期, 台南 : 國家台灣文學館, pp42-44, May. 2007. [Online] Available: http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Phenglun/DADWT/dadwt.asp [Accessed: Jun. 30, 2008].

[12] 楊允言, “台語文/華文辭典,” Dec. 2000. [Online]. Available: http://iug.csie.dahan.edu.tw/q/q.asp [Accessed Jun. 30, 2008].

[13] 劉杰岳, “全羅漢羅對照文本找華語候選詞” Aug. 2007. [Online]. Available: http://iug.csie.dahan.edu.tw/nmtl/dadwt/pos_tagging/clhl_hoagi_hausoansu.asp [Accessed Jun. 30, 2008].

[14] C. Samuelsson, “Statistical methods,” in *the Oxford Handbook of Computational Linguistics*, R. Mitkov, Ed. New York: Oxford Univ. Press, 2003, pp. 358-375.

[15] A. Ratnaparkhi, “A maximum entropy model for part-of-speech tagging,” in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, 1996, pp. 133-142. [Online] Available: http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf [Accessed: Jun. 30, 2008].

[16] A. McCallum, D. Freitag and F. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation,” in *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford Univ., Jun. 2000, pp. 591-598. [Online] Available: http://www.cs.umass.edu/~mccallum/papers/memm-icml2000.ps [Accessed: Jun. 30, 2008]

[17] Y.F. Tsai and K.J. Chen, “Reliable and Cost-Effective Pos-Tagging,” *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 1, Feb. 2004, pp. 83-96. [Online] Available: http://rocling.iis.sinica.edu.tw/CKIP/paper/

Reliable_and_Cost-Effective_PoS-Tagging.pdf [Accessed: Jun. 30, 2008]

[18]戴嘉宏, “台語選詞跟詞性” Jun. 2007. [Online]. Available: http://140.109.19.105/ [Accessed Jun. 30, 2008].

[19]楊允言,劉杰岳,戴嘉宏, “台語文斷詞、詞性標示系統” Aug. 2007. [Online]. Available: http://iug.csie.dahan.edu.tw/TGB/tagging/ [Accessed Jun. 30, 2008].

中文名詞組的辨識：監督式與半監督式學習法的實驗
Chinese NP Chunking: Experiments with Supervised,and Semi-supervised Learning

林晏僖 Yen Hsi Lin
國立台灣大學資訊網路與多媒體研究所
r95944002@ntu.edu.tw

高照明 Zhao Ming Gao
國立台灣大學外國語文學系
zmgao@ntu.edu.tw

高成炎 Cheng Yan Kao
國立台灣大學資訊工程學系
cykao@csie.ntu.edu.tw

摘要

本文先利用 Taku Kudo 所發展的 SVM 工具 Yamcha 訓練中文名詞組辨識的初始模型，並嘗試以不同於多數文獻的 IOB 表示法及前二後二位置的語法標記資訊，找到適用於中文的參數。接著利用半監督式學習法中自我學習的概念，利用網路上未標記的資料，強化 supervised-learning 的模型。實驗結果證明，supervised learning 這個步驟裡，我們選用的參數比前人的更合適；而我們所提出的半監督式學習法，可以提昇辨識結果，特別是在動詞修飾名詞的情形，半監督式學習法可以大幅提高辨識的正確率。

Abstract

This paper utilizes Yamcha, a SVM tool designed by Taku Kudo, to train an NP-chunking model for Chinese. In addition to IOB and two words surrounding the focused word, we experimented on new features and exploited unlabeled data from web pages to enhance the previous model. Our experiments with supervised learning indicate that our chosen feature sets outperform those reported in previous studies. In addition, the proposed method of semi-supervised learning is proved to be effective in distinguishing a noun phrase from a verb phrase both consisting of V N combination, thus enhancing the overall accuracy.

關鍵詞：名詞組辨識、YamCha、監督式學習、半監督式學習

Keywords：NP-chunking、YamCha、supervised learning、semi-supervised learning

一、緒論

名詞組的辨識一直以來都是自然語言研究及其相關領域，如網路探勘（web mining）、文件分類（text categorization）等非常關鍵的一個步驟。在自動問答系統（question answering）中，關鍵詞多半以名詞搜尋爲主。在自然語言問答系統裡，名詞組的辨識也是不可或缺。我們每天都使用的搜尋引擎，大家輸入的關鍵詞以及搜尋引擎統計出來的熱門關鍵詞亦以名詞組居多；大至搜尋引擎背後大量的資料庫、小至普通文本建檔而成的資料庫，製作索引分類時，名詞組的使用也多於動詞、副詞；也有越來越多網頁都在針對網頁中的名詞組做自動偵測以及外部連結 ... 除了這些例子之外，語意角色標記（semantic role labeling）、專有名詞辨識（name entity identification）、文章處理中的回指（coreference），名詞組的辨識都是一個重要的步驟，因此有好的 NP chunker，可以改善許多 NLP 的研究成果以及相關應用。

在這篇論文中，我們先用 Taku Kudo 所提出利用 SVM 的演算法當作一開始的模型，除了許多參考文獻中常用的 IOB 標示法以及位置，我們還嘗試了以不同標示法以及加入不同位置的句子部份資訊當作特徵，證明對於中文的處理，不論是封閉或開放實驗中，IOE 表示法和加入前後兩個位置的詞及中研院簡化標記，是我們利用中研院句法樹庫 Sinica Treebank 所能得到最好的結果的參數。接著，我們利用一個沒有句法結構訊息的大型語料庫 word sketch engine 中的句子，加上半監督式學習法中，自我學習的概念，利用網路上大量未標記的網頁，來彌補 Sinica TreeBank 裡不足的訊息，改善利用監督式學習法實做出的 chunker。

實驗部份，除了封閉測試外，由於中研院樹庫圖中資料有限，我們額外收集了不同類型的句子當作開放測試的語料，以分別比較兩種作法在名詞組辨識的效果及限制。實驗結果顯示，我們選用的參數較前人選用的參數做出的模型在第一階段開放測試中高出了 16 個百分比，在第二個開放測試中也有 70%的 f-rate；加入 unlabeled data 這個步驟的半監督式學習法，也的確提昇監督式學習法的效果，使開放測試的 f-rate 提高至 78.79%，不但保存分類器的優點，也明顯提昇中文在難解的名物化歧義的名詞辨識結果。

接下來的章節中，第二章爲文獻回顧包含 Chunking, SVM, 半監督式學習法的基本介紹；第三章及第四章分別爲實驗方法說明和數據結果討論。最後爲結論與未來展望。

二、 文獻回顧

（一）、規則法

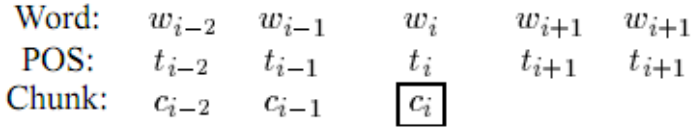
Abney(1995)利用了有限狀態機做出的規則式剖析器。他的實驗利用語意的訊息例如字形變化來當作特徵。他對英文及德文做了測試都成功並且快速的取出主要類型，包括動詞、名詞、介係詞的詞組。不過他還是強調選取的文法的重要性。Kinyon(2000) 提出一個適用於不同語言的 rule-based chunker。即使缺乏大量的訓練語料，只要有一些能夠辨別結構邊界的規則，就能使用 Kinyon 所提出的方法。Igor (2005)比較利用 NLTK 工具實做完成的 rule-based chunker 以及利用 TnT(Trigram and tag) 統計方式這兩種方法做出的 chunker 在名詞組和動詞組上的辨識效果，實驗證明近來鮮少被大家採用的規則方式實做出來的 chunker 不但沒有比利用統計的 chunker 遜色，甚至在召回率（recall）及 f-rate 的表現上要來的更好。在中文方面，雖然 Zhao 等(1999)對某些類型的詞組整理出結構的規則，但 Zhao 等(1999)還是捨棄規則式的作法，使用記憶基礎學習（memory-based learning）的方式。他們的實驗顯示若不加上詞彙本身的訊息，而只有利用詞性的訊息下，效果會比較差。

（二）、監督式學習及統計方法

在大規模語料庫建立之前，名詞組辨識常利用組成名詞組結構規律透過有限狀態機找出符合的模式 pattern，或從標記好詞性的語料庫以統計方式得到，或結合語言規律及語料庫統計；隨著賓州大學樹庫圖（University of Penn TreeBank）開放給大家使用之後，詞組辨識也朝向以機器學習的方法來解決：Skut and Brants(1998)、Koeling (2000) and Osborne (2000)使用最大熵演算法；Park and Zhang 採用規則以及記憶學習（memory-based learning, MBL）綜合的方式；Kudo and Matsumoto(2000,2001)利用 8 個 Support Vector Machine（SVM）系統投票（voting）的方式得出 chunking 模型，其他利用監督式學習（supervised learning）的方法還有 Hidden Markov Model(HMM)（Li (2004)）、transform-based learning(Ramshaw and Marcus (1995))這幾種，大都是利用語料的結構及前後語境的特徵得到的。這些演算法也早已被用在其他跟自然語言處理有關的議題上。

1 、Kudo 的支持向量機演算法

Kudo 等(2000) 第一個將 SVM 有效利用在詞組辨識作業上。它利用周圍的詞、這些詞的詞性以及預測的詞組類別當作訓練、預測過程中的特徵集，利用 SVM 對每個詞做標記的動作。要辨識第 i 個字的詞組類別 Ci，Kudo 採用了如圖一的特徵：



圖一、Kudo 提出的演算法中所使用的特徵[2]

Wi 是出現在第 *i*-th 個位置的詞, *Ti* 是 *Wi* 的詞性 而 *Ci* 是第 *i*-th 個字的詞組類別標記。另外，他們把特徵集中的 (*Ci+1*, *Ci+2*) 換成 (*Ci-1*, *Ci-2*) 以達到反向剖析的效果。由於在測試時，詞組類別標記這個特徵 (正向剖析：*Ci-1, Ci-2*；反向剖析：*Ci+1,Ci+2*) 並不是事先給定，而是利用當下模型決定的結果，因此被稱爲動態特徵；相對的 *Wi* 和 *Ti* 則爲靜態特徵。給定一個句子，例如：這/是/詞組/範例/標記，表一是對應的範例向量，其中 B 和 O 分別表示該詞是名詞組的開始或不在名詞組內。

表一、「這是詞組範例標記」在 Kudo 演算法中對應的範例向量

關注詞的類別	*Wi*	*Wi-2*	*Wi-1*	*Wi+1*	*Wi+2*	*Ti*	*Ti-1*	*Ti+1*
B	1:這	1:0	1:0	1:是	1:詞組	1:NES	1:0	1:SHI
O	1:是	1:0	1:這	1:詞組	1:範例	1:SHI	1:NES	1:NA
B	1:詞組	1:這	1:是	1:範例	1:標記	1:NA	1:SHI	1:NA

張席維、高照明與劉昭麟(2005)仿照 Taku 提出的這個演算法，以 Sinica TreeBank 當作語料，訓練中文的模型。雖然 Taku 在英文的實驗結果有 94%左右的正確率，張席維等只得到了 87.43%。

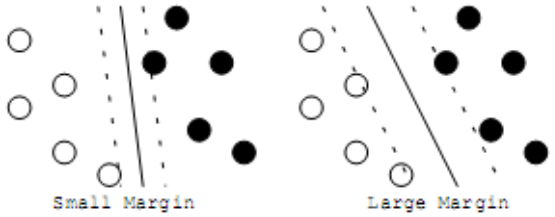
（三）、SVM 以及 YAMCHA

支持向量機（Support Vector Machine）是一個目前被廣泛運用在分類問題上的數學工具，是根據 Vapnik 的 max margin strategy 發展出來的分類器。相較於其他傳統分類器，如：決策樹學習(decision tree learning)、最大熵法(maximum entropy)等，SVM 有以下明顯的優點：

(1)即使在高維特徵向量空間下還是能產生好的效能。

(2)核心函數能將資料映射到更高維的空間而沒有增加計算複雜度。

支持向量機主要的想法是製造一個最佳的平面可以讓訓練範例向量分成兩個類別（positive and negative）並且把這個平面的邊界最大化。圖二中，黑色實線就是兩個可將資料分成兩類的平面，兩條虛線中間的距離就是邊界（margin），也就是 SVM 演算法試著最大化的目標。在虛線兩邊的點稱爲支持向量（support vectors），而且只有在訓練集中的支持向量會影響整個模型的結果。雖然 SVM 的分類準確度十分驚人，計算複雜度跟其他機器學習方法比起來也相對的高了許多。在需要龐大的訓練語料集的狀況下，利用 SVM 的訓練過程不但不夠有效率，甚至有可能因爲需要的訓練時間太久這種因素，實際情況下無法看到成果。



圖二、兩種可能將資料分開的超平面[2]

YamCha（Yet Another Multi-purpose Chunking Annatator）是 Taku Kudo 基於 Taku 等(2000)中的演算法，設計專門用在解決詞組辨識、詞性標記甚至文件分類等自然語言處理應用的工具。整個架構採用的分類方法是 SVM。跟單純的 SVM 分類器不同的地方是，Yamcha 要求的輸入檔案格式比較符合人直觀的想法，把需要做詞組分類的資料如同圖三，每個詞會利用到的特徵簡單的排列，直接交給 YamCha 去執行即可。如果不做任何參數的變動，這個工具就用 Taku 中一樣的預設值，把 (n-2, n-1, n, n+1, n+2) 位置上的字和特徵都當成關注詞(第 n 個詞)的特徵集去訓練。所以跟傳統分類器不同的地方，只在於 YamCha 幫使用者處理了資料格式的問題。另外值得注意的一點是，雖然 SVM 分類的效果非常的好，但其耗費的計算量以及時間也比其他分類器來的大，而 YamCha 在這點上做了改進，使得訓練時間以及分類時間都加速了至少三倍以上。

He	PRP	B-NP
reckons	VBZ	B-VP
the	DT	B-NP
current	JJ	I-NP
account	NN	I-NP
deficit	NN	I-NP

圖三、Yamcha 輸入資料的格式，與 CoNLL 2000 shared task 相同。其中 B, I, O 分別表示該詞是某種詞組的開始，內部，或不在詞組中。

（四）、半監督式學習法

如同表面上的字義一樣，半監督式學習法介於監督式學習法和非監督式學習法之間：利用大量未標記過的資料結合一些已經標記過的資料來做訓練的模型以解決資料量稀少及分散的問題。而在一個自然的考量下，如果有一樣數量標記過的資料（labeled data），我們是不是能利用大量容易取得的未標記過、未處理過的資料（unlabeled data）來建造一個更精確的分類器（classifier）？ 這個問題通常就會被歸類於半監督式學習（semi-supervised learning）。在真實生活中，標記資料不但耗費時間、人工、甚至金錢；相對的，未經過標記的資料量多而且隨手可得。因此在機器學習的領域上，如何利用未標記資料是一個重要的課題，例如：我們可以利用程式取得網路上大量的網頁存檔，卻需要人工處理才能做正確的分類；在語音辨識的研究中，要取得大量的錄音檔非常簡單，但是要去標記這些錄音檔則需要大量的時間和人力去打逐字稿，在這些情況下，如果 unlabeled data 確實能讓模型的效能提高，半監督式學習是非常的有幫助的。

在半監督式學習中，資料型態和不同演算法的模型的配對是十分重要的，不然也有可能造成反效果。EM with generative mixture model、self-training、co-training、TSVM 還有基於圖形的演算法都是 SSL 裡常用的方法並且各有優點。假如標記的類別可以把把資料分開得很明確，使用 EM with generative mixture model 比較好；如果特徵群就足夠隨著資料被分成兩半，那麼 co-training 的方法會比較適合，因爲這個演算法就是對分開的特徵做一些假設，並在不同的特徵集上使用不同的學習工具。若有相同的特徵的

點會被分在同一個類別，而目前的模型也無法被改進時， Mincut, Boltzmann Machine, Tree-based Bayes 等基於圖形的方法會比較適合。

最早引進利用未標記過的資料這個概念應用在分類上的也許就是自我學習這種演算法。一開始只利用少量標記過的資料作訓練，接著利用當下的決策函數（decision function） 從未標記的資料中找出符合的點，加進本來標記過的訓練集，重新訓練並找出新的決策函數。直到在未標記的資料中無法再找到可標示的資料或是整個情形超過一些閥值（threshold）。

Yarowsky（1995）是在提到自我學習時常被引用的有名的例子。Yarowsky 利用這個方法做分辨語意歧異度，跟只用標記過資料和監督式學習法的效果比起來改進很多。在一個篇章中，要如何決定 plant 這個字是植物的意思還是工廠的意思？Yarowsky 的假設是：

（1）一個詞在一章篇幅中只會有一個意思,，（2）一組搭配語中也只會有一個意思。他先從未標記過的資料中，取小量（約爲訓練集的 2%）的句子出來標記答案，例如：句子中的 plant 是植物則此句爲 class A，是工廠則爲 class B。根據（1）的假設，找出同類別句子中的搭配語，例如：一開始選定的 A 中的句子的 plant 旁邊都有 life 這個字，而 B 的 plant 旁邊都有 manufactoring 這個字，利用這個特徵到未標記過的資料中去找有同樣特徵的句子，收進訓練集裡。也同時利用一些決策函數在這些新標記好的句子中找新的搭配語。接著利用（2）的假設，如果同一篇文章中的多個句子都已經被歸到同一個類別，則同篇文章中剩下的句子都可以分到同一個類別，這個假設不但可以擴大訓練集，還可以修正在前面的步驟被歸類錯的句子。重複擴大訓練資料、利用新的資料訓練模型的步驟直到未標記資料的數量不再有太大的變化爲止。實驗結果證明，這個方法訓練出來的模型跟只利用標記過資料及監督式學習法的效果比起來，確實將效能提昇並且少了很多人工標記的動作。

三、作法說明

（一）、名詞組表示法

Inside/Outside： Ramshaw and Marcus(1995)提出並使用了下述三種 class（IOB）表示一個詞在詞組中的位置。

I：這個詞在某個詞組之中

O：這個詞不屬於任何詞組

B：這個詞是緊接著別的詞組的詞組開頭

這個表示法被 Tjong Kim Sang 稱爲 IOB1，另外他還提出了 IOB2/IOE1/IOE2：

IOB2 中 B 是任何詞組的開頭；

IOE1 中 E 是緊鄰著別的詞組的詞組結尾；

IOE2 中 E 是任何任何詞組的結尾；

表二是各種表示法的範例說明。另外還有 start/end 的表示法，由於在 Taku(2000)中實驗結果不佳，因此本實驗中並沒有用到，也不加以介紹。

表二、以各種表示法標記「這是詞組範例標記說明」

	IOB1	IOB2	IOE1	IOE2
這	I	B	I	E
是	O	O	O	O
詞組	I	B	I	I
範例	I	I	I	I

標記	I	I	E	E
說明	B	B	I	E

（二）監督式學習法：Supervised-learning

我們仿照[1]及[2]中的演算法（在 2.2.1 中也有描述），我們將前後兩個詞及詞性分類的訊息及前面兩個已經判別好的詞分類當作特徵集讓 SVM 分類器參考，利用 SVM 分類器判別每個詞是屬於 IOB 中的哪一類。例如「這是一個例句」中有哪些名詞組？要判別「一」這個詞時，將前後兩個詞及個別的詞性標記 這、NEP、是、SHI、個、NF、例句、NA、前面兩個已經判別出來的詞組分類 E（這）、O（是）以及本身一、NEU 這 12 個當作分類的特徵值。最後這五個詞分別的類別可能爲

這	是	一	個	例句
E	O	I	I	E

則可以利用 IOE 分出詞組間的邊界，判斷出此句的名詞組爲「這」及「一個例句」。 因此監督式學習法可說是希望能找出有最佳辨識效果的特徵集。

（三）半監督式學習法

從[1]及監督式學習法的實驗可以看出，只從樹圖資料庫的資訊加上查詢詞典的詞義對我們判斷動詞的名物化現象並沒有幫助，於是我們希望能利用外部的資源幫助我們獲得一些新的訊息，靈感來自於前面提到的 Self-training。
從訓練語料的句子裡，我們對動詞後面接一個名詞的組合做觀察之後發現一些規則，下表是例句及推測：

表三、語料中部份句子及其特性

例句	推測
導遊 發 給 每人 一 ***本*** *導覽 手冊*。	量詞的後面常常會接名詞組。
這個 報導 給予 我 們 無限 ***的*** *想像 空間*。	"的"的後面常常會接名詞組。
大家 看 ***了*** *宣導 短片 之*後 有 什麼 感想 呢？	時態詞的後面常會接名詞組。

表四、從語料庫推測出來的規則

可能出現在動詞後面的詞類	NEQA、NEP、NEU 等量詞
	NG : 時間後置詞
	NC : 位置詞
	NH : 代名詞

先從訓練語料中任選當修飾的動詞接名詞(例如：採購人員、運動精神)，以及動詞加上受詞(例如：祭拜祖先、來自家人) 的這兩種情形，各 100 組 Bi-gram，從 word sketch engine 中搜尋 50 句包含這些 bigram 的句子，如下面幾句是包含採購人員的句子：

...團體 *之 採購人員*。
...需要仰賴的不只有 *總務 採購人員 自身*。
...又蘊含 *了 採購人員 對* 他人和社會...
...大多數 *的 採購人員 會* 蕭規曹隨...

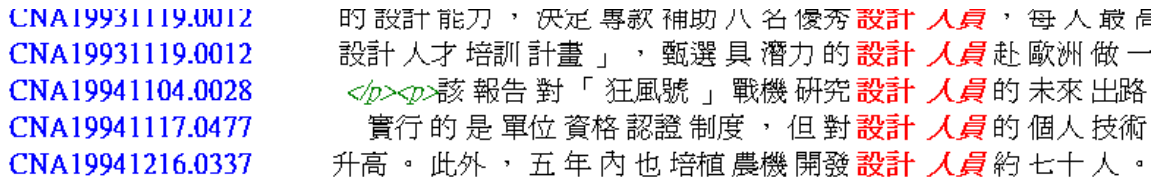
蒐集到某個數量的句子之後，我們對關注的詞組前後緊鄰的詞類做統計。即使收集來的句子前或後會連接許多不同的詞類，我們還是可以分別對這兩種情形歸納出一些可

能的共通點，如下表：

表五、在不同功用的動詞前的詞類比較

類型	可能出現在前面的詞類
動詞	D：副詞
修飾用的動詞	DE：的，之等、DI：時態標記、NEU, NF 等量詞

接著我們進行初步小規模的測試，從封閉測試語料中，選擇動詞後面立即接一個名詞(V1 N1)的組合，一樣的從 WSE 中蒐集句子，接著統計這些句子中緊鄰詞的詞類，再利用上表先前推測出的特徵判斷（V1 N1）是哪一種情形。例如我們蒐集包含「設計人員」的句子（圖四）：



圖四、WSE 中截取出來的例句

統計過後的結果，前面出現"DE"、"DI"及量詞的機會（次數）比出現副詞"D"的機會（次數）高，因此將設計人員歸類爲修飾用的動詞加名詞組合而成的複合詞（class A）。而包含「採購汽車」的句子中，前面緊鄰副詞"D"的機會比緊鄰"DE"、"Di"及量詞的高，則將其歸類於動詞接受詞這類（class B）。還有一種情形是，如「服務社會」這個詞組，在 WSE 裡找不到一個句子精確的包含這個 bi-gram 或是只收入了四五句相關的句子，也就是資料不足（class C）。

針對在 WSE 裡面，沒有出現過的詞語組合，例如：服務社會，或出現次數不夠多，拿資料來分析似乎不夠客觀的詞組，例如：紅燒牛肉，我們利用 google 搜尋引擎尋找內文裡有符合一樣詞組的網頁，得到如同 WSE 中包含關注詞組的短句。跟 WSE 一樣，我們收集前 50 個 google 傳回的 Snippet（各個網頁的摘要），只是在 WSE 裡的句子有做過斷詞以及有部份的標記，而從搜尋引擎得來的是沒被分析過的資料（raw-data）。從 Google 得到資料並經過中研院分詞程式處理後，我們利用跟在 WSE 中作法雷同，統計關注詞組前面的詞的類型。在我們剛剛提到的功能詞的長度大部分都是一個詞或兩個詞，因此我們查詢字典中，關注詞組前面的一個詞以及兩個詞的詞性，查看是否是符合我們假設中模式的詞並做統計。

由於這兩個利用外部語料庫資料的步驟，只能將某些 POS 本來屬於動詞的詞彙功能做分類，因此這個實驗的結果將被當成一個特徵加入 SVM 工具中一起訓練，SVM 工具還是整個標記名詞組過程的核心。

因此此時模型採用的特徵成爲 IOE, Wi-2 .. Wi+2, Pi-2..Pi+2, Vi, Ci-2..Ci-1,

$$\text{其中 } Vi = \left\{ \begin{array}{l} \text{C, Wi 的 pos 不屬於動詞類別} \\ \text{A, Wi 是被名物化的動詞} \\ \text{B, Wi 是動詞} \end{array} \right\}$$

四、實驗結果討論

（一）、實驗語料介紹

本文的實驗中，監督式以及半監督式學習法所需要的訓練語料，使用的是中研院中文句結構樹資料庫 Sinica Treebank3.1。這個資料庫的文章來源分別有直接從平衡語料庫

中取出的文章、國小課本、光華雜誌以及中研院語言所的語音平衡檔案，再經過電腦剖析及人工校對做成樹圖庫。全部包含了六個檔案，分別爲不同的背景、情境，共有 65434 個中文樹圖、392237 個詞（平均一句包含了六個詞）。我們把資料庫中每個檔案的 70%取出整合來當作訓練語料、30%當作封閉測試的測試語料。
檔案中的句子都以下句的形式表示，除了可從中得知結構訊息之外還有中文的語意角色。句子中的詞類標記，是 CKIP 詞類標記，與中研院的字典所使用的詞性標記是同一種（另外有還有簡化標記、精簡標記兩種）。

```
# S(agent:NP(Head:Nca:觀光局)|evaluation:Dbb:還|quantity:Daa:另|Head:VE12:安排|aspect:Di:了|theme:NP(property:NP(quantifier:DM:幾處|Head:Ncb:市郊)|property:Nv4:遊覽|Head:Nac:活動))# 。(PERIODCATEGORY)
```

在訓練語料中，約有 45000 個句子，依照 Church 的定義，所取出的 NP chunk 共有 65009 個，但是每個 chunk 平均只包含了 1.57 個詞。可見得在中研院樹庫圖裡標記得 np-chunk 還是以單詞居多。從上一個句子中，依照樹庫圖中的結構，「觀光局」「幾處市郊」「活動」都成爲一個單獨的 chunk，但少了我們認定的「遊覽活動」甚至有可能是「幾處市郊遊覽活動」。如果只利用 non-recursive chunk 的結構從 Treebank 自動抽取答案出來，會得到很不理想的訓練語料，因此在標記詞組答案上，我們對幾種結構做了修改，標記出較符合我們思維的答案，包含
(1) 名詞+的+名詞 ；(2) 形容詞+的+名詞；(3)量詞+的+名詞

（二）、相關資源介紹

在之後的實驗中，我們使用到一些廣爲人知的工具及資源：

（1）中研院斷詞程式：輸出包含分詞結果，與每一個詞的詞性標記，並可以處理未知詞。

（2）詞彙特性速描系統（Word Sketch Engine）[12]：這是一個包含了中文、英文、法文、德文等多種語言的大型語料庫，並且已經對這些語料做了同義詞、用語索引、搭配語分類的整理。這個語料庫中的句子，是標示好的資料，除了已經做好斷詞，也可以查看詞性標記。如下列兩句：

```
稅捐處 工商 稅科 、 財產稅科 、 稽徵 科 及 稅務 管理 科 等 依照 權責 ，將 分別 全面 查緝 逃漏稅 。
在/P21 賦稅/Naeb 方面/Nac，/COMMACATEGORY 查緝/VC2 逃漏稅/Na 及/Caa 進行/VC2 會計師/Nab 評鑑
```

在詞性標記這部份與中研院斷詞程式不同的地方在於，中研院斷詞程式似乎參考了句子的語境標示每個詞的 POS 而 WSE 沒有，所以在不屬於功能詞（function word）的部份，也就是一個詞可能會有多種詞性的情況下，WSE 的詞類標記的精確度會比較差，並且比較不適合拿來當參考。

（3）Google Soap API：讓使用者合法做關鍵字搜尋，每天最多可做 1000 筆的搜尋，並提供回傳網頁的相關資訊。。

（三）潛在問題

1. 相較於英文有 CoNLL2000 Shared Task 的 Chunking 規格、資料集和答案，中文這方面的訊息並不一致，多數還是從標記好的樹庫中抽取初以標記好的 chunk（Xia 等(2000)）或是自己再從樹圖標示的結構定義及標記 chunk 當作訓練語料的答案（Li (2003), Zhao and Huang (1999)）。若同 CoNLL2000 shared task 按照 Church(1996)在英文中的定義，將 chunk 視爲沒有包含其它種 chunk 的詞組，也就是不重疊（non-overlap）、不遞迴（non-recursive）

的詞組合而成的，即 NP-chunk 可簡單的想成不包含別種詞組的名詞組。但回頭看語料庫中標記好的 NP-chunk 有絕大部分是屬於單詞，反而不符合我們一般的思維（4.1 節），因此若利用語料庫標記好的 NP-chunk，除了很難將全部的答案改成如同我們所希望看到的規則，還會有下面提到的長詞組的問題。

2. 由於中文是一種沒有屈折語素（inflectional morpheme）的語言，例如英文中被動式的動詞會有一個變化型，轉爲名詞的用法則字尾變成 ing、加上 tion 等等，但是在中文裡則是加個「被」字以表達被動式，其他的情形在前後不一定有加入的關鍵詞，必須由對中文有一定了解程度的人自己對語境做推測來判斷每個詞的詞性及功用。例如：從下面這個句子

The experiment involved the *combining* of the two chemicals。

可以很清楚的看出 combining 是名詞的用法，但是在以下這兩個句子，無法直接看出進口和喜愛的詞性。

政府編定汽車管理制度使進口汽車得以合法化。 他深得學生的喜愛。

這也是張席維等、Ding 等(2005)中提及的名物化現象。由於在 Sinica TreeBank 裡有這種現象的詞組只有不到 3000 組，因此張席維等也根據實驗結果強調利用監督式學習法辨識中文的名詞組時，能否找出被名物化的動詞是一個提昇正確率的關鍵。

3. 從 Sinica TreeBank 中取出標記好的 chunk 的平均長度不到兩個詞，Cheng 等(2005)提到中文實際上有非常多由數個名詞組合而成得名詞組，例如：行政院/國家/科學/委員會、電腦/人體/模型...等等，這些在日常生活中都不是令人陌生的詞語。因此使用 Sinica TreeBank 當做一種 gold standard 或是訓練語料時，很難解決長詞的問題。

（四）開放測試集

由於上面說明了很多在訓練及封閉測試語料中無法觀察到的情形，因此我們利用開放測試的結果作爲不同方式設計的模型間比較的準則。在監督式及半監督式學習方法也各有一組的開放測試資料作爲該次實驗內的參數比較。監督式學習法中的開放測試語料，大部分是包含"形容詞接名詞"、"量詞接名詞"、"量詞接形容詞接名詞"、名詞中有所有格的句子，例如：這是最新的車款、事情發生在去年的夏天、班上有一名天才學生..等等；半監督式學習法的開放測試語料強調動詞的判別，因此測試語料中的名詞組包含一些已轉化爲別的作用的動詞，例如："他在拍賣網站上買東西"，"警察透過銷贓管道抓到小偷"等等。

實驗結果我們採用與 CoNLL 2000 shared task 一樣的評量方法，直接利用他們提供的評量工具，分別算出詞標記正確率（tag accuracy）以及詞組正確（Precision）、詞組召回率（Recall）以及 F-rate = 2PR/P+R，而 F-rate 還是爲主要考量。

（五）監督式學習法 Supervised-learning

在[1]中，作者利用 IOB 表示法來做名詞組的標記，並指出簡化標記及精簡標記對名詞辨識的影響度；當其餘的詞類使用大分類，而保留簡化標記的動詞次分類時可使學習效果提昇。樹庫中的 CKIP 詞類標記比簡化標記的分類更細，也代表 CKIP 的詞類標記透露出更多的語言訊息，這樣是否能讓 SVM 的學習效果更好呢？另外，由於中文的名詞詞組的中心語（Head）傾向出現在最後面（head-final），那麼 IOE 的表示法是否比 IOB 的恰當？因此我們先針對詞性標記以及名詞組表示法做選擇。(CKIP 詞類標記和簡化標記的對照和代表意義可參考 http://godel.iis.sinica.edu.tw/CKIP/paper/poslist.pdf)

表六是封閉測試的數據，從此表中可以發現這四個模型之間並沒有明顯數字上的差距，

因此我們轉向開放測試的結果。下列是一些開放測試的句子：

我們買了一張很貴的票。

我聘了一個很優秀的職員。

阿忠的那一間房子。

表六、比較 IOB,IOE,CKIP,simplified（簡化標記）四種特徵在封閉測試時的結果

feature combination	tag accuracy	precision	recall	f-rate
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset , T(n-2 .. n-1), IOB	91.21%	84.85%	86.98%	85.90%
W(n-2 .. n+2) , P(n-2 .. n+2) in CKIP tagset, T(n-2 .. n-1), IOB	90.89%	84.44%	86.60%	85.50%
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset, T(n-2 .. n-1), IOE	92.06%	84.65%	86.28%	85.46%
W(n-2 .. n+2) , P(n-2 .. n+2) in CKIP tagset, T(n-2 .. n-1), IOE	91.93%	84.34%	86.09%	85.20%
W(n-2 .. n+2) , P(n-2 .. n+2) V in CKIP, others in Simplified, T(n-2 .. n-1), IOE	92.11%	84.67%	86.23%	85.44%

由於[1]已經說明監督式學習訓練出來的模型對名物化詞類沒有好的辨識效果，因此在這部份的開放測試，我們先選擇一些句子裡有名詞組中基本形式，像量詞接名詞、形容詞接名詞、量詞接形容詞加名詞，加上一些包含長複合詞組以及有所有格的句子。表七是這個實驗的結果。

表七、IOB IOE 初步開放測試比較結果

feature combination	accruracy	precision	recall	f-rate
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset , T(n-2 .. n-1), IOE	92.95%	86.67%	89.66%	88.14%
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset , T(n-2 .. n-1), IOB	88.93%	68.75%	75.86%	72.13%

從表六中，我們可以看出（1）CKIP 詞類標記總共有多達約 230 個分類，簡化標記約有 45 個分類，因此不論是只有動詞採用 CKIP 的次分類或是整體的詞類都利用次分類來標記，可能由於 CKIP 詞類分項太細造成分類器中資料稀疏的問題，使得採用 CKIP 詞類標記的表現並沒有以用簡化標記的表現好，以及（2）雖然在封閉測試中兩種表示法精確度不相上下，從開放測試的表七來看，以 IOE 表示法的結果會比 IOB 表示法來的精確，所以我們固定這兩種參數來進行之後的實驗（包括半監督式學習法）。利用簡化標記的另一個好處是：當我們用中研院的斷詞程式對開放測試的資料做前處理時，得到的標記與訓練出來的模型使用的一致。由於我們從這個開放測試的結果發現：

1. 初始模型對長詞的偵測不太敏感：語料中有"一名騎機車的年輕人"及"一名高級官員"。這兩句中的「一名」都是某個 NP 的一部分，但是前者的詞組標記爲一 O/ 名 O 而後者爲一 I / 名 I。因此量詞後面被 tagging 過程考慮進來的詞性變得十分重要。由於我們初始模型只將前後各兩個詞加入特徵集，在開放測試裡「有一間很漂亮的教師休息室」的「一間」就有辨識錯誤的可能，因爲在量詞後面的兩個詞都還不見名詞的蹤影。
2. 即使有一些句子有著非常類似的形式（例如：很漂亮的衣服，很貴的票），但是模型輸出的結果卻不相同，而我們發現這或許是因爲 "很漂亮" 在訓練集中出現過爲名詞一部分的用法，而"貴"在訓練集裡只有當動詞用。

因此我們考慮：

1. 往前後看不同長度的詞及語意特徵作爲特徵集；
2. 某個位置的詞及語意特徵不必同時存在。

表八是特徵集的符號以及其代表的意義；

表八、實驗中採用的特徵代表符號以及相對的意義

代表符號	代表意義	代表符號	代表意義
Wn	第 n 個詞	IOB	利用 IOB 表示法標記名詞組
Ln	第 n 個詞的 POS	IOE	利用 IOE 表示法標記名詞組
Tn	第 n 個詞的 tag 標記	Hn	第 n 個詞在 hownet 中的義元
Simplified	簡化標記	CKIP	CKIP 標記
F	forward parsing	B	backward parsing

表九是不同模型所使用的特徵組合，由於這些組合在封閉測試的結果都十分相近，所以不將數據一一列出；

表九、模型及其利用的特徵對照表

model 1	F, Wi-2.. Wi+2, Pi-2..Pi+2 Ti-2..Ti-1
model 2	F, Wi-2.. Wi+2, Pi-2..Pi+2,Hi-2..Hi+2, Ti-2..Ti-1
model 3	F, Wi-4.. Wi+2, Pi-4..Pi+2 Ti-2..Ti-1
model 4	F, Wi-1.. Wi+1, Pi-1..Pi+1 Ti-2..Ti-1
model 5	F, Wi, Pi-2..Pi+2 Ti-2..Ti-1
model 6	B, Wi-2.. Wi+2, Pi-2..Pi+2 Ti-2..Ti-1
model 7	F, Pi-2..Pi+2 Ti-2..Ti-1

表十是各個模型開放測試的結果，從開放測試的輸出來看每個模型都有明顯不足的地方，「監察人員」這種型態的詞更沒有一個模型判斷正確。model 2 除了語料庫的內部訊息之外，加入了每個詞在 HowNet 中的義元（可以視爲語意特徵或類別）當做一個特徵，雖然比 model 1 的 F-rate 好上約 0.8 個百分比，訓練模型的時間卻也增加爲約 1.8 倍；model 3 參考前四個詞及後兩個詞，雖然對上面提到包含量詞的長名詞組合有些幫助，但是對其他問題沒有太大的影響；backward parsing(model6) 在 closed test 的部份，雖然有最高的 F-rate, 但是在開放測試的部份卻沒有特別好的表現。

表十、各模型開放測試比較結果

model	tag accuracy(%)	precision(%)	recall(%)	F-rate(%)
model 1	92.95	86.67	89.66	88.14
model 2	93.43	87.67	90.63	88.97
model 3	91.95	83.37	87.93	85.59
model 4	91.14	86.19	88.26	87.21
model 5	89.95	81.97	81.14	81.55
model 6	92.30	84.78	85.34	85.06
model 7	90.55	80.54	80.44	80.49

這階段開放測試的句子中，有很多詞（或詞性標記(pos)序列）是重複的，加入不同的搭配語或修飾語，或是以不同的語序組合，是測試模型對不同形式的語句的準確

度及穩定度。

除了受限於訓練語料這個問題之外，在訓練語料方面除了之前提到的缺點以外，歧義、名物化、未知詞這些實際生活中的現象，在語料庫裡是沒有標記的；由於語料庫的組成大多是長篇文章切成的句子，某些用法或語句，會因爲出自於同一篇文章而重複出現很多次，實際的訓練語料並不如統計過後數字上得多。雖然在這個語料庫裡共有六萬多句的句子，但是因爲分句是以標點符號爲原則，所以有很多句子其實只包含單詞，或只由名詞組成，或比正常情況書寫來的短，無法從中看出結構的訊息，反而可能變成分類器的雜訊（noise）。例如：「爸爸說：山路不難走」這個句子在語料庫中被分爲「爸爸/說」和「山路/不/難/走」兩句，但實際生活中多數的寫法還是會以長句子爲主。中研院句法樹庫中的詞及詞組標記由於經過人工校對，所以相當精確、可信度高，但是在實際測試時發現，有些詞的詞性標記利用中研院斷詞程式執行出來的結果，由於分詞程式分詞或標記錯誤或其它原因，並不會出現。例如，在語料庫中有一類 NV 的詞如電腦（NA）/打字（NV4）/及（CAA）/排版（NV4），而中研院斷詞系統並沒有辦法即時判斷出名物化(NV)的現象，因此斷詞後的結果爲「電腦(Na)　打字(VA)　及(Caa)　排版(VA)」；在訓練集中，有「兩（NEU）/者（NA）/同等重要」這樣的句子，中研院的線上程式斷詞的結果是，「兩者（NH）/同等重要」，顯現出開放測試和封閉測試的語料有一定的差異。

然而這個實驗的數據顯示，利用監督式學習法處理中文 NP-chunking 時，只有訓練語料中的資訊可以被利用的時候，IOE 名詞組表示法、簡化詞類標記、關注詞本身及前後兩個位置的詞，還有他們的詞性（pos）是最好的特徵組合。也因此我們固定這幾個特徵，當成下一個實驗的基本特徵集。

（六）實驗：半監督式學習法 Semi-supervised learning

我們對這個方法分別做了封閉及開放測試。由於此處著重在改進辨識名物化的動詞，這部份的開放測試語料除了上個實驗的測試語料，我們還加入包含不曾出現在訓練語料中的名物化動詞詞組的句子，佔全部測試語料的 50%。表十一是這個方法的實驗結果，supervised II 是將詞的前一個詞性單獨拿出拿作爲一個特徵。在封閉測試中，半監督式學習法實驗結果的 F-rate 爲 85.46%，雖然只比監督式學習作法的高出了 0.2%。但在開放測試的部份，半監督式學習法明顯比監督式學習法的 F-rate 高出了 8.79%之多。當我們只透過此實驗的作法描述中利用未標記資料判別封閉測試裡 VN Pair 中的動詞類別時約有八成的正確率（包含錯誤的被修正以及本來正確的維持正確），但是透過分類器的預測結果之後，VN pairs 這部份卻只有約 5%的改善，要如何有效應用分類器來突顯出新找到的特徵的重要性著實爲一個議題；由於轉爲名詞的動詞在樹庫圖中的詞類標記爲 NV 所以封閉測試語料中並不需要考慮表面上爲名詞接動詞的情形。開放測試中有 21 個名物化動詞的詞組，監督式學習法中正確判斷出六組，而半監督式學習法判斷出九組，剩餘的 12 組中有 7 組在作法描述中利用未標記資料判斷動詞類別時的分類是正確的，但經過分類器分類之後變成誤判的情形。

表十一、監督式及半監督式學習法實驗結果

	Tag accuracy	Precision	Recall	F-rate
封閉測試				
supervised	92.06%	84.65%	86.28%	85.46%
supervised II	91.76%	81.71%	86.05%	83.82%
semi-supervised	92.19%	84.85%	86.64%	85.73%
開放測試				
supervised	89.03%	67.31%	72.92%	70%

supervised II	83.83%	63.06%	69.23%	66%
semi-supervised	91.61%	76.47%	81.25%	78.79%

半監督式學習法中，我們只用了蒐集來的語料裡特定詞組前後的詞性做判別，但這與在監督式學習法裡將前後的詞性獨立分類(supervised II)作爲特徵有何不同呢? 由於在訓練語料中包含這些變化動詞的詞組約爲 5%，也就是訓練語料太少；另外句子中的前後詞性非常可能只是恰好在當句中出現在隔壁，而不是真正具有辨別作用的成分，如我們可能需要大量的 DE 或 DI 類的詞來佐證一個 VN 組合爲名詞的複合詞組，但是在語料庫中這個詞組的前面是個普通名詞，如同我們蒐集來的句子也有非常多是無法在我們統計過程中能被拿來參考的。因此半監督式學習法的成果還是比監督式學習法好。另外對於 WSE 和 Google 這兩個語料庫的比較上則各有優缺點。WSE 中的資料不但有斷詞還有詞類標記非常方便拿來直接使用，而從 Google 得到的則是 raw data，必須再經過斷詞的處理；WSE 中的句子經過挑選，形式較單純，而網路上的網頁還包含了關鍵字可能在標題、連結或搜尋到的網頁的重複性太高，可能有的特定的句子，現在很流行、或被很多人引用過，那麼出現在搜尋頁上很多筆都在描述同一筆資料，也就是同一個句子，而且在關注詞前面出現的詞，是在我們期待之外的 pattern、或無法讓我們利用的 pattern，分別形成錯誤資料太多、能用的資料太少的情況。因此無法在我搜尋的範圍內，有足夠符合我預期的模式的資料；在 WSE 中無法找到新的詞彙及不適合被收錄在語料庫中的詞彙，從 Google 得到的資料則沒有這個問題。

五、結論及未來展望

我們利用不同的特徵並將原本的模型加以改善過後，利用監督式學習法在小型開放測試有 70%的 f-rate，但在利用未標記過的網頁當作特徵之後，將模型的 f-rate 提昇至 78.79%，比原本高出了 8.79%。雖然模型的 performance 還是會受到訓練語料的影響使得結果不穩定，偶爾會有與預期不符合的情形發生，整體來說，我們提出的利用半監督式學習方法善用了網路上隨手可得的資源並且的確增進名詞辨識的效果。
本篇論文的貢獻在於：

(1)雖然詞組辨識一直是許多自然語言處理議題中的重要步驟，但在中文方面並沒有看到太多相關的研究。由於中文的名詞組結構相較於其它語言的名詞組結構都要複雜的多，因此本文只專注於名詞組的辨識，並且證明之前用在類似作法以及在其他語言中被普遍採用的特徵，並不完全適合用在中文語言上。而我們也找到了更適合的特徵。

(2)我們提出一個簡單的半監督式學習法，改善了監督式學習法中資料稀疏 (data sparseness) 及只依賴訓練語料時無法解決的問題。並且跟原本的 chunker 相比之下提高不少準確度及實用性。對於一些自然語言處理中，需要利用較高比例的名詞組的應用，例如：句子剖析 parsing、語意角色 semantic role labeling、文件分類 text categorization 等等，都有實質上的幫助。

未來我們希望能夠找到更好的特徵，加上監督式機器學習的方式，以解決更長的複合詞組以及包含有兩個以上名物化動詞的複合詞組。另外，由於網路上大量未標記過的資料隨手可得，因此我們也希望能提出非監督式(unsupervised)的演算法，以突破受限於少量人工標記過訓練語料的限制。

參考文獻

[1] 張席維，高照明，劉昭麟 （ 2005 ）利用向量支撐機辨識中文基底名詞組的初步研究。第十七屆自然語言與語音處理研討會。 pp. 317-332
[2] Kudo, Taku, and Matsumoto, Yuji. (2000). Use of Support Vector Learning for Chunk

Identification. In Proceedings of CoNLL-2000, pp. 142-144.
[3] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (pp. 189–196).
[4] Kudo, Taku, and Matsumoto, Yuji. (2001). Chunking with Support Vector Machine. In Proceedings of NAACL 2001, pp. 192-199.
http://chasen.org/~taku/software/YamCha/
[5] Chang, Chih-Chung and Lin, Chih-Jen. (2004) LIBSVM -- A Library for Support Vector Machines.[On line]. Available.
http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[6] Guang-Lu Sun, Chang-Ning Huang, Xiao-Long Wang, and Zhi-Ming Xu .Chinese Chunking Based on Maximum Entropy Markov Models. Computational Linguistics and Chinese Language Processing Vol. 11, No. 2, June 2006, pp. 115-136
[7] R. K. Ando and T. Zhang. A high-performance semi-supervised learning method for text chunking. In Proceedings of the Annual Meetings of the Association for Computational Linguistics (ACL), pages 1-9. 2005
[8] Semi-supervised learning book
http://www.kyb.tuebingen.mpg.de/ssl-book/
[9] Xiaojin Zhu, Semi-supervised literature survey, December 14, 2007
http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
[10] Yuchang CHENG and Masayuki ASAHARA and Yuji MATSUMOTO,“Machine Learning-based Dependency Analyzer for Chinese”, Journal of Chinese Language and Computing 15 (1): (13-24) ,2005
[11] CoNLL 2000 Shared Task http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.tx
[12] The Sketch Engine http://www.sketchengine.co.uk

強健性語音辨識中能量相關特徵之改良式正規化技術的研究

Study of the Improved Normalization Techniques of Energy-Related Features for Robust Speech Recognition

潘吉安 Chi-an Pan
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
s95323544@ncnu.edu.tw

杜文祥 Wen-hsiang Tu
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
aero3016@ms45.hinet.net

洪志偉 Jeih-weih Hung
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
jwhung@ncnu.edu.tw

摘要

隨著科技的發展，自動語音辨識技術也逐漸成熟，而達實際應用的階段；但當一自動語音辨識系統使用於現實環境中時，往往會受到雜訊的干擾，而造成辨識率大幅的下降；因此，環境相關的語音強健技術就顯得格外重要。本論文是針對在加成性雜訊所造成的辨識系統之訓練與辨識環境不匹配作爲主要探討的課題，除了概述許多語音特徵之強健性處理技術外，主要重點在於介紹我們所新發展的能量相關特徵強健化演算法—靜音特徵正規化法。在此，我們以較嚴謹的數學分析，探討加成性雜訊對能量相關特徵造成的失真現象；接著根據這些現象，我們發展相對應的一套新技術，即靜音特徵正規化法，來降低這些失真。透過這一系列的辨識實驗，證實我們所提出的新技術能夠有效提升各種加成性雜訊環境下的語音辨識率，並與其它許多強健性技術有良好的加成性。

Abstract

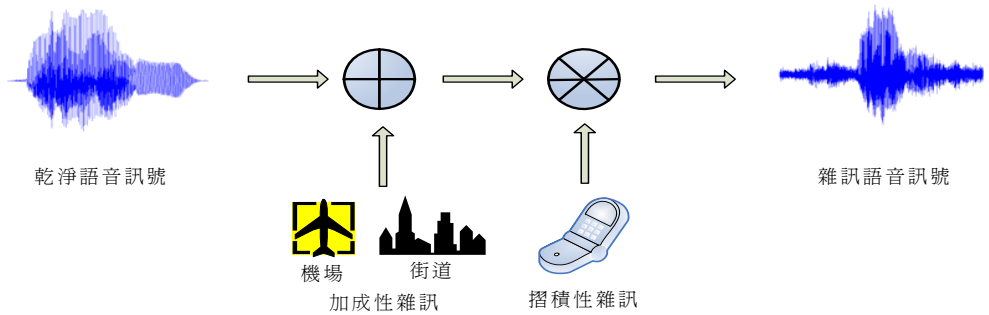
The rapid development of speech processing techniques has made themselves successfully applied in more and more applications, such as automatic dialing, voice-based information retrieval, and identity authentication. However, some unexpected variations in speech signals deteriorate the performance of a speech processing system, and thus relatively limit its application range. Among these variations, the environmental mismatch caused by the embedded noise in the speech signal is the major concern of this paper. In this paper, we provide a more rigorous mathematical analysis for the effects of the additive noise on two energy-related speech features, i.e. the logarithmic energy (log$E$) and the zeroth cepstral coefficient ($c0$). Then based on these effects, we propose a new feature compensation scheme, named silence feature normalization (SFN), in order to improve the noise robustness of the above two features for speech recognition. It is shown that, regardless of its simplicity in implementation, SFN brings about very significant improvement in noisy speech recognition, and it behaves better than many well-known feature normalization approaches. Furthermore,

SFN can be easily integrated with other noise robustness techniques to achieve an even better recognition accuracy.

關鍵詞：自動語音辨識、對數能量特徵、第零維倒頻譜特徵係數、強健性語音特徵
Keywords: speech recognition, logarithmic energy feature, the zeroth cepstral coefficient, robust speech features

一、緒論

近年來科技發展迅速，但是自動語音辨識仍然是一門相當具有挑戰性的課題。通常一自動語音辨識系統在不受外在雜訊干擾的研究室環境下，都可以獲得極高的辨識效能，但若是應用到實際的環境中，系統辨識效能則通常會大幅降低，這主要是被現實環境中許多的變異性(variation)所影響。而語音辨識的變異性種類繁多，例如訓練環境與測試環境間存在的環境不匹配(environmental mismatch)、語者變異(speaker variation)以及發音的變異(pronunciation variation)等。對於環境不匹配而言，其相關的變數可概略分爲下列幾項類型：加成性雜訊(additive noise)、摺積性雜訊(convolutional noise)以及頻寬的限制(bandwidth limitation)等。圖一爲乾淨語音訊號受到雜訊干擾之示意圖。



圖一、乾淨語音受雜訊干擾之示意圖

本論文是以上述所提及的環境不匹配中的加成性雜訊因素，作爲主要探討的主題，以期將加成性雜訊對語音辨識的影響降低。在特徵參數抽取步驟時，我們經常計算語音的能量值作爲特徵之一；根據過去的文獻指出[1][2]，語音訊號的能量特徵(energy feature)所包含的辨識資訊大過於其它特徵，且能量特徵的計算複雜度很低。所以根據上述能量特徵的優勢，在本論文中，我們特別對其強健性技術加以分析、討論與發展。

近年來，有許多成功的強健性對數能量特徵(logarithmic energy, log$E$)的技術相繼被提出，例如，對數能量動態範圍正規化法(log-energy dynamic range normalization, LEDRN)[3]其目標是使訓練與測試的語音資料其對數能量值之動態範圍一致化；對數能量尺度重刻法(log-energy rescaling normalization, LERN)[4]則是將對數能量特徵乘上一個介於 0 與 1 間的權重值，試圖重建出乾淨語音的對數能量特徵；而本實驗室先前所提出的靜音音框對數能量正規化法(silence energy normalization, SLEN)[5]，是將判別爲非語音音框(non-speech frame)的對數能量特徵設定爲一極小值的常數。上述的三種方法，皆傾向於將非語音部分的對數能量數值調低，並將語音部分的對數能量值保持不變；其主要的原因是一段語音特徵中，能量較低的部分通常會比能量較高的部分更容易受到雜訊的影響。本論文依據前人所發表的文獻加以改進，且針對語音訊號能量相關的特徵如何受到雜訊影響，以較嚴謹的數學理論加以分析，並提出一套新的強健技術，稱爲「靜音特徵正規化法」(silence feature normalization, SFN)，此方法可以有效地降低加成性雜訊對語音能量相關特徵的干擾，進而提高系統的辨識效能。

本論文其它章節概要如下：在第二章中，我們先主要將對能量相關特徵受雜訊影響的效應，做進一步的分析與探討，接著介紹本論文所新提出的之靜音特徵正規化法(SFN)；第三章包含了各種針對能量相關特徵之處理技術的語音辨識實驗數據及相關討論，其中除了介紹語音辨識實驗環境外，主要是評估靜音特徵正規化法的效能，並與其他方法作比較，藉此驗證我們所提出新方法能有效提升能量相關特徵在雜訊環境下的強健性。在第四章中，我們嘗試將所提的新方法結合其它的強健性特徵技術，對此類的結合作辨識實驗所得到的辨識率加以探討與分析，以驗證我們所提出的靜音特徵正規化法是否與其它技術有良好的加成性。第五章則爲本論文結論與未來展望。

二、靜音特徵正規化法

首先，我們在第一節中，針對語音能量相關特徵：對數能量(logarithmic energy, log$E$)與第零維倒頻譜係數($c0$)受到環境雜訊干擾的變異現象做較深入的觀察分析與探討，接著在第二節中，我們根據這些結果，提出靜音特徵正規化法的新強健性技術。

（一）對數能量特徵及第零維倒頻譜特徵係數受加成性雜訊干擾之現象的探討

加成性雜訊對於能量相關特徵(log$E$ 與 $c0$)造成的效應可由圖二看出端倪。圖二(a)、(b)與(c)分別表示一乾淨語音訊號(Aurora-2.0 資料庫中的"MAH_1390A"檔)的波形圖、對數能量(log$E$)曲線圖與第零維倒頻譜特徵係數($c0$)曲線圖；而(b)與(c)中紅色實線、綠色虛線與藍色點線則分別爲乾淨語音、訊雜比 15dB 的語音及訊雜比 5dB 的語音所對應的曲線。由這三張圖中，可以很明顯地看出，在有語音存在的區域，log$E$ 與 $c0$ 特徵值較大，較不容易受到雜訊的影響而失真，而且隨時間上下振盪的情況較爲明顯；反之，在無語音存在的區段，其特徵值前後變化較平緩，且受到雜訊的干擾後，其值會很明顯地被改變許多。接下來，我們就以較嚴謹的數學理論，對以上兩種失真現象加以分析與探討。首先，我們探討加成性雜訊對於 log$E$ 特徵的影響。假設一段受加成性雜訊干擾的語音(noisy speech)中，第 n 個音框的訊號 $x_n[m]$ 可表示爲：

$$x_n[m] = s_n[m] + d_n[m] \ , \qquad \text{式(2-1)}$$

其中 $s_n[m]$ 與 $d_n[m]$ 分別表示第 n 個音框之乾淨語音訊號(clean speech)以及雜訊(noise)，則此音框之 log$E$ 特徵值可用下式表示：

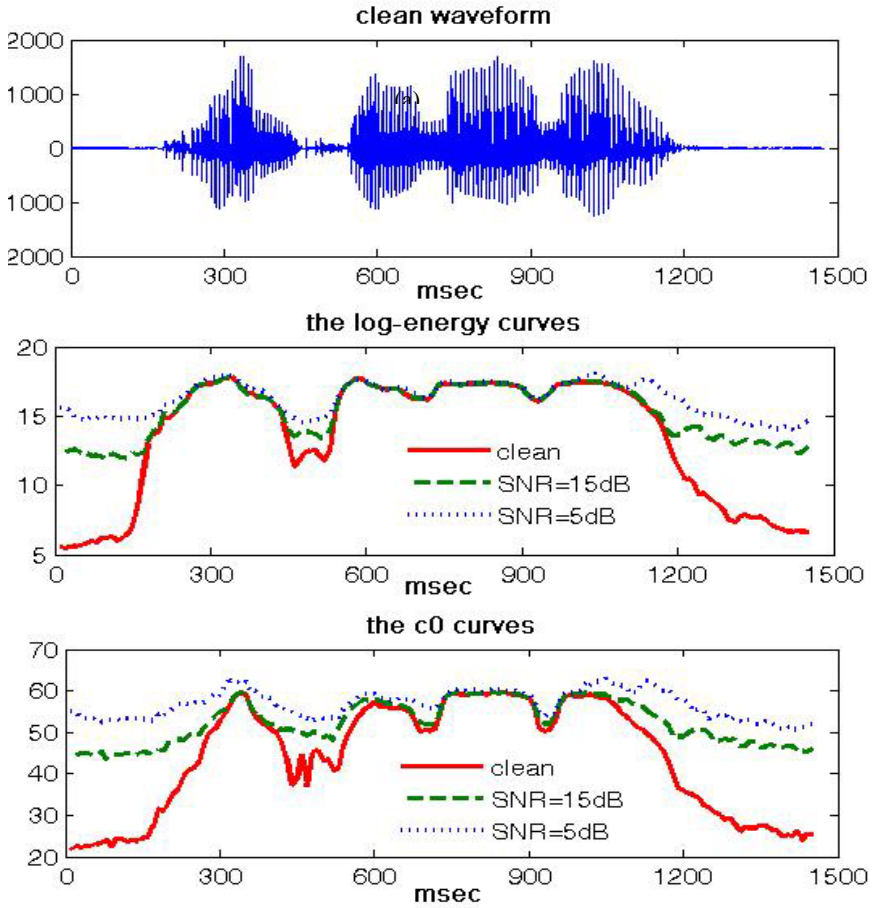
$$\begin{aligned} E^{(x)}[n] &= \log\left(\sum\nolimits_m x_n^2[m]\right) \approx \log\left(\sum\nolimits_m s_n^2[m] + \sum\nolimits_m d_n^2[m]\right) \\ &= \log\left(\exp\left(E^{(s)}[n]\right) + \exp\left(E^{(d)}[n]\right)\right) \ , \end{aligned} \qquad \text{式(2-2)}$$

其中 $E^{(x)}[n]$ 、 $E^{(s)}[n]$ 與 $E^{(d)}[n]$ 分別爲 $x_n[m]$ 、 $s_n[m]$ 以及 $d_n[m]$ 所對應之 log$E$ 特徵值。因此，受到雜訊干擾所導致雜訊語音與乾淨語音訊號兩者間 log$E$ 特徵的差異 $\Delta E[n]$ 可用下式表示：

$$\Delta E[n] = E^{(x)}[n] - E^{(s)}[n] \approx \log\left(1 + \exp\left(E^{(d)}[n] - E^{(s)}[n]\right)\right) \ 。 \qquad \text{式(2-3)}$$

由式(2-3)可觀察出，若在相同的雜訊能量($E^{(d)}[n]$)下，此差異值 $\Delta E[n]$ 與乾淨語音訊號之 $E^{(s)}[n]$ 兩者呈現負相關的關係，當 $E^{(s)}[n]$ 愈大時， $\Delta E[n]$ 愈小，反之則愈大。根據上述的推導，可以看出一雜訊語音訊號中，含有語音成份的音框($E^{(s)}[n]$ 較大)相較於純雜訊音框($E^{(s)}[n]$ 較小)而言，其 log$E$ 特徵被雜訊影響的情況較小(即失真量 $\Delta E[n]$ 較小)。

接下來，我們探討加成性雜訊對於語音訊號的 log$E$ 特徵序列於調變頻譜(modulation spectrum)上的影響。首先，我們將式(2-2)以泰勒級數(Taylor series)展開，其展開的中心點設定爲 $\left(E^{(s)}[n], E^{(d)}[n]\right) = (0,0)$ ，展開階層爲 2 階，如式(2-4)所示：



(a)

(b)

(c)

圖二、在不同 SNR 下，一語音訊號之波形圖及能量相關特徵時間序列圖，其中(a)爲乾淨語音波形、(b)爲 log$E$ 特徵曲線、(c)爲 $c0$ 特徵曲線

$$E^{(x)}[n] \approx \log\left(\exp\left(E^{(s)}[n]\right)+\exp\left(E^{(d)}[n]\right)\right)$$
$$\approx \log 2+\frac{1}{2}\left(E^{(s)}[n]+E^{(d)}[n]\right)+\frac{1}{8}\left(\left(E^{(s)}[n]\right)^2+\left(E^{(d)}[n]\right)^2-E^{(s)}[n]E^{(d)}[n]\right). \quad \text{式(2-4)}$$

因此，若將上式(2-4)取傅立葉轉換，則此雜訊語音的對數能量序列$\left\{E^{(x)}[n]\right\}$的調變頻譜可用下式表示：

$$X(j\omega) \approx (2\pi\log 2)\delta(\omega)+\frac{1}{2}\left(S(j\omega)+D(j\omega)\right)$$
$$+\frac{1}{16\pi}\left(S(j\omega)*S(j\omega)+D(j\omega)*D(j\omega)-S(j\omega)*D(j\omega)\right), \quad \text{式(2-5)}$$

式中$X\left(j\omega\right)$、$S\left(j\omega\right)$以及$D\left(j\omega\right)$分別爲雜訊語音之log$E$序列$\left\{E^{(x)}[n]\right\}$、乾淨語音之log$E$序列$\left\{E^{(s)}[n]\right\}$與雜訊之log$E$序列$\left\{E^{(d)}[n]\right\}$的調變頻譜。假設$\left\{E^{(s)}[n]\right\}$與$\left\{E^{(d)}[n]\right\}$兩序列皆爲低通(low-pass)訊號，且$B_s$與$B_d$爲其相對應之頻寬(bandwidth)，則式(2-5)中$D\left(j\omega\right)*D\left(j\omega\right)$與$S\left(j\omega\right)*D\left(j\omega\right)$兩項的頻寬分別爲$2B_d$與$B_s+B_d$；這意味著雜訊語音之log$E$序列$\left\{E^{(x)}[n]\right\}$相較於雜訊的log$E$序列$\left\{E^{(d)}[n]\right\}$將擁有更大的頻寬。換言之，對log$E$序列而言，雜訊語音比雜訊擁有較多高頻的調變頻譜成份；這便可以解釋爲何在一雜訊

語音訊號中含有語音的區段，比起純雜訊的區段看起來振盪情形(fluctuating)更爲明顯。

接著我們探討加成性雜訊對於 *c*0 特徵的影響。假設雜訊語音中第 n 個音框的 *c*0 特徵值以 $c_0^{(x)}[n]$ 做表示，而 $c_0^{(s)}[n]$ 與 $c_0^{(d)}[n]$ 分別表示此音框之所含乾淨語音訊號及純雜訊的 *c*0 特徵值，則它們可被推導如下三式：

$$c_0^{(x)}[n]=\sum\nolimits_k \log\left(M^{(x)}[k,n]\right)\approx\sum\nolimits_k \log\left(M^{(s)}[k,n]+M^{(d)}[k,n]\right)\text{，} \qquad \text{式(2-6)}$$

$$c_0^{(s)}[n]=\sum\nolimits_k \log\left(M^{(s)}[k,n]\right)\text{，} \qquad \text{式(2-7)}$$

$$c_0^{(d)}[n]=\sum\nolimits_k \log\left(M^{(d)}[k,n]\right)\text{，} \qquad \text{式(2-8)}$$

其中，$M^{(x)}[k,n]$、$M^{(s)}[k,n]$ 與 $M^{(d)}[k,n]$ 分別爲式(2-1)中雜訊語音訊號 $x_n[m]$、乾淨語音訊號 $s_n[m]$ 以及雜訊 $d_n[m]$ 於轉換成梅爾倒頻譜特徵時，第 k 個梅爾濾波器的輸出值。因此我們可推導出，由於加成性雜訊干擾所導致雜訊語音與乾淨語音訊號兩者之 *c*0 特徵值的差異 $\Delta c_0[n]$ 如下式所示：

$$\begin{aligned}\Delta c_0[n]&=c_0^{(x)}[n]-c_0^{(s)}[n]\approx\sum\nolimits_k \log\left(1+\frac{M^{(d)}[k,n]}{M^{(s)}[k,n]}\right)\\&=\sum\nolimits_k \log\left(1+\frac{1}{SNR[k,n]}\right)\text{，}\end{aligned} \qquad \text{式(2-9)}$$

式中 $SNR[k,n]$ 定義爲第 n 個音框中第 k 維梅爾頻帶的訊雜比，即

$$SNR[k,n]=\frac{M^{(s)}[k,n]}{M^{(d)}[k,n]}. \qquad \text{式(2-10)}$$

由式(2-9)可看出，若多數梅爾頻帶的訊雜比 $SNR[k,n]$ 都比較大時，差異值 $\Delta c_0[n]$ 也相對變小，因此這可約略解釋含語音之音框(*SNR* 較大)相對於純雜訊音框(*SNR* 較小)而言，*c*0 特徵值較不易受到影響的現象。

以下我們將探討加成性雜訊對於 *c*0 特徵序列之調變頻譜(modulation spectrum)上的影響。首先爲了推導起見，我們將式(2-6)、式(2-7)與式(2-8)改寫成下列三式：

$$c_0^{(x)}[n]=\sum\nolimits_k \tilde{M}^{(x)}[k,n]\approx\sum\nolimits_k \log\left(\exp\left(\tilde{M}^{(s)}[k,n]\right)+\exp\left(\tilde{M}^{(d)}[k,n]\right)\right), \qquad \text{式(2-11)}$$

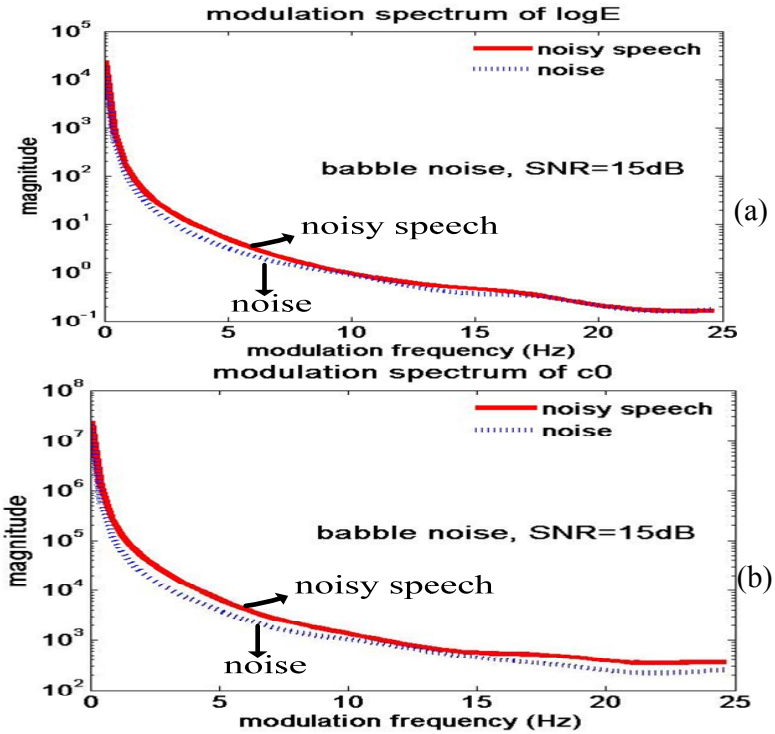
$$c_0^{(s)}[n]=\sum\nolimits_k \tilde{M}^{(s)}[k,n], \qquad \text{式(2-12)}$$

$$c_0^{(d)}[n]=\sum\nolimits_k \tilde{M}^{(d)}[k,n], \qquad \text{式(2-13)}$$

其中 $\tilde{M}^{(x)}[k,n]=\log\left(M^{(x)}[k,n]\right)$、$\tilde{M}^{(s)}[k,n]=\log\left(M^{(s)}[k,n]\right)$、$\tilde{M}^{(d)}[k,n]=\log\left(M^{(d)}[k,n]\right)$。類似將式(2-11)與式(2-2)作比較，可看出雜訊語音、乾淨語音與純雜訊三者的關係在 log*E* 與 *c*0 兩特徵中十分類似，因此藉由前面之式(2-4)與式(2-5)對於 log*E* 特徵序列之調變頻譜的推導，我們可以發現對每個梅爾濾波器輸出的對數值序列 $\left\{\tilde{M}^{(x)}[k,n]\right\}$ 而言，其頻寬仍是大於 $\left\{\tilde{M}^{(d)}[k,n]\right\}$，也就是說 $\left\{c_0^{(x)}[n]\right\}$ 比起 $\left\{c_0^{(d)}[n]\right\}$ 將擁有更大的頻寬，因此，類似 log*E* 特徵的結果，我們同樣歸納出雜訊語音之 *c*0 特徵序列比純雜訊之 *c*0 特徵序列擁有較多高頻的調變頻譜成份，亦即前者比後者有更明顯的上下振盪現象。

圖三(a)與圖三(b)分別爲一句語音訊號之 log*E* 特徵及 *c*0 特徵的功率頻譜密度(power spectral density, PSD)曲線圖，其中的語音訊號及雜訊爲 Aurora-2.0 資料庫中的 "FAC_5Z31ZZ4A"檔與人聲雜訊(babble noise)，訊雜比爲 15dB。由這兩圖我們可以很明

顯地看出，雜訊語音相對於純雜訊而言，其 log$E$ 特徵序列與 $c0$ 特徵序列都有較大的頻寬，此亦驗證了我們之前的推導。



圖三、能量相關特徵之功率頻譜密度圖，(a)爲 log$E$ 特徵、(b)爲 $c0$ 特徵

綜合上述的推導及圖例，我們驗證了一段雜訊語音中含有語音的音框其 log$E$ 特徵與 $c0$ 特徵相對於純雜訊音框而言，失真程度較小，且擁有較大的頻寬，亦即具有較明顯的上下振盪現象。基於上述觀察，我們將提出新的強健性語音特徵處理技術─靜音特徵正規化法(silence feature normalization, SFN)，其具有兩種模式，分述於之後的兩節中。

（二）靜音特徵正規化法 I (silence feature normalization I, SFN-I)

在本節中，我們介紹第一種模式的靜音特徵正規化法，稱之爲「靜音特徵正規化法 I」 (silence feature normalization I, SFN-I)；此方法是針對原靜音音框對數能量正規化法 (SLEN) [5]加以改良，目的是希望對 log$E$ 與 $c0$ 之能量相關特徵做處理，使一段訊號中非語音(non-speech)部份的特徵值做正規化，而含有語音之區域的特徵值則保持不變，以達到重建出乾淨語音訊號之能量相關特徵的效果。

首先，我們假設 $\{x[n]\}$ 爲一段雜訊語音訊號之 log$E$ 特徵或 $c0$ 特徵之序列；根據我們於上一小節所得到的結論，雜訊語音中含有語音的區段相較於純雜訊區段，其 log$E$ 與 $c0$ 特徵序列將擁有更高的調變頻譜成份；因此我們設計一高通無限脈衝響應濾波器 (high-pass infinite impulse response filter)來處理此段序列，其轉換函數如下：

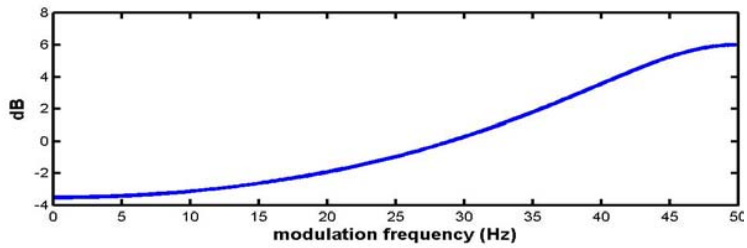
$$H(z)=\frac{1}{1+\alpha z^{-1}} \qquad 0<\alpha<1\text{。} \qquad \text{式(2-14)}$$

而此濾波器之輸入輸出關係式如下所示：

$$y[n]=-\alpha y[n-1]+x[n]\text{，} \qquad \text{式(2-15)}$$

式中 $y[n]$ 爲濾波器的輸出，且我們將其初始值設定爲 $y[0]=0$ 。式(2-14)之濾波器其強度響應(magnitude response)如圖四所示，由圖中可以發現，此濾波器能夠有效地降低特徵序列中接近直流(near-DC)的成份，並將較高頻率的部份加以強調，此較高頻率的成份

可突顯出語音與純雜訊的差異。因此經過此濾波器轉換後所得到的$\{y[n]\}$將比原始參數$\{x[n]\}$擁有更好的效能來判斷語音與非語音區段。



圖四、式(2-14)之高通濾波器的強度響應($\alpha = 0.5$)

根據式(2-15)所得之$y[n]$，我們可作一段訊號中語音與非語音音框的判別，並進而將其非語音的音框做正規化處理，此即爲靜音特徵正規化法 I (silence feature normalization I, SFN-I)，其式如下：

SFN-I: $\tilde{x}[n] = \begin{cases} x[n] & \text{if } y[n] > \theta \\ \log(\varepsilon) + \delta & \text{if } y[n] \le \theta \end{cases}$， 式(2-16)

其中θ、ε與δ分別爲門檻值、一極小的正數以及一平均值爲 0 且變異數很小的隨機變數，$\tilde{x}[n]$爲經過 SFN-I 處理後所得到的新特徵參數。其門檻值θ計算式如下：

$$\theta = \frac{1}{N}\sum_{n=1}^{N} y[n]\ ,$$ 式(2-17)

式中N爲此段語音的音框總數。因此，門檻值即爲整段語音所有$y[n]$的平均值，其計算十分簡便，且無需額外特別設計之處。

從式(2-16)看出，若$y[n]$大於門檻值θ，則將其所對應之音框判斷爲語音，且原特徵參數保持不變；反之則將其歸類爲非語音音框，並將原特徵參數正規化成一極小的隨機變數；相較於之前靜音音框對數能量正規化法(SLEN)[5]而言，靜音特徵正規化法 I 可避免將非語音部份的特徵正規化爲一定值，而可能導致之後所訓練的聲學模型中的變異數(variance)變爲 0 的錯誤現象產生。我們可以透過圖五來觀察 SFN-I 法的作用。圖五中，(a)與(b)分別爲原始的 log$E$ 特徵序列以及 $c0$ 特徵序列曲線；(c)與(d)分別爲經過靜音特徵正規化法 I 處理後所得到之 log$E$ 特徵序列以及 $c0$ 特徵序列曲線，其中紅色實線是對應至乾淨語音(Aurora-2.0 資料庫中的"FAK_3Z82A"檔)、綠色虛線與藍色點線則分別爲對應至訊雜比 15dB 與 5dB 的雜訊語音。由這些圖明顯地看出，SFN-I 法處理過後之能量相關特徵值可以較趨近於原始乾淨語音訊號之特徵值，達到降低失真的目的。

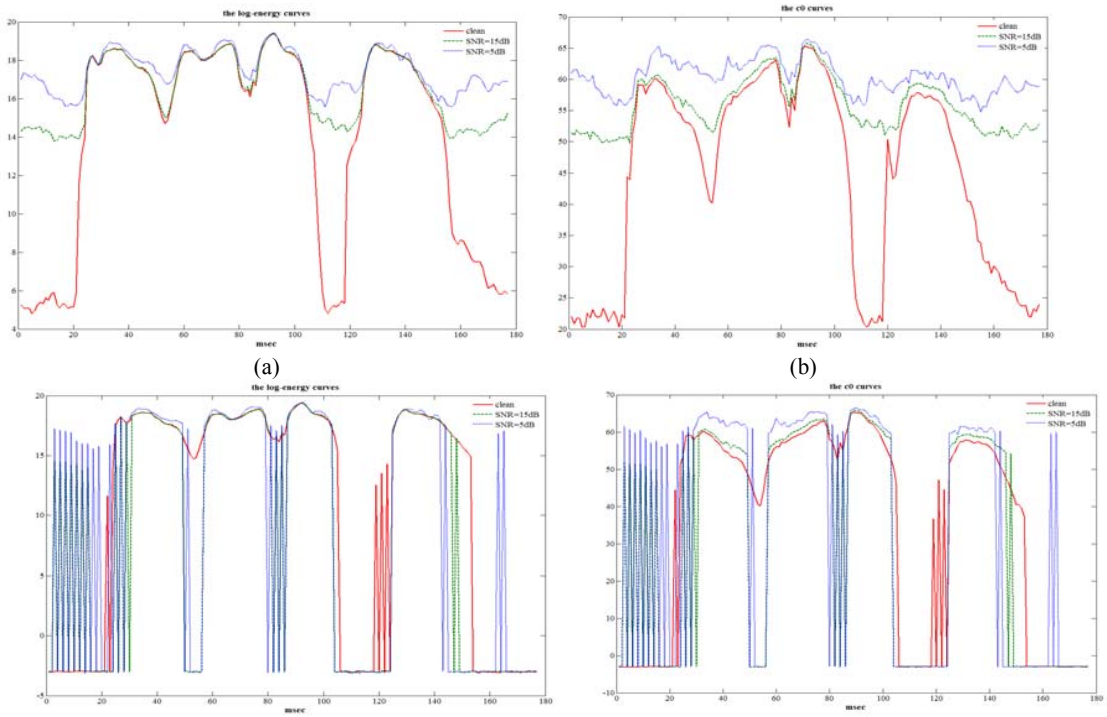
（三）靜音特徵正規化法 II (silence feature normalization II, SFN-II)

在本節中，我們將介紹第二種模式的靜音特徵正規化法，稱之爲「靜音特徵正規化法 II」(silence feature normalization II, SFN-II)，SFN-II 法與前一節之 SFN-I 法最大的差異在於，SFN-II 是將原能量相關特徵$\{x[n]\}$乘上一權重值(weight)，而得到新特徵值$\{\tilde{x}[n]\}$。SFN-II 的演算法如下式所示：

SFN-II: $\tilde{x}[n] = w[n]x[n]$, 式(2-18)

其中，

$$w[n] = \begin{cases} 1/\left(1+\exp\left(-(y[n]-\theta)/\beta\sigma_1\right)\right) & \text{if } y[n] > \theta \\ 1/\left(1+\exp\left(-(y[n]-\theta)/\beta\sigma_2\right)\right) & \text{if } y[n] \le \theta \end{cases}\ ,$$ 式(2-19)



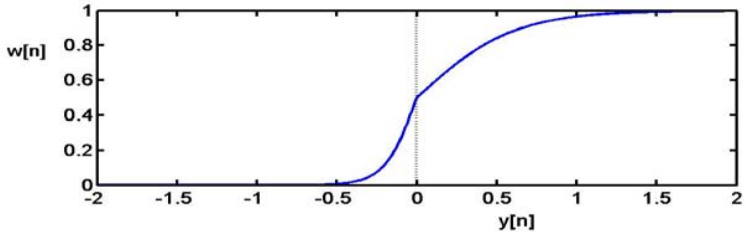
圖五、靜音特徵正規化法 I 處理前((a)與(b))與處理後((c)與(d))能量相關特徵序列曲線圖，其中(a)與(c)爲 log$E$ 特徵序列曲線，(b)與(d)爲 $c0$ 特徵序列曲線

其中$y[n]$如前一節之式(2-15)所示，爲$\{x[n]\}$通過一高通濾波器之輸出值，θ爲門檻值、σ_1與σ_2分別爲$\{y[n]|y[n]>\theta\}$(大於門檻值θ之所有的$y[n]$)以及$\{y[n]|y[n]\le\theta\}$(小於或等於門檻值θ之所有的$y[n]$)所對應之標準差、β爲一常數。SFN-II 之門檻值θ跟 SFN-I 相同，計算式如下所示：

$$\theta=\left(1/N\right)\sum_{n=1}^{N}y[n], \qquad \text{式(2-20)}$$

式中N爲此段語音中音框總數。

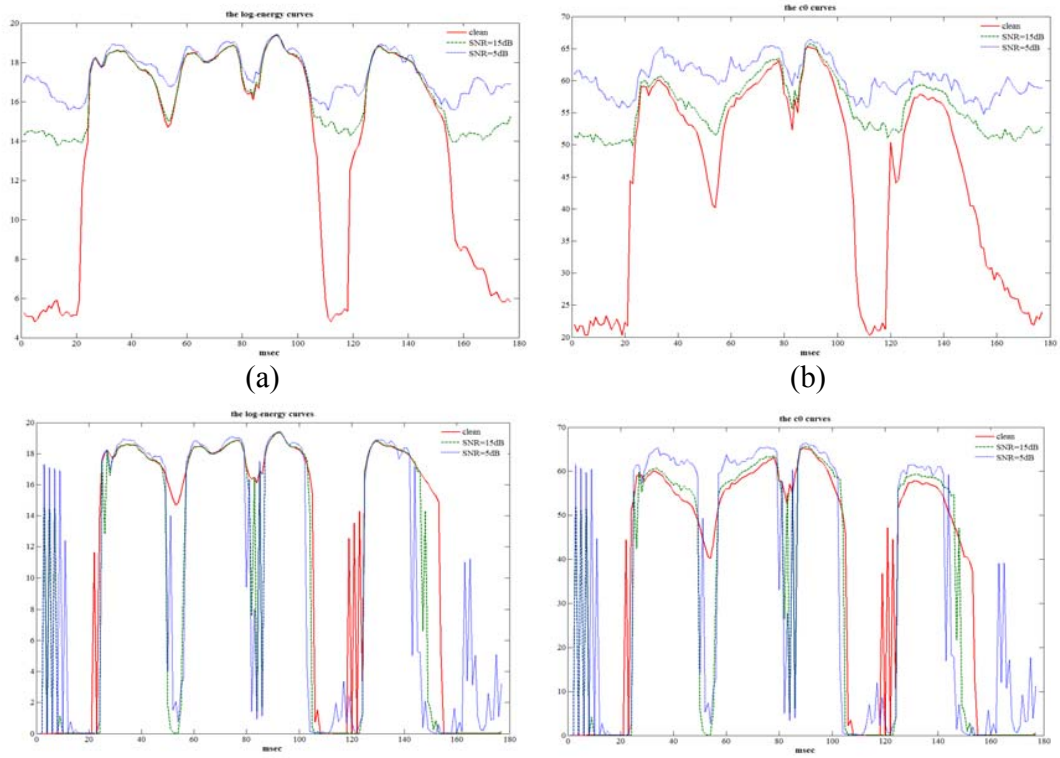
式(2-19)的權重值$w[n]$如圖六所示，其中假設$\theta=0$、$\sigma_1=1$、$\sigma_2=3$以及$\beta=0.1$。由圖六可以發現，權重值函數$w[n]$爲一個左右不對稱之遞增的 S 形曲線(sigmoid curve)，其值介於 0 和 1 之間。此權重值所代表的意義與 SFN-I 法相似，我們希望新得到的能量相關特徵$\tilde{x}[n]$能在原始特徵值很大時，儘量維持不變；而原始值較小時，則使其變得更小。SFN-II 法和 SFN-I 法不同之處在於，SFN-II 法具有"軟式"的語音端點偵測決策(soft-decision VAD)，而 SFN-I 法則爲"硬式"的語音端點偵測決策(hard-decision VAD)；因此 SFN-II 法相較於 SFN-I 法而言，其 VAD 判定錯誤的影響可能相對來得比較小，效能也會比較好，這推想將會在之後的章節驗證。



圖六、權重值函數$w[n]$曲線示意圖

圖七爲 SFN-II 法處理前與處理後能量相關特徵之曲線圖。與之前的圖三類似，(a)與(b)

分別爲原始的 log$E$ 特徵序列以及 c0 特徵序列曲線；(c)與(d)分別爲經過靜音特徵正規化法 II 處理後所得到之 log$E$ 序列以及 c0 序列曲線，其中紅色實線是對應至乾淨語音(Aurora-2.0 資料庫中的"FAK_3Z82A"檔)、綠色虛線與藍色點線則分別爲對應至訊雜比 15dB 與 5dB 的雜訊語音。很明顯地，經由 SFN-II 處理過後之雜訊語音的能量相關特徵，皆類似 SFN-I 法的效果，可以更趨近於原始乾淨語音之特徵，有效降低雜訊造成的失真。



(a) (b)

(c) (d)

圖七、靜音特徵正規化法 II 處理前((a)與(b))與處理後((c)與(d))能量相關特徵序列曲線圖，其中(a)與(c)爲 log$E$ 特徵序列曲線，(b)與(d)爲 c0 特徵序列曲線

三、能量相關特徵處理技術之實驗結果與討論

（一）、語音資料庫簡介

本論文中的語音辨識實驗所使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)發行的 Aurora-2.0 語料庫[7]。它是一套藉由以人工的方式錄製的連續英文數字字串，語者爲美國成年男女，加上八種加成性雜訊，分別爲地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，以及不同程度的訊雜比，分別爲 20dB、15dB、10dB、5dB、0dB 以及-5dB，附加上乾淨(clean)語料。

（二）、特徵參數的設定與辨識系統的訓練

本論文根據 Aurora-2.0 實驗語料庫標準設定[7]，語音特徵參數主要是使用梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)及能量相關特徵，附加上其一階差量與二階差量。爲了分析能量相關特徵的影響，於本論文中採用兩組不同的特徵參數；第一組是 12 維梅爾倒頻譜特徵值(c1～c12)加上 1 維的對數能量(log$E$)，另一組則是使用 12 維梅爾倒頻譜特徵值(c1～c12)加上第零維倒頻譜特徵係數(c0)；而每組皆會再加上一階與二階差量，故兩組皆用了 39 維的特徵參數。詳細的特徵參數設定，如表一所示。

我們利用 HTK 程式[8]來訓練聲學模型，產生了 11(oh, zero, one~nine)個數字模型以

及靜音模型，每個數字模型包含 16 個狀態(states)，而每個狀態是由 20 個高斯密度函數混合(Gaussian mixtures)所組成。

表一、本論文中所使用之語音特徵參數設定

取樣頻率	8kHz	
音框長度(Frame Size)	25ms, 200 點	
音框平移(frame Shift)	10ms, 80 點	
預強調濾波器	$1-0.97z^{-1}$	
視窗形式	漢明窗(Hamming window)	
傅立葉轉換點數	256 點	
濾波器組(filters)	梅爾刻度三角濾波器組， 共 23 個三角濾波器	
特徵向量 (feature vector)	第一組： $\{c_i \mid 1 \le i \le 12\}$, $\{\Delta c_i \mid 1 \le i \le 12\}$, $\{\Delta^2 c_i \mid 1 \le i \le 12\}$, $\log E, \Delta \log E, \Delta^2 \log E$ 共計 39 維	第二組： $\{c_i \mid 1 \le i \le 12\}$, $\{\Delta c_i \mid 1 \le i \le 12\}$, $\{\Delta^2 c_i \mid 1 \le i \le 12\}$, $c0, \Delta c0, \Delta^2 c0$ 共計 39 維

（三）語音辨識實驗結果

在這一節中，我們將執行各種針對能量相關特徵之強健性技術的語音辨識，並比較其效能。除了我們所新提出的靜音特徵正規化法（SFN-I與SFN-II）外，我們同時實驗了平均與變異數正規化法(mean and variance normalization, MVN)[9]、平均與變異數正規化附加ARMA濾波器法(MVN plus ARMA filtering, MVA)[10]、統計圖等化法(histogram equalization, HEQ)[11]、對數能量動態範圍正規化法(log-energy dynamic range normalization, LEDRN)[3]、對數能量尺度重刻法 (log-energy rescaling normalization, LERN)[4]與靜音對數能量正規化法(silence log-energy normalization, SLEN)[5]，值得注意的是，原始之MVN、MVA與HEQ三方法雖是設計於所有種類的特徵上，我們爲了評估其在能量相關特徵的效能，在這裡只將它們運用於log*E*與*c*0特徵的正規化上，另外，LEDRN法有分線性與非線性兩種，在這裡我們分別以LEDRN-I與LEDRN-II表示，而LERN亦有兩種版本，我們分別以LERN-I與LERN-II表示。

1、針對對數能量特徵(log*E*)之強健式語音技術綜合分析

此小節之實驗所用到語音特徵爲前述之第一組的特徵參數，即 12 維梅爾倒頻譜特徵值(*c*1～*c*12)加上 1 維的對數能量(log*E*)，附加其一階與二階差量，共 39 維。而這裡所用到的十種特徵強健性方法，皆是單純處理 log*E* 特徵，不考慮其它 12 維的梅爾倒頻譜係數，表二列出了基礎實驗及這十種方法所得之平均辨識率（20dB、15dB、10dB、5dB 與 0dB 五種訊雜比下的辨識率平均），其中 AR 與 RR 分別爲相較基礎結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。從表二的數據，我們可觀察到下列幾點現象：

①原始作用於所有種類特徵之 MVN、MVA 與 HEQ 法單純作用於 log*E* 特徵時，其提供的改進效果也十分明顯，相較於基礎實驗結果，分別具有 10.18%、11.70%與 14.97%的辨識率提升。相對於 MVN 而言，由於 MVA 多使用了一個 ARMA 低通濾波器以強調語音的成分，而 HEQ 額外對語音特徵的高階動差(higher-order moments)作正規化，所以兩者效果皆比 MVN 還來得好。

②以往文獻所提出之針對 log$E$ 特徵作補償的各種方法：LEDRN-I、LEDRN-II、LERN-I、LERN-II 與 SLEN，都能帶來十分顯著的辨識率提升，其中線性 LEDRN(LEDRN-I)明顯優於非線性 LEDRN(LEDRN-II)，其平均辨識率相差了大約 4%，兩種版本的 LERN(LERN-I 與 LERN-II)，效果則十分接近，且表現優於 LEDRN。而本實驗室過去所提出的 SLEN 法，相對於基礎實驗的平均辨識率而言，有 15.19%的提升，明顯優於之前所提之 LEDRN 與 LERN 等方法。

③ 本論文所提出的兩種靜音特徵正規化法，SFN-I 與 SFN-II，相對於基礎實驗結果而言，平均辨識率分別提升了 15.38%與 16.11%，相對錯誤降低率都在 50%以上，相較於之前所提的各種方法，SFN-I 與 SFN-II 都有更優異的表現，此驗證了我們所提的兩個新方法，都能有效地提昇 log$E$ 特徵在加成性雜訊環境下的強健性，且優於目前許多著名的 log$E$ 特徵正規化技術。此外，我們也發現，SFN-II 所得之辨識率比 SFN-I 更好，此可能原因如之前所述，由於 SFN-II 在語音偵測(voice activity detection)的決策機制與 SFN-I 並不相同，語音偵測之錯誤在 SFN-II 中相對影響較小，而使其相對表現較佳。

表二、針對 log$E$ 特徵之強健式語音技術之辨識率的綜合比較表(%)

	Method	Set A	Set B	average	AR	RR
(1)	Baseline	71.98	67.79	69.89	—	—
(2)	MVN	79.04	81.08	80.06	10.18	33.79
(3)	MVA	80.53	82.64	81.59	11.70	38.85
(4)	HEQ	83.91	85.79	84.85	14.97	49.69
(5)	LEDRN-I	82.01	79.70	80.86	10.97	36.43
(6)	LEDRN-II	77.21	75.53	76.37	6.49	21.53
(7)	LERN-I	83.64	83.35	83.50	13.61	45.19
(8)	LERN-II	82.71	81.94	82.33	12.44	41.31
(9)	SLEN	84.87	85.27	85.07	15.19	50.42
(10)	**SFN-I**	**85.02**	**85.50**	**85.26**	**15.38**	**51.05**
(11)	**SFN-II**	**85.67**	**86.32**	**86.00**	**16.11**	**53.49**

2、針對第零維倒頻譜特徵係數($c0$)之強健式語音技術綜合分析

此小節之實驗所用到語音特徵為前述之第二組的特徵參數，即 12 維梅爾倒頻譜特徵值($c1$～$c12$)加上第零維倒頻譜特徵係數($c0$)，附加其一階與二階差量，共 39 維。類似前一小節，我們將原始針對 log$E$ 特徵的十種特徵強健性方法，作用於 $c0$ 特徵上，其它 12 維的梅爾倒頻譜係數則維持不變。雖然目前處理的是 $c0$ 特徵，但為了簡明起見，這裡我們不將原本各種技術的名稱作修改，例如 LEDRN 法，我們並不特別將其改名為 $c0$-DRN 法，而仍沿襲其名，其他方法名稱依此類推。

表三列出了基礎實驗及這十種方法所得之平均辨識率（20dB、15dB、10dB、5dB 與 0dB 五種訊雜比下的辨識率平均），而其中的 AR 與 RR 分別為相較於基礎實驗結果之絕對錯誤降低率和相對錯誤降低率。從表三的數據，我們可觀察到下列幾點現象：

①類似之前的表二之結果，各種方法作用於 $c0$ 特徵時，都能帶來提昇辨識率的效果，其中，LEDRN-I 與 LEDRN-II 的表現比其他方法稍差，尤其是 LEDRN-II，只有 3.57% 之絕對錯誤降低率(AR)，其可能原因在於，LEDRN 原本是針對 log$E$ 特徵所設計，若我們直接將其套用於 $c0$ 特徵處理上，其所使用的參數並非是最佳化而得，導致效果不彰。

②三種原本作用於所有種類特徵之方法：MVN、MVA 與 HEQ 法，單純作用於 $c0$ 特徵

時，仍然以 HEQ 表現最好，MVA 法次之，MVN 法較差，但彼此表現的差距並未如之前在表二來得明顯。此外，LERN-I、LERN-II 與 SLEN 都有十分顯著的改進效果，唯與表二的數據不同之處，在於三種方法的效能十分接近，而 LERN-I 略優於 SLEN。

③本論文所提出的兩種靜音特徵正規化法，SFN-I 與 SFN-II，相對於基礎實驗結果而言，平均辨識率分別提升了 13.79%與 14.13%，相對錯誤降低率約爲 46%，類似表二的結果，SFN-II 仍然優於 SFN-I，且這兩種方法之表現仍優於其他所有的方法。此結果驗證了我們所提的兩個新方法，能有效地提昇 *c*0 特徵在加成性雜訊環境下的強健性。

表三、針對 *c*0 特徵之強健式語音技術之辨識率的綜合比較表(%)

	Method	Set A	Set B	Average	AR	RR
(1)	Baseline	71.95	68.22	70.09	—	—
(2)	MVN	80.80	82.95	81.88	11.79	39.41
(3)	MVA	81.76	84.04	82.90	12.82	42.84
(4)	HEQ	82.89	84.59	83.74	13.66	45.65
(5)	LEDRN-I	79.04	77.36	78.20	8.11	27.13
(6)	LEDRN-II	74.08	73.22	73.65	3.57	11.92
(7)	LERN-I	83.81	83.65	83.73	13.65	45.61
(8)	LERN-II	83.03	82.53	82.78	12.70	42.44
(9)	SLEN	82.94	84.28	83.61	13.53	45.21
(10)	**SFN-I**	**83.04**	**84.70**	**83.87**	**13.79**	**46.08**
(11)	**SFN-II**	**83.29**	**85.14**	**84.22**	**14.13**	**47.23**

雖然 SFN 法有效地降低雜訊對 *c*0 造成的失真，進而提昇辨識率，但當我們比較表二與表三時，發現無論是 SFN-I 或 SFN-II，作用於 log*E* 特徵可得到的辨識率會高於作用在 *c*0 特徵所得之辨識率；由此，我們推斷由 log*E* 特徵所得之 SFN-I 法與 SFN-II 法其中的語音端點偵測(VAD)結果，可能會比由 *c*0 所得結果來的好。根據此推想，我們將原來針對 *c*0 特徵的兩種 SFN 法稍作修改。於 SFN-I 中，我們先利用 log*E* 對音框做語音/非語音的分類，再將此判別結果套用於 *c*0 上，對非語音音框的 *c*0 做如式(2-16)之正規化處理；而 SFN-II 也是利用相同的方式，先利用 log*E* 對音框做語音/非語音的分類，再將其結果轉換至 *c*0 上，並對語音與非語音音框的 *c*0 特徵序列求取其式(2-19)所用的標準差 σ_1 與 σ_2，然後作式(2-18)之正規化處理。我們將以上的修正作法分別稱作針對 *c*0 特徵之修正式 SFN-I 法(modified SFN-I)與修正式 SFN-II 法(modified SFN-II)。

針對 *c*0 特徵之修正式 SFN-I 法與修正式 SFN-II 法，其所得之平均辨識率如表四所示，如我們所預期的，修正式 SFN 法相對於原始 SFN 法，能有更進一步的改進效果，對 SFN-I 而言，前者相較於後者額外提昇了 1.29%的平均辨識率，而對 SFN-II 而言，前者相較於後者額外提昇了的 1.33%平均辨識率。此結果部分驗證了我們的推想，即利用 log*E* 特徵來執行語音端點偵測(VAD)，其效果會比 *c*0 特徵來的好。

表四、針對 *c*0 特徵之原始 SFN 法與修正式 SFN 法之辨識率比較表(%)

Method	Set A	Set B	Average	AR	RR
Baseline	71.95	68.22	70.09	—	—
SFN-I	83.04	84.70	83.87	13.79	46.08
modified SFN-I	**84.54**	**85.79**	**85.17**	**15.08**	**50.41**
SFN-II	83.29	85.14	84.22	14.13	47.23
modified SFN-II	**85.03**	**86.06**	**85.55**	**15.46**	**51.68**

四、靜音特徵正規化法與其它特徵強健法結合之實驗結果與討論

前一章之一系列的實驗，主要是探討各種能量相關特徵處理技術效能，進而突顯出我們所新提出之靜音特徵正規化(SFN)法的優異表現，這些實驗中，只有 logE 與 $c0$ 兩種能量相關特徵被處理，剩餘的梅爾倒頻譜特徵係數($c1$~$c12$)則維持不變。在這一章中，我們嘗試將作用於 log$E$ 與 $c0$ 特徵的 SFN 法與作用於 $c1$~$c12$ 之梅爾倒頻譜特徵係數的強健性技術加以結合，藉以觀察兩者之間是否有加成性，能進一步改進語音辨識率。

在這裡，我們選擇之前所提之 MVN[9]、MVA[10]以及 HEQ[11]三種強健性技術，分別作用於 $c1$~$c12$ 之梅爾倒頻譜特徵係數上，而將我們所提之 SFN-I 或 SFN-II 法作用於能量相關特徵(log$E$ 或 $c0$)上，我們將其上述所有的實驗結果分別彙整成表五與表六。

針對第一組特徵(log$E$, $c1$~$c12$)處理之表五的數據中，列(2)~(4)是利用單一強健技術(MVN, MVA 或 HEQ)處理全部特徵參數之結果，而列(5)~(10)則分別爲靜音特徵正規化法(SFN)結合其它方法之結果。當我們將列(2)、列(5)與列(8)的結果相比較、列(3)、列(6)與列(9)的結果相比較，及列(4)、列(7)與列(10)的結果相比較，都可以看出將 SFN-I 或 SFN-II 使用於 log$E$ 特徵，並用其他方法使用在 $c1$～$c12$ 特徵上，所得到的辨識率比單獨使用一種方法處理全部特徵的辨識結果高出許多，例如列(9)之『SFN-II (log$E$) + MVA ($c1$~$c12$)』法，其平均辨識率高達 89.97%，超越了列(4)之『HEQ (log$E$, $c1$~$c12$)』法所得之 87.44%的平均辨識率。同時，我們也看出 SFN-II 的效能普遍優於 SFN-I，此結果跟前一章的結論是一致的。而當我們將表五與表二的數據相比較時，也可以看出，使用 SFN 處理 log$E$ 特徵結合使用 MVN、MVA 或 HEQ 法額外處理 $c1$～$c12$ 特徵，可以比單獨使用 SFN 處理 log$E$ 特徵得到更佳的辨識效果，此結果驗證了 SFN 法與 MVN、MVA 或 HEQ 法的確具有加成性。

表五、SFN 法作用在 log$E$ 特徵結合其它語音強健技術作用於 $c1$～$c12$ 特徵參數之平均辨識率的綜合比較表(%)

	Method	Set A	Set B	average	AR	RR
(1)	Baseline	71.98	67.79	69.89	—	—
(2)	MVN (log$E$, $c1$~$c12$)	83.55	83.75	83.65	13.77	45.71
(3)	MVA (log$E$, $c1$~$c12$)	86.69	86.89	86.79	16.91	56.13
(4)	HEQ (log$E$, $c1$~$c12$)	87.15	87.72	87.44	17.55	58.28
(5)	SFN-I (log$E$) + MVN ($c1$~$c12$)	87.33	87.81	87.57	17.69	58.72
(6)	SFN-I (log$E$) + MVA ($c1$~$c12$)	88.40	88.84	88.62	18.74	62.21
(7)	SFN-I (log$E$) + HEQ ($c1$~$c12$)	87.93	88.04	87.99	18.10	60.10
(8)	SFN-II (log$E$) + MVN ($c1$~$c12$)	88.45	88.88	88.67	18.78	62.36
(9)	SFN-II (log$E$) + MVA ($c1$~$c12$)	89.82	90.12	89.97	20.09	66.69
(10)	SFN-II (log$E$) + HEQ ($c1$~$c12$)	89.29	89.33	89.31	19.43	64.50

針對第二組特徵($c0$, $c1$~$c12$)處理之表六的數據中，列(2)~(4)是利用單一強健技術(MVN, MVA 或 HEQ)處理全部特徵參數之結果，而列(5)~(16)則分別爲靜音特徵正規化法(SFN)結合其它方法之結果。類似表五中列(1)~(10)所呈現的結果，從表六中之列(1)~(10)與表三的數據相較，使用 SFN 處理 $c0$ 特徵結合使用 MVN、MVA 或 HEQ 法額外處理 $c1$～$c12$ 特徵，可以比單獨使用 SFN 處理 $c0$ 特徵得到更佳的效能，然而我們發現，將 SFN-I 或 SFN-II 使用於 $c0$ 特徵，並用其他方法使用在 $c1$～$c12$ 特徵時，所得到的辨識率並非總是優於單獨使用一種方法處理全部特徵的辨識結果(這些較差的數據在表中以*號加以註記)，例如列(6)之『SFN-I ($c0$) + MVA ($c1$~$c12$)』法，其平均辨識率爲

87.77%，相較於列(3)之『MVA (*c*0, *c*1~*c*12)』法所得之 88.46%來得差。此現象的可能原因，在前一章已經提到，即利用 *c*0 特徵執行 SFN 法中的語音端點偵測(VAD)會比較不精確，進而降低 SFN 的效能。因此，類似前一章，在這裡我們使用針對 *c*0 特徵之修正式的 SFN 法，來與 MVN、MVA 或 HEQ 法作結合，這些結果列於表六的列(11)~(16)中。

表六、SFN 法作用在 *c*0 特徵結合其它語音強健技術作用於 *c*1～*c*12 特徵參數之平均辨識率綜合比較表(%)

	Method	Set A	Set B	Average	AR	RR
(1)	Baseline	71.95	68.22	70.09	—	—
(2)	MVN (*c*0, *c*1~*c*12)	85.03	85.54	85.29	15.20	50.81
(3)	MVA (*c*0, *c*1~*c*12)	88.11	88.81	88.46	18.38	61.42
(4)	HEQ (*c*0, *c*1~*c*12)	86.99	88.13	87.56	17.48	58.42
(5)	SFN-I (*c*0) + MVN (*c*1~*c*12)	85.62	86.62	86.12	16.04	53.60
(6)	SFN-I (*c*0) + MVA (*c*1~*c*12)	87.38*	88.16*	87.77*	17.69	59.12
(7)	SFN-I (*c*0) + HEQ (*c*1~*c*12)	85.95*	86.53*	86.24*	16.16	54.00
(8)	SFN-II (*c*0) + MVN (*c*1~*c*12)	86.92	87.69	87.31	17.22	57.56
(9)	SFN-II (*c*0) + MVA (*c*1~*c*12)	89.04	89.61	89.33	19.24	64.32
(10)	SFN-II (*c*0) + HEQ (*c*1~*c*12)	87.43	87.88*	87.66	17.57	58.73
(11)	modified SFN-I (*c*0) + MVN (*c*1~*c*12)	87.49	87.89	87.69	17.61	58.85
(12)	modified SFN-I (*c*0) + MVA (*c*1~*c*12)	89.30	89.54	89.42	19.34	64.63
(13)	modified SFN-I (*c*0) + HEQ (*c*1~*c*12)	88.10	88.39	88.25	18.16	60.71
(14)	modified SFN-II (*c*0) + MVN (*c*1~*c*12)	88.25	88.33	88.29	18.21	60.86
(15)	modified SFN-II (*c*0) + MVA (*c*1~*c*12)	89.87	89.98	89.93	19.84	66.32
(16)	modified SFN-II (*c*0) + HEQ (*c*1~*c*12)	89.25	89.46	89.36	19.27	64.42

將表六之列(11)~(16)的數據與列(1)~(10)相比較，我們可以明顯看出針對 *c*0 特徵之修正式 SFN 法(modified SFN-I 與 modified SFN-II)，比原始 SFN 法的效能高出許多，且與 MVN、MVA 或 HEQ 一併使用後，其結果必然優於 MVN、MVA 或 HEQ 處理所有特徵的結果，其中以列(15)之『modified SFN-II (*c*0) + MVA (*c*1~*c*12)』法所得到的平均辨識率最高，爲 89.93%，與之前表五中最佳辨識率 89.97%（列(9)的『SFN-II (log*E*) + MVA (*c*1~*c*12)』法）十分接近，此結果明顯驗證了修正式 SFN 法確實更進一步改進了 *c*0 特徵在加成性雜訊環境下的強健性。

由第三章與第四章之全部的實驗數據中，我們可以充分驗證所提出的兩種靜音特徵正規化法(SFN-I 與 SFN-II)對於能量相關特徵具有良好的強健化效果，而 SFN-II 所得到的辨識率總是比 SFN-I 高，其可能原因如第二章所陳述，因爲 SFN-II 法具有"軟式"決策之語音端點偵測(soft-decision voice activity detection)的機制，相較於 SFN-I 法"硬式"決策之語音端點偵測(hard-decision voice activity detection)的機制，前者的語音/非語音判別錯誤所造成的影響相對較小。然而，總括而言，SFN-I 法 SFN-II 法的共同優點在於執行上十分簡易（即複雜度極低）且效果很優異，因此極具實用的價值。

五、結論

在本論文中，我們提出一個新的語音強健技術 —「靜音特徵正規化法」(silence feature normalization, SFN)，此方法執行上十分簡易且效果優異。它是針對能量相關特

徵(log$E$ 與 c0)因加成性雜訊造成的失真現象作適當的補償。SFN 法利用了一個高通濾波器去處理原始能量相關特徵序列，並將通過此高通濾波器所的之輸出特徵序列拿來作語音/非語音的分類，並應用簡單且有效的方法來處理非語音部份的特徵，將雜訊對語音特徵的干擾降低，以期提升訓練與測試環境匹配度，進而提升雜訊環境下的語音辨識率。

由實驗數據中可發現，就處理能量相關特徵而言，SFN 法比基本實驗以及許多強健式語音技術得到更好的辨識率；由此可知針對能量相關特徵做適當的補償，在穩定以及非穩定雜訊環境下皆得到十分顯著的辨識率提升，顯示了能量相關特徵所含的語音鑑別資訊是影響辨識率的一個重要指標。此外，當我們將 SFN 法與其它強健式語音技術做結合，發現其辨識率比單獨使用一種強健式語音技術所得到的辨識率更高，其中又以 SFN-II 法結合 MVA 法得到的辨識率最高，可達到將近 90%的平均辨識率。

能量相關特徵雖然具高度語音鑑別力，但是雜訊對其干擾程度也相對很大，因此能量相關特徵處理的好壞，將會很直接地影響到系統的辨識效能，由此可知能量相關特徵的強健化處理在未來仍是值得探討的一大課題；我們希望未來可以將所發展的技術，擴展測試至其它較大字彙量的語音辨識系統上，探討這類技術在不同複雜度之語音辨識系統的效能。另外，未來我們仍可朝向消除加成性雜訊的方向繼續深入研究，也可以針對消除通道性雜訊的方法去作相關的探討，並嘗試將兩者結合，使得語音辨認系統能更有效地降低各類雜訊的干擾，而擁有令人滿意之辨識率。

參考文獻

[1] Bocchieri , E. L., and Wilpon, J. G., "Discriminative Analysis for Feature Reduction in Automatic Speech Recognition", 1992 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992).

[2] Julien Epps and Eric H.C. Choi, "An Energy Search Approach to Variable Frame Rate Front-End Processing for Robust ASR", 2005 European Conference on Speech Communication and Technology (Interspeech 2005—Eurospeech).

[3] Weizhong Zhu and Douglas O'Shaughnessy, "Log-Energy Dynamic Range Normalization for Robust Speech Recognition", 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005).

[4] Hung-Bin Chen, "On the Study of Energy-Based Speech Feature Normalization and Application to Voice Activity Detection", M.S. thesis, National Taiwan Normal University, Taiwan, 2007.

[5] C-F. Tai and J-W. Hung, "Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments", 2006 International Conference on Spoken Language Processing (Interspeech 2006—ICSLP).

[6] Tai-Hwei Hwang and Sen-Chia Chang, "Energy Contour Enhancement for Noisy Speech Recognition", 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP 2004).

[7] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," Proceedings of ISCA IIWR ASR2000, Paris, France, 2000

[8] http://htk.eng.cam.ac.uk/

[9] S. Tiberewala and H. Hermansky, "Multiband and Adaptation Approaches to Robust Speech Recognition", 1997 European Conference on Speech Communication and Technology (Eurospeech 1997)

[10] C-P. Chen and J-A. Bilmes, "MVA Processing of Speech Features", IEEE Trans. on Audio, Speech, and Language Processing, 2006

[11] A. Torre, J. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-Linear Transformations of the Feature Space for Robust Speech Recognition", 2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)

Robust Features for Effective Speech and Music Discrimination

Zhong-hua Fu[1], Jhing-Fa Wang[2]

School of Computer Science
Northwestern Polytechnical University, Xi'an, China[1]
Department of Electrical Engineering
National Cheng Kung University, Tainan, Taiwan[1, 2]
mailfzh@mail.ncku.tw[1], wangjf@csie.ncku.edu.tw[2]

Abstract

Speech and music discrimination is one of the most important issues for multimedia information retrieval and efficient coding. While many features have been proposed, seldom of which show robustness under noisy condition, especially in telecommunication applications. In this paper two novel features based on real cepstrum are presented to represent essential differences between music and speech: Average Pitch Density (APD), Relative Tonal Power Density (RTPD). Separate histograms are used to prove the robustness of the novel features. Results of discrimination experiments show that these features are more robust than the commonly used features. The evaluation database consists of a reference collection and a set of telephone speech and music recorded in real world.

Keywords: Speech/Music Discrimination, Multimedia Information Retrieval, Real Cepstrum.

1. Introduction

In applications of multimedia information retrieval and effective coding for telecommunication, audio stream always needs to be diarized or labeled as speech, music or noise or silence, so that different segments can be implemented in different ways. However, speech signals often consist of many kinds of noise, and the styles of music such as personalized ring-back tone may differ in thousands ways. Those make the discrimination problem more difficult.

A variety of systems for audio segmentation or classification have been proposed in the past and many features such as Root Mean Square (RMS) [1], Zero Crossing Rate (ZCR) [1,4,5], low frequency modulation [2,4,5], entropy and dynamism features [2,3,6], Mel Frequency Cepstral coefficients (MFCCs) have been used. Some features need high quality audio signal or refined spectrum detail, and some cause long delay so as not fit for telecommunication applications. While the classification frameworks including nearest neighbor, neural network, Hidden Markov Model (HMM), Gaussian Mixture Modal (GMM) and Support Vector Machine (SVM) have been adopted as the back end, features are still the crucial factor to the final performance. As shown in the following part of this paper, the discrimination abilities of some common features are poor with noisy speech. The main reason may explain as that they do not represent the essential difference between speech and music.

In this paper, two novel features, called as Average Pitch Density (APD) and Relative Tonal

Power Density (RTPD) are proposed, which are based on real cepstrum analysis and show better robustness than the others. The evaluation database consists of two different data sets: one comes from Scheirer and Slaney [5], the other is collected from real telecommunication situation. The total lengths for music and speech are about 37 minutes and 28.7 minutes respectively.

The rest of this paper is organized as follows: Section 2 introduces the novel features based on real cepstrum analysis. Section 3 describes the evaluation database and the comparative histograms of different features. The discrimination experiments and their results are given in section 4. Section 5 concludes this paper.

2. Features Based on Real Cepstrum

There are tremendous types of music, and the signal components of which can be divided into two classes: tonal-like and noise-like. The tonal-like class consists of tones played by all kinds of musical instruments, and these tones are catenated to construct the melody of music. The noise-like class is mainly played by percussion instruments such as drum, cymbal, gong, maracas, etc. The former class corresponds to the musical system, which construct by a set of predefined pitches according to phonology. The latter class can not play notes with certain pitch and is often used to construct rhythm.

The biggest difference between speech and music lies on the pitch. Because of the restriction of musical system, the pitch of music usually can only jump between discrete frequencies, except for vibratos or glissandi. But pitch of speech can change continuously and will not keep on a fixed frequency for a long time. Besides the difference of pitch character, the noise part of music, which is often played by percussion instrument, also has different features from speech. That part of music does not have pitch, but it usually has stronger power. This phenomenon seldom exists in speech signal, because generally the stronger part of speech is voiced signal, which does have pitch.

In order to describe the differences of pitch between speech and music, we use real cepstrum instead of spectrogram. Cepstrum analysis is a more powerful tool to analysis the detail of spectrum, which can separate pitch information from spectral envelop. The real cepstrum is defined as (Eq. (2) gives the Matlab expression)

$$RC_x \hat{=} real\left(\frac{1}{2}\int_{-\pi}^{\pi}\log\left|X\left(e^{j\omega}\right)\right|e^{j\omega n}d\omega\right) \tag{1}$$

$$RC_x = real\left(ifft\left(\log\left(abs\left(fft\left(x\right)\right)\right)\right)\right) \tag{2}$$

Where x is a frame of audio signal weighted by hamming window, of which the discrete Fourier transform is $X\left(e^{j\omega}\right)$. $real(\cdot)$ denotes extracting the real part of the complex results. RC_x are the coefficients of real cepstrum. The coefficients that near zero origin reflect the big scale information of power spectrum such as the spectrum envelop, and those far from the zero origin show the spectrum detail. Figure 1 uses the latter to demonstrate the differences of pitch between speech and music. It is clear that the music pitches are jumped discretely while speech pitches do not. Figure 2 uses spectrogram to show the noise-like feature of a rock music segment, where most ictus have no pitch.

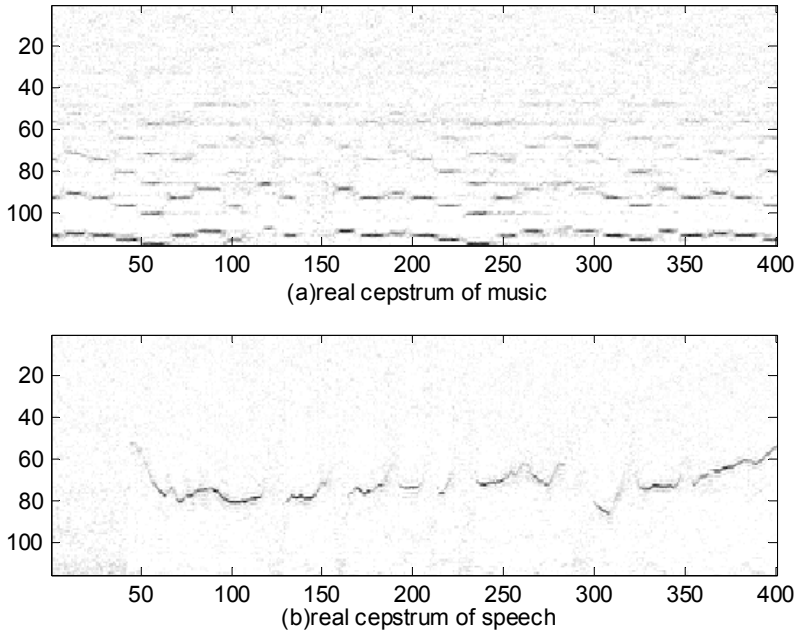


Figure 1. Pitch different between music (a) and speech (b) by means of real cepstrum. Only coefficients far from the zero origin are used.

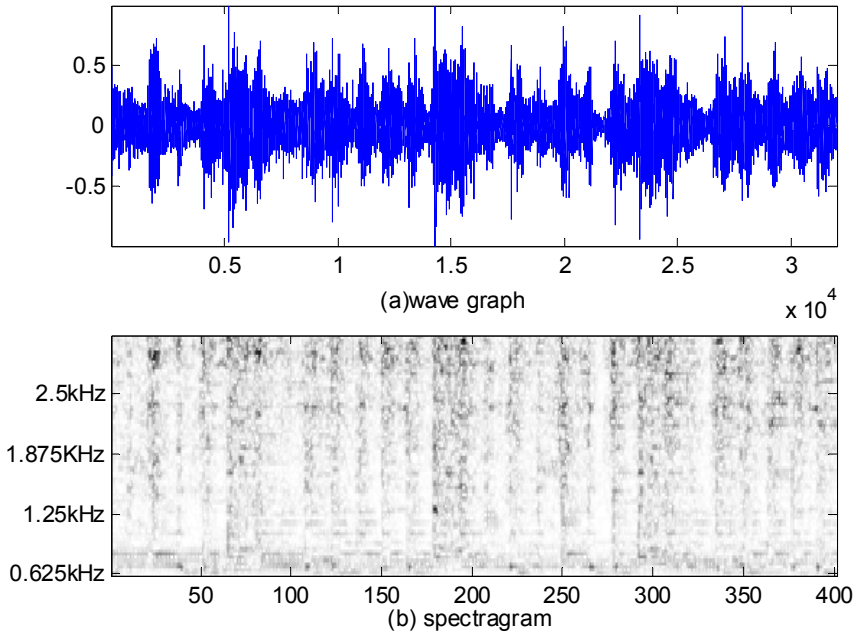


Figure 2. Waveform and spectrogram of a segment of rock music. It is clear to find that most ictus have no pitch.

To parameterize the above conclusion, we propose two novel features: Average Pitch Density (APD) and Relative Tonal Power Density (RTPD).

A. APD feature

Because of the musical instruments and polyphony, the average pitch usually is higher than speech. The APD feature is independent with signal power and reflects the details about spectrum, which is defined as

$$APD(K) = \sum_{i=K*N+1}^{K*N+N} \frac{1}{L} \sum_{j=l_1}^{l_2} \left| RCx_i(j) \right| , \text{ where } \quad L = l_2 - l_1 + 1 \tag{3}$$

where K means the K-th analysis segment, and N is the length of it. L is number of RC_x coefficients that far from zero origin, whose range is l_1 to l_2. This feature is relative simple, but it does prove to be robust for discrimination between speech and music. The histogram in figure 3 (e) demonstrate this conclusion.

B. RTPD feature

While the detail information about spectrum can be used to discriminate tonal or song from speech, the variation of energy combined with pitch information may be used to separate percussive music from noisy speech. In clean or noisy speech signal, the segments that show clear pitch usually are voiced speech, which are likely to have bigger energy. So if all segments with pitch are labeled as tonal parts and the others are label as non-tonal parts, we can probably say that if the energy of tonal parts is smaller than that of non-tonal parts, then the segment may not be speech, otherwise the segment can be speech or music.

In order to label tonal and non-tonal parts, we still use real cepstrum. Since if clear pitch does exist, a distinct stripe will appear in real cepstrum, even if in noise condition. We use the peak value of RCx that far from zero origin to judge tonal or non-tonal. The threshold we choose is 0.2. Frames whose peak value is bigger than 0.2 are labeled as tonal, or else are labeled as non-tonal. Thus the RTPD can be defined as

$$RTPD(K) = \underset{i \in \Theta}{mean}\left(RMS_i\right) \Big/ \underset{j \in \Psi}{mean}(RMS_j) \tag{4}$$

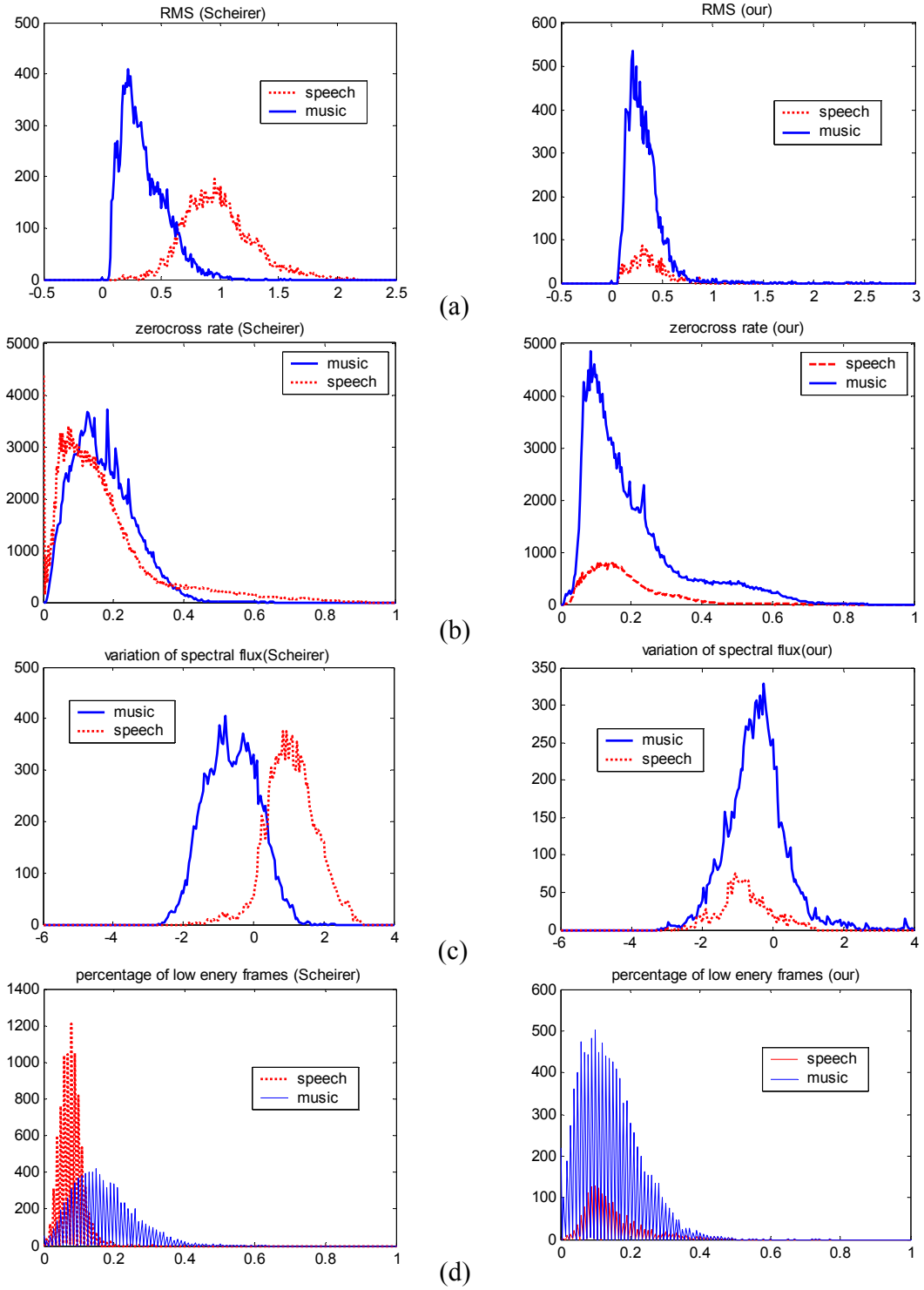
where Θ consists of all tonal frames of K-th analysis segment, and Ψ is the entire set of frames of the segment. RMS_i is the root mean square of the i-th frame.

3. Discrimination Ability

Due to the lack of a standard database for evaluation, the comparisons between different features are not easily. Our evaluation database consists of two parts: one comes from collection of Scheirer and Slaney[5], the other comes from the real records from telecommunication application. The former includes speech sets and music sets. Each set contains 80 15-second long audio samples. The samples were collected by digitally sampling an FM tuner (16-bit monophonic samples at a 22.05 kHz sampling rate), using a variety of stations, content styles, and noise levels. They made a strong attempt to collect as much of the breadth of available input signals as possible (See [5] for details). The latter set is recorded by us based on telecommunication application, which has 25 music files and 174 noisy speech files, 17 and 11.7 minutes in length respectively. Especially, the speech signals of the latter set consist of many kinds of live noises, which are non-stationary with different SNR.

Based on the two data sets above, we build an evaluation corpus by concatenating those files

randomly into two columns: CLN-Mix and ZX-Mix. CLN-Mix contains 20 mixed files, each concatenates 2 speech samples and 2 music samples which are all extracted from Scheirer's database. ZX-Mix uses the same way except that all samples are chosen from our records. With these databases, we compared 4 commonly used features with our prompted ones. They are (1) RMS; (2)zero crossing rate; (3)variation of spectral flux; (4)percentage of "low-energy" frames. Figure 3 shows the discrimination abilities of each feature with Scheirer's and our database. It is clear that those 4 features show poor performance in noise situation, while APD and RTPD show more robust



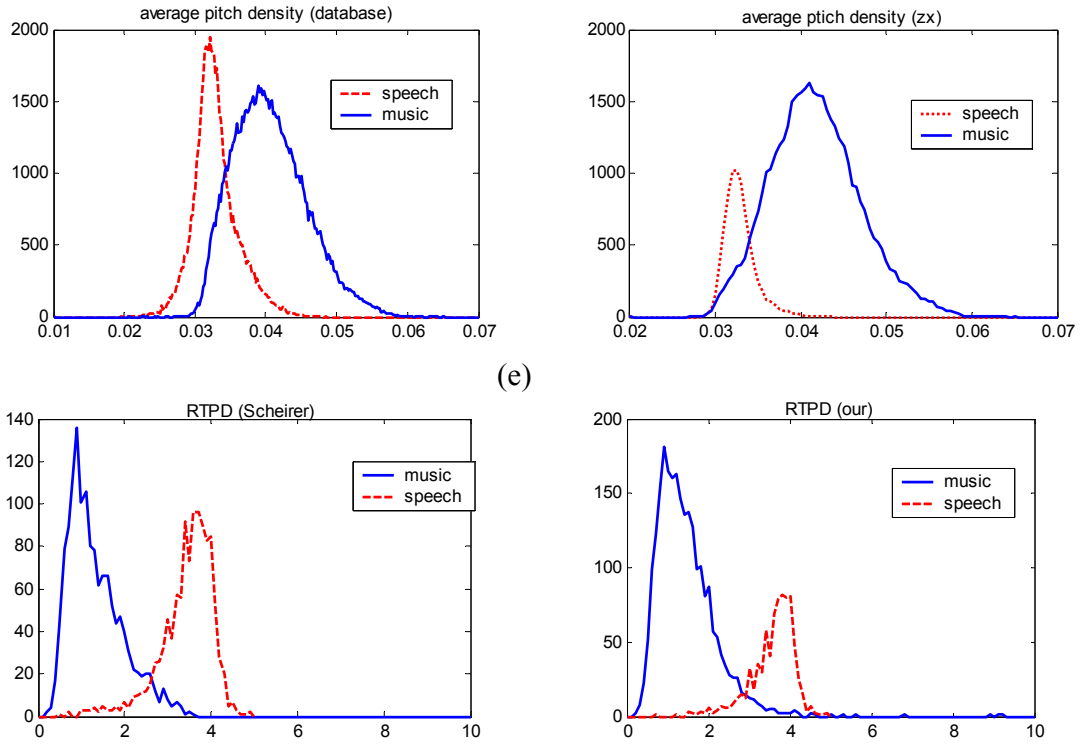


Figure 3. Histograms of different features for speech/music discrimination. (a)-(f) are RMS, ZCR, variation of spectral flux, percentage of "low-energy" frames, APD, RTPD.

4. Discrimination Experiments

In many speech and music discrimination system, GMM is commonly used for classification. A GMM models each class of data as the union of several Gaussian clusters in the feature space. This clustering can be iteratively derived with the well-known EM algorithm. Usually the individual clusters are not represented with full covariance matrices, but only the diagonal approximation. GMM uses a likelihood estimate for each model, which measurers how well the new data point is modeled by the entrained Gaussian clusters.

We use 64 components GMM to modal speech and music signal separately. The feature vector consists of: (1) APD; (2) RTPD; (3) log of variance of RMS; (4) log of variance of spectral centroid; (5) log of variance of spectral flux; (6) 4Hz modulation energy; (7) dynamic range. Training data consists of the training part of Scheirer's database and 8 minutes of noisy speech recorded. CLN-Mix and ZX-Mix are used for evaluation.

The frame length is 10ms, and the analysis windows for proposed features extraction is 1 second (100 frames) with 10 new input frames each time. For comparison, MFCC + delta + acceleration (MFCC_D_A) feature for each frame is also examined. GMM with 64 mixtures is used for speech and music respectively. For classification, every proposed feature vector is used to calculate the log likelihood score, and correspondingly, 10 frames MFCC_D_A features are used. The experimental results are list in Table 1. Furthermore, we also use the adjacent 10 proposed feature vectors for one decision and 100 frames of MFCC_D_A features are used as well. The results are shown in Table 2.

It is clear that MFCC _D_A features have good ability for discrimination with CLN-Mix data, but drop distinctly with ZX-mix, especially for music signals. But on both data sets, our

proposed features work well and express robustness in noise condition.

Table 1. Speech/Music Discrimination Accuracies in Every 100ms

Accuracy	MFCC_D_A		Proposed	
	Speech	Music	Speech	Music
CLN-Mix	91.56%	89.81%	93.78%	91.48%
ZX-Mix	99.91%	64.41%	94.19%	93.13%

Table 2. Speech/Music Discrimination Accuracies in Every Second

Accuracy	MFCC_D_A		Proposed	
	Speech	Music	Speech	Music
CLN-Mix	93.98%	95.11%	95%	92.86%
ZX-Mix	100%	67.39%	100%	94.45%

5. Conclusion

Two novel features have been presented in this paper for robust discrimination between speech and music, named Average Pitch Density (APD) and Relative Tonal Power Density (RTPD). As shown in separate histograms, many other commonly used features do not work in noisy condition, but the novels show more robustness. When combined with the other 5 robust features, the accuracies of discrimination are higher than 90%. The results mean that the novel features may represent some essential differences between speech and music.

There are many interesting directions in which to continue pursuing this work. Since the real cepstrum can show many differences between speech and music, there will be other novel features which represent the holding and changing characters of pitches. What's more, more researches are needed for better classification and feature combinations.

References

[1] C. Panagiotakis, G. Tziritas, *A Speech/Music Discriminator Based on RMS and Zero-Crossings*, IEEE Transactions on Multimedia, Vol.7(1), February 2005.

[2] O. M. Mubarak, E. A. Ambikairajah, J. Epps, *Novel Features for Effective Speech and Music Discrimination*, Proc. IEEE International Conference on Engineering of Intelligent Systems, pp.1-5, April 2006.

[3] J. E. Muñoz-Expósito , S. García-Galán , N. Ruiz-Reyes , P. Vera-Candeas, *Adaptive Network-based Fuzzy Inference System vs. Other Classification Algorithms for Warped LPC-based Speech/Music Discrimination*, Engineering Applications of Artificial Intelligence, Vol. 20(6), pp.783-793, September, 2007.

[4] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, *A Comparison of Features for Speech, Music Discrimination*, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp. 149-152, March 1999.

[5] E. Scheirer, M. Slaney, *Construction and Evaluation of a Robust Multifeature Speech /Music Discriminator*, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp. 1331-1334, April 1997.

[6] T. Zhang, J. Kuo, *Audio Content Analysis for On-line Audiovisual Data Segmentation and Classification*, IEEE Transactions on Speech Audio Processing, Vol. 9 (3), pp. 441-457, May 2001.

Robust Voice Activity Detection Based on Discrete Wavelet Transform

Kun-Ching Wang

Department of Information Technology & Communication
Shin Chien University
kunching@mail.kh.usc.edu.tw

Abstract

This paper mainly addresses the problem of determining voice activity in presence of noise, especially in a dynamically varying background noise. The proposed voice activity detection algorithm is based on structure of three-layer wavelet decomposition. Appling auto-correlation function into each subband exploits the fact that intensity of periodicity is more significant in sub-band domain than that in full-band domain. In addition, Teager energy operator (TEO) is used to eliminate the noise components from the wavelet coefficients on each subband. Experimental results show that the proposed wavelet-based algorithm is prior to others and can work in a dynamically varying background noise.

Keywords: voice activity detection, auto-correlation function, wavelet transform, Teager energy operator

1. Introduction

Voice activity detection (VAD) refers to the ability of distinguishing speech from noise and is an integral part of a variety of speech communication systems, such as speech coding, speech recognition, hand-free telephony, and echo cancellation. Although the existed VAD algorithms performed reliably, their feature parameters are almost depended on the energy level and sensitive to noisy environments [1-4]. So far, a wavelet-based VAD is rather less discussed although wavelet analysis is much suitable for speech property. S.H. Chen et al. [5] shown that the proposed VAD is based on wavelet transform and has an excellent performance. In fact, their approach is not suitable for practical application such as variable-level of noise conditions. Besides, a great computing time is needed for accomplishing wavelet reconstruction to decide whether is speech-active or not.

Compared with Chen's VAD approach, the proposed decision of VAD only depends on three-layer wavelet decomposition. This approach does not need any computing time to waste the wavelet reconstruction. In addition, the four non-uniform subbands are generated from the wavelet-based approach and the well-known "auto-correlaction function (ACF)" is adopted to detect the periodicity of subband. We refer the ACF defined in subband domain as subband auto-correlation function (SACF). Due to that periodic property is mainly focused on low frequency bands, so we let the low frequency bands have high resolution to enhance the periodic property by decomposing only low band on each layer. In addition to the SACF, enclosed herein the Teager energy operator (TEO) is regarded as a pre-processor for SACF. The TEO is a powerful nonlinear operator and has been successfully used in various speech processing applications [6-7]. F. Jabloun et al. [8] displayed that TEO can suppress the car engine noise and be easily implemented through time domain in Mel-scale subband. The later experimental result will prove that the TEO can further enhance the detection of subband periodicity.

To accurately count the intensity of periodicity from the envelope of the SACF, the Mean-Delta (MD) method [9] is utilized on each subband. The MD-based feature parameter has been presented for the robust development of VAD, but is not performed well in the non-stationary noise shown in the followings. Eventually, summing up the four values of MDSACF (Mean-Delta of Subband Auto-Correlation Function, a new feature parameter called "speech activity envelope (SAE)" is further proposed. Experimental results show that the envelope of the new SAE parameter can point out the boundary of speech activity under the poor SNR conditions and it is also insensitive to variable-level of noise.

This paper is organized as follows. Section 2 describes the concept of discrete wavelet transform (DWT) and shows the used structure of three-layer wavelet decomposition. Section 3 introductions the derivation of Teager energy operator (TEO) and displays the efficiency of subband noise suppression. Section 4 describes the proposed feature parameter, and the block diagram of proposed wavelet-based VAD algorithm is outlined in Section 5. Section 6 evaluates the performance of the algorithm and compare to other two wavelet-based VAD algorithm and ITU-T G.729B VAD. Finally, Section 7 discusses the conclusions of experimental results.

2. Wavelet transform

The wavelet transform (WT) is based on a time-frequency signal analysis. The wavelet analysis represents a windowing technique with variable-sized regions. It allows the use of long time intervals where we want more precise low-frequency information, and shorter regions where we want high-frequency information. It is well known that speech signals contain many transient components and non-stationary property. Making use of the multi-resolution analysis (MRA) property of the WT, better time-resolution is needed a high frequency range to detect the rapid changing transient component of the signal, while better frequency resolution is needed at low frequency range to track the slowly time-varying formants more precisely [10]. Figure 1 displays the structure of three-layer wavelet decomposition utilized in this paper. We decompose an entire signal into four non-uniform subbands including three detailed scales such as D1, D2 and D3 and one appropriated scale such A3.

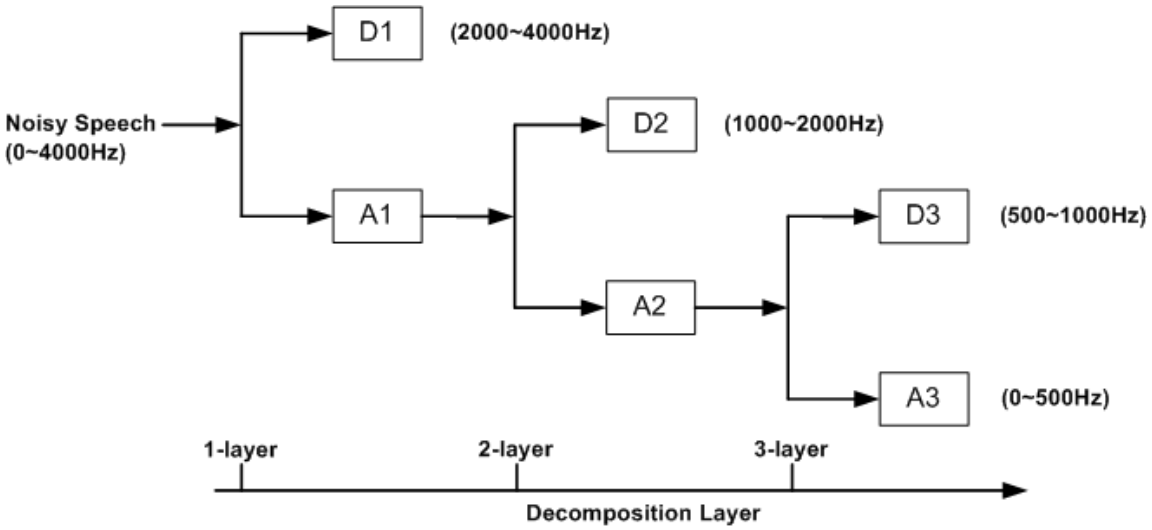


Figure 1. Structure of three-layer wavelet decomposition

3. Mean-delta method for subband auto-correlation function

The well-known definition of the term "Auto-Correlation Function (ACF)" is usually used for measuring the self-periodic intensity of signal sequences shown as below:

$$R(k) = \sum_{n=0}^{p-k} s(n)s(n+k), \quad k = 0,1,.....p\,, \tag{1}$$

where p is the length of ACF. k denotes as the shift of sample.

In order to increase the efficiency of ACF about making use of periodicity detection to detect speech, the ACF is defined in subband domain, which called "subband auto-correlation function (SACF)". Figure 2 clearly illustrates the normalized SACFs for each subband when input speech is contaminated by white noise. In addition, a normalization factor is applied to the computation of SACF. This major reason is to provide an offset for insensitivity on variable energy level. From this figure, it is observed that the SACF of voiced speech has more obviously peaks than that of unvoiced speech and white noise. Similarly, for unvoiced speech the ACF has greater periodic intensity than white noise especially in the approximation $A3$.

Furthermore, a Mean-Delta (MD) method [9] over the envelope of each SACF is utilized herein to evaluate the corresponding intensity of periodicity on each subband. First, a measure which similar to delta cepstrum evaluation is mimicked to estimate the periodic intensity of SACF, namely "Delta Subband Auto-Correlation Function (DSACF)", shown below:

$$\dot{R}_M(k) = \frac{\sum_{m=-M}^{M} m\left(\frac{R(k+m)}{R(0)}\right)}{\sum_{m=-M}^{M} m^2}, \tag{2}$$

where $\dot{R}_M$ is DSACF over an M-sample neighborhood ($M = 3$ in this study).

It is observed that the DSACF measure is almost like the local variation over the SACF. Second, averaging the delta of SACF over a M-sample neighborhood $\dot{R}_M$, a mean of the absolute values of the DSACF (MDSACF) is given by

$$\overline{R}_M = \frac{1}{N}\sum_{k=0}^{N-1}\left|\dot{R}_M(k)\right|. \tag{3}$$

Observing the above formulations, the Mean-Delta method can be used to value the number and amplitude of peak-to-valley from the envelope of SACF. So, we just only sum up the four values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale, a robust feature parameter called "speech activity envelope (SAE)" is further proposed.

Figure 3 displays that the MRA property is important to the development of SAE feature parameter. The proposed SAE feature parameter is respectively developed with/without band-decomposition. In Figure 3(b), the SAE without band-decomposition only provides obscure periodicity and confuses the word boundaries. Figure 3(c)~Figure 3(f) respectively show each value of MDSACF from D1 subband to A3 subband. It implies that the value of MDSACF can provide the corresponding periodic intensity for each subband. Summing up the four values of MDSACFs, we can form a robust SAE parameter. In Figure 3(g), the SAE with band-decomposition can point out the word boundaries accurately from its envelope.

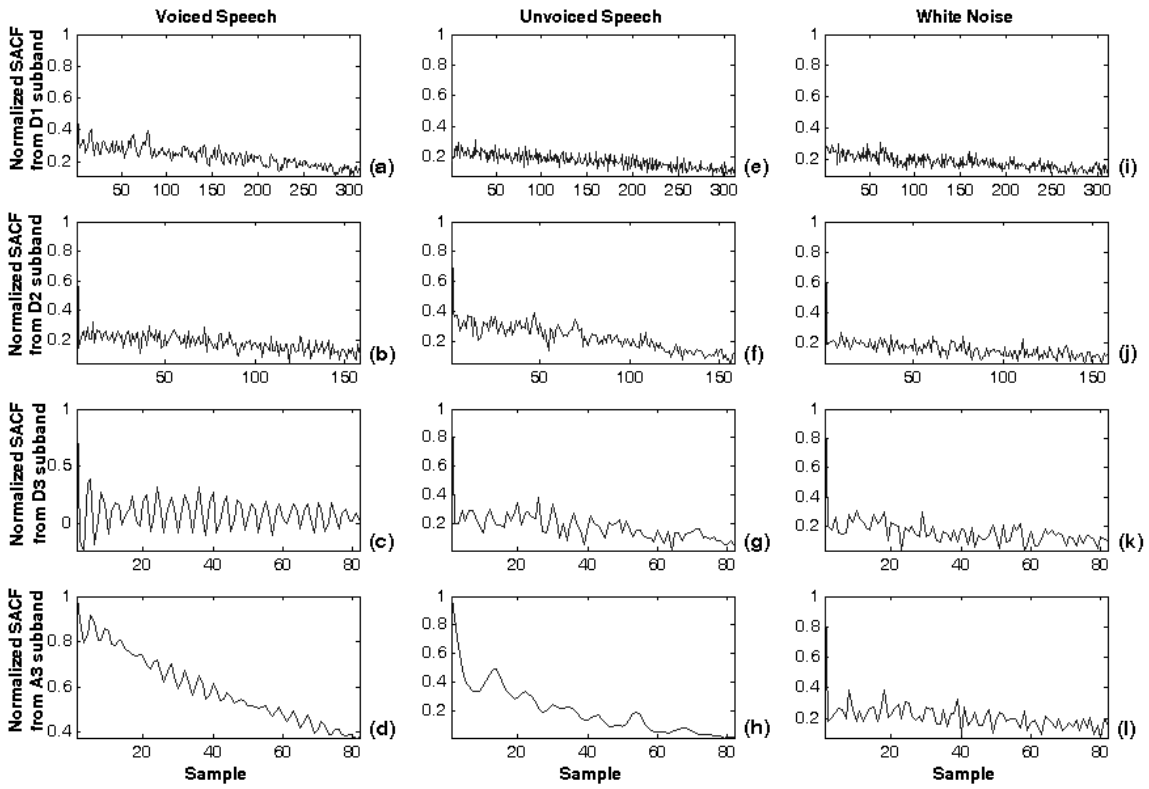


Figure 2. SACF on voiced, unvoiced signals and white noise

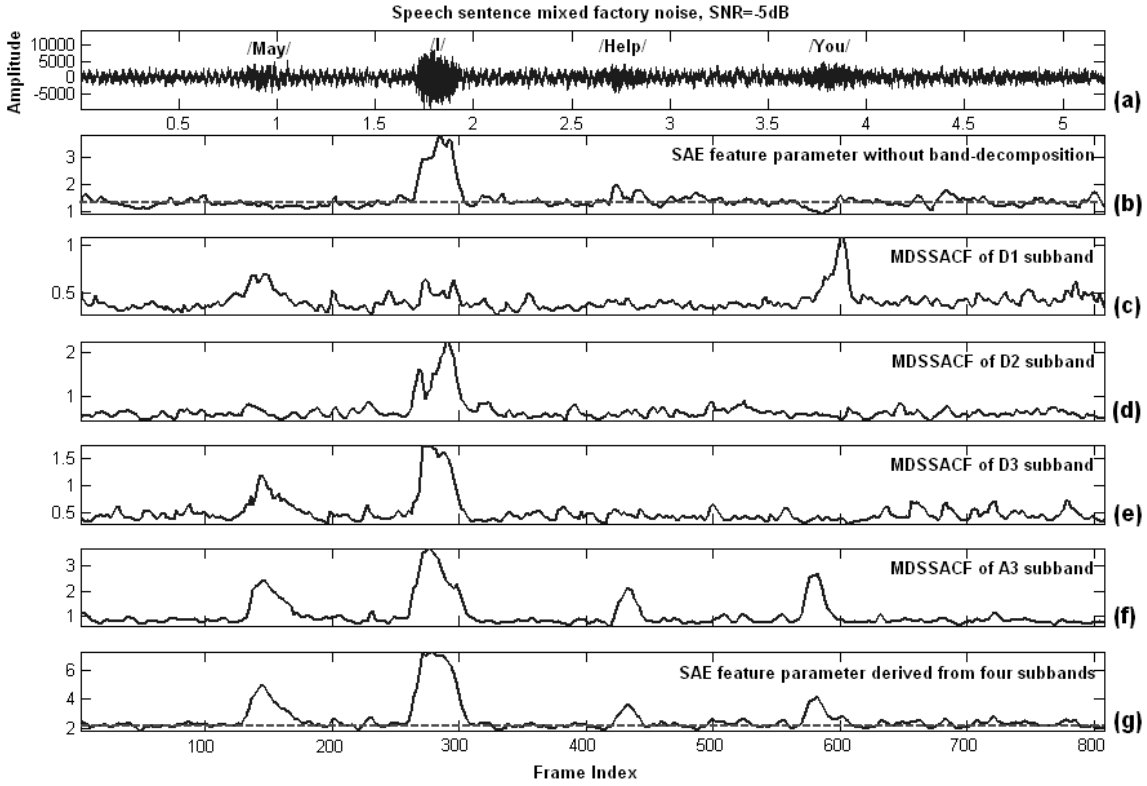


Figure 3. SAE with/without band-decomposition

4. Teager energy operator

The Teager energy operator (TEO) is a powerful nonlinear operator, and can track the modulation energy and identify the instantaneous amplitude and frequency [7-10].

In discrete-time, the TEO can be approximate by

$$\Psi_d[s(n)] = s(n)^2 - s(n+1)s(n-1)\,, \tag{4}$$

where $\Psi_d[s(n)]$ is called the TEO coefficient of discrete-time signal $s(n)$.

Figure 4 indicates that the TEO coefficients not only suppress noise but also enhance the detection of subband periodicity. TEO coefficients are useful for SACF to discriminate the difference between speech and noise in detail.

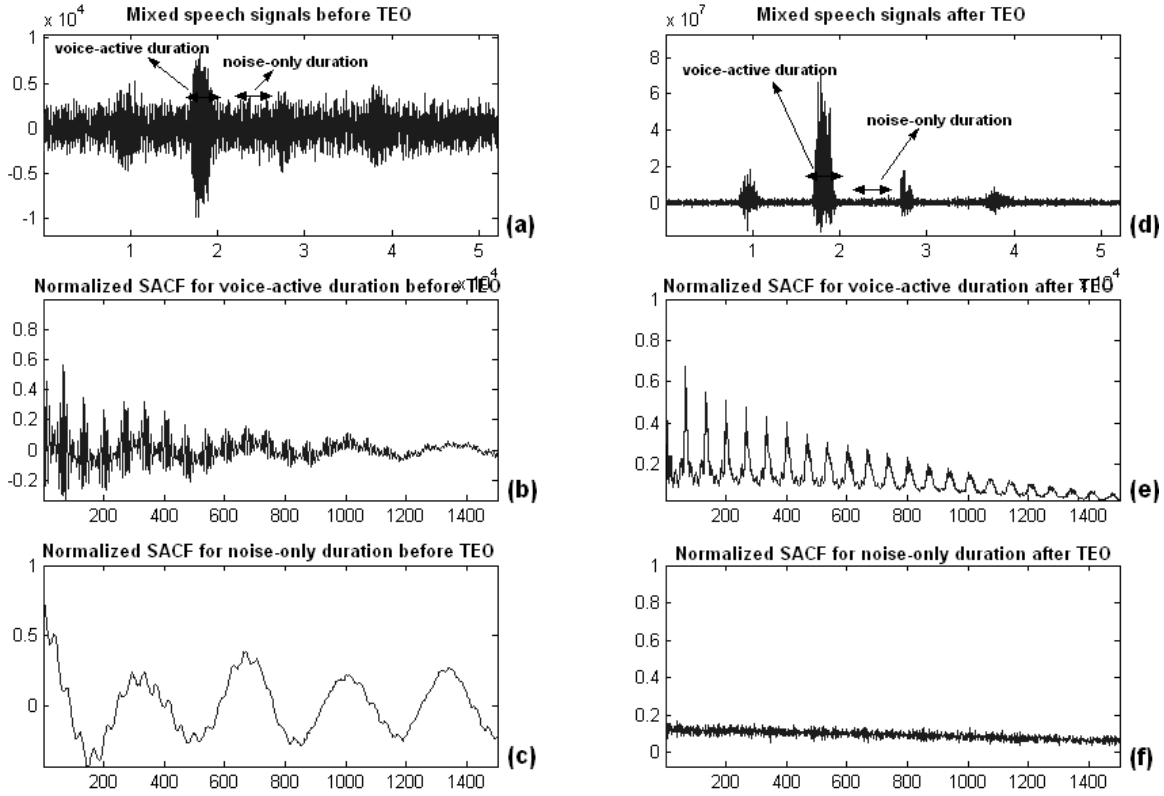


Figure 4. Illustration of TEO processing for the discrimination between speech and noise by using periodicity detection

5. Proposed voice activity detection algorithm

In this section, the proposed VAD algorithm based on DWT and TEO is presented. Fig. 8 displays the block diagram of the proposed wavelet-based VAD algorithm in detail. For a given layer j, the wavelet transform decomposed the noisy speech signal into $j+1$ subbands corresponding to wavelet coefficients sets $w_{k,n}^{j}$. In this case, three-layer wavelet decomposition is used to decompose noisy speech signal into four non-uniform subbands including three detailed scales and one appropriated scale. Let layer $j=3$,

$$w_{k,m}^{3} = DWT\{s(n),3\}, \quad n=1....N,\ k=1....4, \tag{5}$$

where $w_{k,m}^{3}$ defines the m^{th} coefficient of the k^{th} subband. N denotes as window length. The decomposed length of each subband is $N/2^{k}$ in turn.

For each subband signal, the TEO processing [8] is then used to suppress the noise

component, and also enhance the periodicity detection. In TEO processing,

$$t_{k,m}^{3} = \psi_d[w_{k,m}^{3}], \quad k = 1...4. \tag{6}$$

Next, the SACF measures the ACF defined in subband domain, and it can sufficiently discriminate the dissimilarity among of voiced, unvoiced speech sounds and background noises from wavelet coefficients. The SACF derived from the Teager energy of noisy speech is given by

$$R_{k,m}^{3} = R[t_{k,m}^{3}], \quad k = 1...4. \tag{7}$$

To count the intensity of periodicity from the envelope of the SACF accurately, the Mean-Delta (MD) method [9] is utilized on each subband.

The DSACF is given by

$$\dot{R}_{k,m}^{3} = \Delta[R_{k,m}^{3}], \quad k = 1...4. \tag{8}$$

where $\Delta[\cdot]$ denotes the operator of delta.

Then, the MDSACF is obtained by

$$\overline{R}_{k}^{3} = E[\dot{R}_{k,m}^{3}]. \tag{9}$$

where $E[\cdot]$ denotes the operator of mean.

Finally, we sum up the values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale and denote as SAE feature parameter given by

$$SAE = \sum_{k=1}^{4} \overline{R}_{k}^{3}. \tag{10}$$

6. Experimental results

In our first experiment, the results of speech activity detection are tested in three kinds of background noise under various values of the SNR. In the second experiment, we adjust the variable noise-level of background noise and mix it into the testing speech signal.

6.1. Test environment and noisy speech database

The proposed wavelet-based VAD algorithm is based on frame-by-frame basis (frame size = 1024 samples/frame, overlapping size = 256 samples). Three noise types, including white noise, car noise and factory noise, are taken from the Noisex-92 database in turn [11]. The speech database contains 60 speech phrases (in Mandarin and in English) spoken by 32 native speakers (22 males and 10 females), sampled at 8000 Hz and linearly quantized at 16 bits per sample. To vary the testing conditions, noise is added to the clean speech signal to create noisy signals at specific SNR of 30, 10, -5 dB.

6.2. Evaluation in stationary noise

In this experiment we only consider stationary noise environment. The proposed wavelet-based VAD is tested under three types of noise sources and three specific SNR values mentioned above. Table 1 shows the comparison between the proposed wavelet-based VAD and other two wavelet-based VAD proposed by Chen et al. [5] and J. Stegmann [12] and ITU standard VAD such as G.729B VAD [4], respectively. The results from all the cases involving various noise types and SNR levels are averaged and summarized in the bottom row of this table. We can find that the proposed wavelet-based VAD and Chen's VAD algorithms are all superior to Stegmann's VAD and G.729B over all SNRs under various types of noise. In terms of the average correct and false speech detection probabilities, the proposed wavelet-based VAD is comparable to Chen's VAD algorithm. Both the algorithms are based on the DWT and TEO processing. However, Chen et al. decomposed the input speech signal into 17 critical-subbands by using perceptual wavelet packet transform (PWPT). To obtain a robust feature parameter, called as "VAS" parameter, each critical subband after their processing is synthesized individually while other 16 subband signals are set to zero values. Next, the VAS parameter is developed by merging the values of 17 synthesized bands. Compare to the analysis/synthesis of wavelet from S. H. Chen et al., we only consider analysis of wavelet. The structure of three-layer decomposition leads into four non-uniform bands as front-end processing. For the development of feature parameter, we do not again waste extra computing power to synthesize each band. Besides, Chen's VAD algorithm must be performed in entire speech signal. The algorithm is not appropriate for real-time issue since it does not work on frame-based processing. Conversely, in our method the decisions of voice activity can be accomplished by frame-by-frame processing. Table 2 indicates that the computing time for the listed VAD algorithms running Matlab programming in Celeron 2.0G CPU for processing 118 frames of an entire recording. It is found that the computing time of Chen's VAD is nearly four times greater than that of other three VADs. Besides, the

computing time of Chen's VAD is closely relative to the entire length of recording.

Table 1. Comparison performance.

Noise Conditions		The probability of *correctly* detecting speech frames (%)				The probability of *falsely* detecting speech frames (%)			
Type	SNR(dB)	Proposed VAD	Chen's VAD	Stegmann's VAD	G.729B VAD	Proposed VAD	Chen's VAD	Stegmann's VAD	G.729B VAD
Car Noise	*30*	99.3	97.3	90.2	94.5	6.1	6.9	8.2	6.3
	10	97.8	96.1	85.3	90.3	8.4	9.3	13.5	12.3
	-5	92.9	93.5	79.1	82.7	10.5	10.9	16.6	17.5
Factory Noise	*30*	97.4	97.2	94.5	97.2	7.3	10.3	11.2	7.1
	10	93.2	94.1	83.1	88.4	8.8	13.2	14.6	13.4
	-5	87.8	85.6	75.3	80.7	10.7	15.4	17.3	19.2
White Noise	*30*	99.4	97.2	95.3	98.3	1.2	1.9	4.5	2.3
	10	98.6	98.1	90.1	86.3	1.3	1.8	6.7	2.9
	-5	93.4	92.9	85.8	84.3	1.5	2.3	10.1	3.7
Average		**95.5**	**94.7**	**86.5**	**89.2**	**6.2**	**8**	**11.4**	**9.4**

Table 2. Illustrations of subjective listening evaluation and the computing time

VAD types	Computing time (sec)
Proposed VAD	0.089
Chen's VAD [5]	0.436
Stegmann's VAD [12]	0.077
G.729B VAD [4]	0.091

6.3. Evaluation in non-stationary noise

In practice, the additive noise is non-stationary in real-world, since its statistical property change over time. We add the decreasing and increasing level of background noise on a clean speech sentence in English and the SNR is set 0 dB. Figure 6 exhibits the comparisons among proposed wavelet-based VAD, other one wavelet-based VAD respectively proposed by S. H. Chen et al. [5] and MD-based VAD proposed by A. Ouzounov [9]. Regarding to this figure, the mixed noisy sentence "May I help you?" is shown in Fig. 9(a). The increasing noise-level and decreasing noise-level are added into the front and the back of clean speech signal. Additionally, an abrupt change of noise is also added in the middle of clean sentence. The three envelopes of VAS, MD and SAE feature parameters are showed in Figure 6(b)~Figure

6(d), respectively. It is found that the performance of Chen's VAD algorithm seems not good in this case. The envelope of VAS parameter closely depends on the variable level of noise. Similarly, the envelope of MD parameter fails in variable level of noise. Conversely, the envelope of proposed SAE parameter is insensitive to variable-level of noise. So, the proposed wavelet-based VAD algorithm is performed well in non-stationary noise.

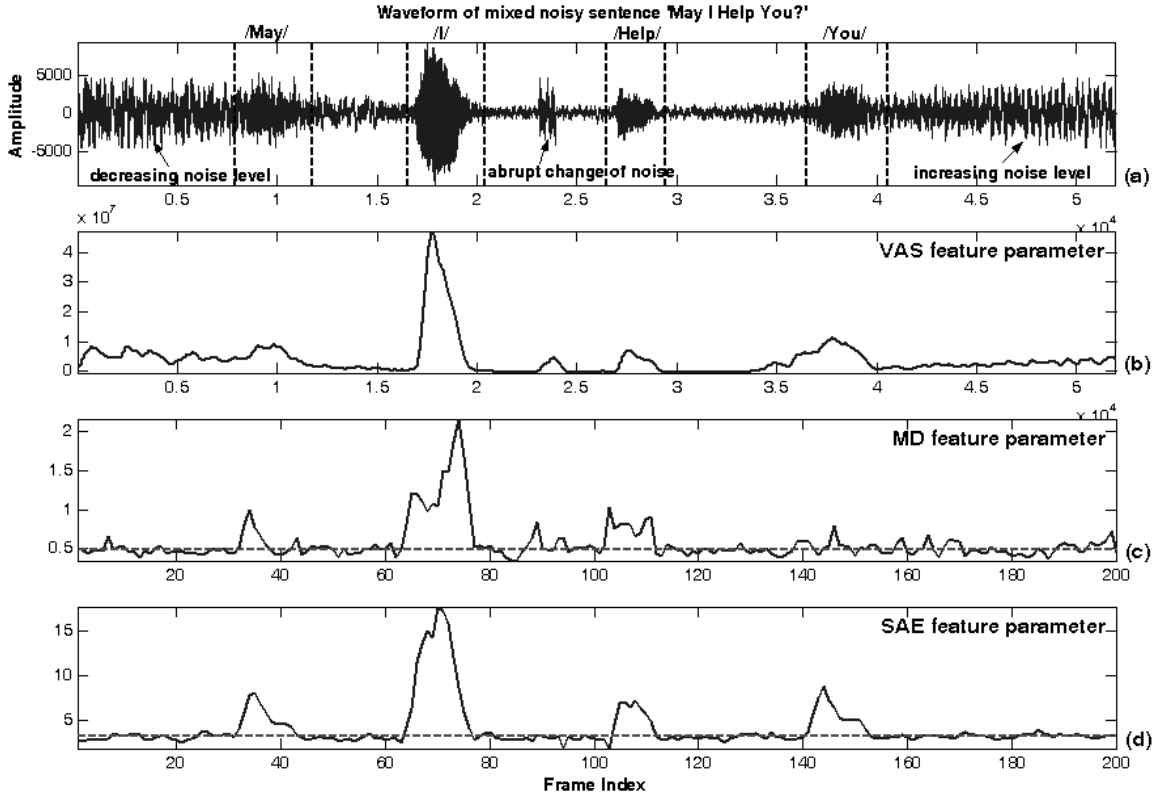


Figure 6. Comparisons among VAS, MD and proposed SAE feature parameters

7. Conclusions

The proposed VAD is an efficient and simple approach and mainly contains three-layer DWT (discrete wavelet transform) decomposition, Teager energy operation (TEO) and auto-correlation function (ACF). TEO and ACF are respectively used herein in each decomposed subband. In this approach, a new feature parameter is based on the sum of the values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale, and it has been shown that the SAE parameter can point out the boundary of speech activity and its envelope is insensitive to variable noise-level environment. By means of the MRA property of DWT, the ACF defined in subband domain sufficiently discriminates the dissimilarity among of voiced, unvoiced speech sounds and background

noises from wavelet coefficients. For the problem about noise suppression on wavelet coefficients, a nonlinear TEO is then utilized into each subband signals to enhance discrimination among speech and noise. Experimental results have been shown that the SACF with TEO processing can provide robust classification of speech due to that TEO can provide a better representation of formants resulting distinct periodicity.

References

[1] Cho, Y. D. and Kondoz, A., "Analysis and improvement of a statistical model-based voice activity detector", *IEEE Signal Processing Lett.*, Vol 8, 276-278, 2001.

[2] Beritelli, F., Casale, S. and Cavallaro, A., "A robust voice activity detector for wireless communications using soft computing", *IEEE J. Select. Areas Comm.*, Vol 16, 1818-1829, 1998.

[3] Nemer, E., Goubran, R. and Mahmoud, S., "Robust voice activity detection using higher-order statistics in the LPC residual domain", *IEEE Trans. Speech and Audio Processing*, Vol. 9, 217-231, 2001.

[4] Benyassine, A., Shlomot, E., Su, H. Y., Massaloux, D., Lamblin, C. and Petit, J. P., "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications", *IEEE Communications Magazine*, Vol. 35, 64-73, 1997.

[5] Chen, S. H. and Wang, J. F., "A Wavelet-based Voice Activity Detection Algorithm in Noisy Environments", *2002 IEEE International Conference on Electronics, Circuits and Systems (ICECS2002)*, 995-998, 2002.

[6] Kaiser, J. F., "On a simple algorithm to calculate the 'energy' of a signal", *in Proc. ICASSP'90*, 381-384, 1990.

[7] Maragos, P., Quatieri, T., and Kaiser, J. F., "On amplitude and frequency demodulation using energy operators", *IEEE Trans. Signal Processing*, Vol. 41, 1532-1550, 1993.

[8] Jabloun, F., Cetin, A. E., and Erzin, E., "Teager energy based feature parameters for speech recognition in car noise", *IEEE Signal Processing Lett.*, Vol. 6, 259-261, 1999.

[9] Ouzounov, A., "A Robust Feature for Speech Detection", *Cybernetics and Information*

Technologies, Vol. 4, No 2, 3-14, 2004.

[10] Stegmann, J., Schroder, G., and Fischer, K. A., "Robust classification of speech based on the dyadic wavelet transform with application to CELP coding", *Proc. ICASSP*, Vol. 1, 546 - 549, 1996.

[11] Varga, A. and Steeneken, H. J. M., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", *Speech Commun.*, Vol. 12, 247-251, 1993.

[12] Stegmann, J. and Schroder, G., "Robust voice-activity detection based on the wavelet transform", *IEEE Workshop on Speech Coding for Telecommunications Proceeding*, 99 - 100, 1997.

組合式倒頻譜統計正規化法於強健性語音辨識之研究
Associative Cepstral Statistics Normalization Techniques for Robust Speech Recognition

杜文祥 Wen-hsiang Tu
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
aero3016@ms45.hinet.net

吳光杰 Kuang-chieh Wu
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
s95323529@ncnu.edu.tw

洪志偉 Jeih-weih Hung
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
jwhung@ncnu.edu.tw

摘要

一套自動語音辨識系統，在雜訊環境下其辨識效果通常會受到明顯影響，該如何有效地克服這樣的問題，一直以來都是此領域研究的重點，本論文即是針對此問題加以研究，而提出幾種改進技術。在過去的研究中，有一系列的改進技術，是藉由正規化語音特徵的統計特性來降低雜訊的影響，例如：倒頻譜平均消去法、倒頻譜平均值與變異數正規化法與統計圖等化法等，這些方法被證明皆有明顯的效能，可以有效提升語音特徵在雜訊環境下的強健性。本論文即是以這三種倒頻譜特徵參數正規化技術爲背景，發展一系列改進之強健性方法。

前面所提到的三種特徵參數正規化技術中所須用到的特徵統計值，通常是由整段的語句或片段的語句所包含的特徵求得，而在過去本實驗室的研究中，曾運用以碼簿(codebook)爲基礎的方式來求取這些統計值，發現相對於之前的作法能有明顯進步。在本論文第一部分，我們提出一改良式的碼簿建構程序，其中使用語音偵測(voice activity detection, VAD) 技術來分隔訊號中的語音成分與非語音成分，然後利用語音部分的特徵來建構碼簿，同時對所建立之碼簿中的每個碼字(codeword)賦予權重(weight)，此程序所建構的碼簿，經實驗證實，可以提升原始碼簿式(codebook-based)特徵參數正規化法的效能。而在第二部份，我們則是整合上述之碼簿式(codebook-based)與整段式(utterance-based)兩類方法所得到之特徵統計資訊，發展出所謂的組合式(associative)特徵參數正規化法。此類組合式的新方法相較於整段式與碼簿式的方法，能得到更好的效果，更有效地提升加成性雜訊環境下語音的辨識精確度。

Abstract

The noise robustness property for an automatic speech recognition system is one of the most important factors to determine its recognition accuracy under a noise-corrupted environment. Among the various approaches, normalizing the statistical quantities of speech features is a

very promising direction to create more noise-robust features. The related feature normalization approaches include cepsral mean subtraction (CMS), cepstral mean and variance normalization (CMVN), histogram equalization (HEQ), etc. In addition, the statistical quantities used in these techniques can be obtained in an utterance-wise manner or a codebook-wise manner. It has been shown that in most cases, the latter behaves better than the former.

In this paper, we mainly focus on two issues. First, we develop a new procedure for developing the pseudo-stereo codebook, which is used in the codebook-based feature normalization approaches. The resulting new codebook is shown to provide a better estimate for the features statistics in order to enhance the performance of the codebook-based approaches. Second, we propose a series of new feature normalization approaches, including associative CMS (A-CMS), associative CMVN (A-CMVN) and associative HEQ (A-HEQ). In these approaches, two sources of statistic information for the features, the one from the utterance and the other from the codebook, are properly integrated. Experimental results show that these new feature normalization approaches perform significantly better than the conventional utterance-based and codebook-based ones. As the result, the proposed methods in this paper effectively improve the noise robustness of speech features.

關鍵詞：自動語音辨識、碼簿、強健性語音特徵

Keywords: automatic speech recognition, codebook, robust speech feature

一、緒論

本論文所討論及提出的強健式技術，主要是在加成性雜訊環境下，對訓練與測試二者的語音特徵參數的統計特性加以正規化，以降低兩環境的不匹配。其中我們利用梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)做爲語音特徵，結合語音偵測技術(voice activity detection, VAD)[1]與特徵統計值正規化的諸多技術，來提升語音特徵在加成性雜訊環境下的強健性。本論文中所討論的特徵參數正規化法分別爲：

（一）整段式(utterance-based)特徵參數正規化法

即傳統的整段式倒頻譜平均消去法(utterance-based cepstral mean subtraction, U-CMS)[2]、整段式倒頻譜平均值與變異數正規化法(utterance-based cepstral mean and variance normalization, U-CMVN)[3]與整段式統計圖等化法(utterance-based histogram equalization, U-HEQ)[4]。此類方法是以一整段語句爲基準去估算每一維特徵參數的統計特性，並執行特徵參數正規化。

（二）碼簿式(codebook-based)特徵參數正規化法

此類方法是藉由碼簿來幫助我們估算出代表訓練語音特徵與測試語音特徵的統計值，藉此執行語音特徵正規化。在過去的研究裡[5][6][7]，發現此類的方法，包括碼簿式倒頻譜平均消去法(codebook-based cepstral mean subtraction, C-CMS)與碼簿式倒頻譜平均值與變異數正規化法(codebook-based cepstral mean and variance normalization, C-CMVN)等，其效果都比前一類之整段式特徵正規化法來的好。

本論文根據以上所述的二類方法提出一系列改進的技術，分述如下：

① 在過去碼簿式特徵正規化法中[5-7]，碼簿取得方式是將全部的訓練語料轉換的特徵參數作向量量化，這樣的方式可能會使其中許多碼字是對應到非語音的靜音(silence)

或雜訊成份，而使這些碼字較缺乏語音特徵的代表性，同時，每個碼字的權重被設爲相等，這樣可能會使之後所欲計算的特徵統計值較不精確。在本論文中，我們應用端點偵測(voice activity detection, VAD)技術偵測出一段訊號的語音(speech)成分與非語音(silence)成分，然後只使用語音成分的特徵去製作碼簿，同時，不同的碼字根據其對應的原始特徵數目多寡設定其權重(weight)，這種新的碼簿建構程序應可以改善上述之缺點，進而提升各種碼簿式特徵正規化法的效能。

② 我們提出了一新方法，稱爲組合式(associative)特徵正規化法，其主要程序是我們整合前述之碼簿式與整段式兩方所使用的特徵統計資訊，來計算特徵的統計值，藉此來執行特徵的正規化。實驗結果發現此類組合式的方法比碼簿式與整段式的兩類方法，能達到更佳的效果。可能原因在於，組合式的方法降低了碼簿式方法中只取每段訊號前幾個音框作爲純雜訊估測的不準確效應，而使所得的特徵統計值更爲精確。

在之後的第二章裡，我們將簡單介紹整段式(utterance-based)特徵正規化技術。第三章將說明新的虛擬雙通道碼簿的建立程序，藉此改進碼簿式(codebook-based)特徵正規化法的效能。在第四章中，我們敘述本論文所新提出的組合式(associative)特徵正規化法。第五章包含了本論文之實驗所使用之語料庫介紹與本論文所提到的各種特徵正規化技術之實驗結果與相關的討論分析。最後，第六章爲結論與未來展望。

二、整段式(utterance-based)特徵參數正規化技術

本章我們簡要介紹三種在強健性語音辨識中，常被應用的特徵參數正規化技術，分別爲整段式倒頻譜平均消去法(utterance-based cepstral mean subtraction, U-CMS)[2]、整段式倒頻譜平均值與變異數正規化法(utterance-based cepstral mean and variance normalization, U-CMVN)[3]與整段式倒頻譜統計圖等化法(utterance-based cepstral histogram equalization, U-HEQ)[4]。

（一）整段式倒頻譜平均消去法 (U-CMS)

倒頻譜平均消去法(CMS)的目的是希望一語音特徵序列中，每一維度的倒頻譜係數長時間平均值爲0。假設其值不爲0時，我們就將此視爲通道雜訊而加以扣除，此種方法對於降低通道雜訊效應是一種簡單且有用的技術，但是有時對於降低加成性雜訊上也有一定的效果。在多數的作法上，首先我們將整段語音每一維的倒頻譜係數取平均值，然後將每一維的係數減掉其平均值，如此即得到補償後之新特徵，此稱爲整段式倒頻譜平均消去法(utterance-based cepstral mean subtraction, U-CMS)。根據這樣的原則，我們假設 $\{X[n], n = 1,2,...,N\}$ 爲一段語音所擷取到的某一維倒頻譜特徵參數序列，在經過整段式倒頻譜平均消去法(U-CMS)處理後，得到新的經過補償的特徵參數序列 $\{X_{U-CMS}[n], n = 1,2,...,N\}$，其數學式如下所示：

$$X_{U-CMS}[n] = X[n] - \mu_X, \quad n = 1,2,...,N. \qquad \text{式(2.1)}$$

其中 $\mu_X = \frac{1}{N}\sum_{n=1}^{N} X[n]$，$N$ 爲整段語音的音框個數。

因此，在 U-CMS 法中，用以正規化的平均值 μ_X 是由原始整段的特徵序列所得。

（二）整段式倒頻譜平均值與變異數正規化法 (U-CMVN)

語音訊號在經過加成性雜訊的干擾之後，其倒頻譜之平均值和原本乾淨語音倒頻譜平均值之間通常會存在一偏移量(bias)，同時其變異數相對於乾淨語音倒頻譜參數的變異數而言則通常會有縮小的現象，如此便造成了訓練與測試特徵的不匹配，而降低辨識

效果。使用倒頻譜平均值與變異數正規化法(CMVN)的目的是把每一維的倒頻譜特徵參數之平均值正規化爲 0，並將其變異數正規化爲 1，如此便能降低上述的失真，以達到提升倒頻譜特徵參數的強健性。

在倒頻譜平均值與變異數正規化法(CMVN)的作法上，我們是先利用倒頻譜平均消去法(CMS)去作處理(使處理過後的每一維倒頻譜係數平均值爲0)，然後再將處理後的每一維倒頻譜係數除以其標準差，如此得到新的特徵序列。在U-CMVN(utterance-based cepstral mean and variance normalization)法中，假設$\{X[n], n=1,2,...,N\}$是一段語音的某一維倒頻譜特徵參數序列，在經過U-CMVN處理後，得到新的特徵參數$\{X_{U-CMVN}[n], n=1,2,...,N\}$，其數學式如下所示：

$$X_{U-CMVN}[n]=\frac{X[n]-\mu_X}{\sigma_X},\quad n=1,2,...,N. \qquad \text{式(2.2)}$$

其中 $$\mu_X=\frac{1}{N}\sum_{n=1}^{N}X[n],\ \sigma_X=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(X[n]-\mu_X\right)^2}$$

因此，在U-CMVN中，所用的平均值μ_X與標準差σ_X皆由整段語音的特徵序列而得。

（三）整段式統計圖等化法(U-HEQ)

統計圖等化法(HEQ)的目的，是希望用以訓練與測試之語音特徵兩者能夠具有相同的統計分佈特性，藉由此匹配的轉換過程，降低測試特徵與訓練特徵之間由於雜訊影響所造成的不匹配情形。其作法是將測試語音特徵與訓練語音特徵的機率分佈同時逼近一參考機率分佈。在本論文中所使用的參考機率分佈爲一標準常態分佈。

根據上述，我們假設$\{X[n], n=1,2,...,N\}$爲一段語音某一維倒頻譜特徵參數序列；$F_X(x)$爲$X[n]$的機率分佈$\left(F_X(x)=P(X\le x)\right)$，它是由整段之特徵$\{X[n], n=1,2,...,N\}$求得；$F_N(x)$爲參考機率分佈。則整段式統計圖等化法(utterance-based histogram equalization, U-HEQ)的數學轉換式如下所示：

$$X_{U-HEQ}[n]=F_N^{-1}\left(F_X\left(X[n]\right)\right), \qquad \text{式(2.3)}$$

其中$X_{U-HEQ}[n]$即爲經過整段式統計圖等化法處理後的新特徵參數。

三、改良式碼簿式特徵參數正規化技術

運用所謂的虛擬雙通道碼簿(pseudo stereo codebooks)來估算乾淨語音與含雜訊語音之特徵統計特性，進而執行特徵參數正規化技術，能有效提升雜訊環境下語音辨識率。在過去研究中[5-7]所提出之倒頻譜統計補償法(cepstral statistics compensation)，是對含雜訊之語音倒頻譜係數做轉換，使得經過轉換後的語音倒頻譜特徵之統計值更相似於乾淨訓練語音倒頻譜的統計值，這種方式只針對雜訊語音特徵作倒頻譜正規化補償。而在本論文所提出之方式，則是同時針對乾淨語音與雜訊語音倒頻譜特徵參數作正規化處理。另外，在之前的倒頻譜統計補償法中，所用的每個碼字(codeword)是利用未處理的乾淨語音特徵訓練而得，且每個碼字的比重相同，而在我們改進的方法上，我們應用了語音偵測技術(voice activity detection, VAD)[1]處理乾淨語音訊號，將訊號中的語音區段與非語音區段區隔出來，然後利用純語音區段的語音特徵來訓練碼字，此外，這些碼字根據其涵蓋的特徵數目賦予不同的權重(weight)，因此，由這些碼字所計算的語音特徵統計值，應該更爲精確、更能代表語音特徵的特性。實驗證明，這樣的修正方式能帶來更好的辨識率。

在上一章，我們介紹了三種整段式 (utterance-based)特徵參數正規化技術，分別爲：U-CMS、U-CMVN與U-HEQ。在這裡，我們將利用新修正的碼簿建立方法，建立虛擬雙通道碼簿，執行一系列改良的碼簿式(codebook-based)特徵參數正規化技術。

（一）虛擬雙通道碼簿之建立方式

在原始的碼簿式特徵參數正規化法 [5-7] 中，碼簿之建立方式是將訓練語料庫裡所有的乾淨語音訊號，在轉換至梅爾倒頻譜特徵參數之過程中，保留下語音與雜訊具備線性相加特性的中介特徵參數(intermediate feature)，並且將這些乾淨語音之中介特徵參數訓練成一組碼簿(codebook)，此一乾淨語音碼簿，大致上可以代表乾淨語音在中介特徵參數的特性。在測試語音方面，對於每一句含雜訊的測試語音，假設其前端部分爲純雜訊，然後將這段純雜訊轉換至上述的中介特徵參數，由於乾淨語音與純雜訊在中介特徵參數域具有線性相加(linearly additive)的特性，因此將這些純雜訊的中介特徵參數直接線性相加於先前訓練好的乾淨語音的每個碼字上，便得到了代表雜訊語音(noisy speech)在中介特徵參數的碼簿。最後，將這兩組分別代表乾淨語音與雜訊語音在中介特徵參數域中的碼字轉換至倒頻譜域，所得的兩組倒頻譜特徵碼簿，就稱爲虛擬雙通道碼簿。

在本論文中所提出改良式的碼簿建立法，與原始的碼簿建立法的兩個不同點在於：

(1) 將訓練語料庫裡所有的乾淨語音訊號，先利用文獻[1]所提之語音偵測技術(voice activity detection, VAD)偵測出語音(speech)與靜音(silence)成分，然後只使用語音部分的中介特徵參數來訓練乾淨語音的碼簿。而在原始的方法裡，僅是使用未上述處理的乾淨語音訊號之中介特徵訓練碼簿。

(2) 不同的碼字根據其涵蓋的特徵量，指定不同的權重(weight)，亦即涵蓋較多量特徵的碼字，所佔的權重也就愈大，此意味著每個碼字的出現機率並不相同。這些權重可以用來幫助後續的特徵統計正規化法裡，估測更精準的特徵統計量。而在原始的方法裡，每個碼字未被賦予權重，其隱含了每個碼字的出現機率是均等(uniform)的。

以下，我們詳述此虛擬雙通道碼簿之建立過程：

我們先將語料庫中每一句乾淨語料，透過語音偵測技術[1]區隔出乾淨訓練語料中，屬於語音區段的部份，然後經由梅爾倒頻譜特徵參數(mel-frequency cepstral coefficients, MFCC)擷取流程的前半部，將此屬於語音區段的部份，轉換成一中介特徵向量(intermediate feature vector)的序列，此中介特徵爲梅爾濾波器之輸出值，也就是平緩化後之線性頻譜(linear spectrum)，這些由乾淨語料所得的中介特徵向量，透過向量量化(vector quantization, VQ)後，建立一組包含 M 個碼字的集合，以 $\{\tilde{x}[n] \mid 1 \le n \le M\}$ 來表示，同時，其對應的權重爲 $\{w_n \mid 1 \le n \le M\}$。這組在中介特徵參數域上的乾淨語音碼簿之所有碼字，再由 MFCC 擷取流程的後半部轉換至倒頻譜域，如下式所示：

$$x[n] = f\left(\tilde{x}[n]\right) \qquad \text{式(3.1)}$$

其中 $f(.)$ 代表轉換程序，因此，$\{x[n], w_n \mid 1 \le n \le M\}$ 爲轉換至倒頻譜的碼簿及權重值，這就是乾淨語音的倒頻譜碼簿及權重值。

在雜訊語音方面，我們藉由乾淨語音在中介特徵參數域上的碼字，來建立對應至該段含雜訊之測試語音的碼簿。我們將每一測試語音估測到的純雜訊，在中介特徵參數域（線性頻譜域）上用一組向量 $\{\tilde{n}[p] \mid 1 \le p \le P\}$ 來表示，由於乾淨語音與純雜訊在中介特徵參數域上具有線性相加的特性，因此雜訊語音的碼字可表示成下式：

$$\tilde{y}[m]\big|_{m=(n-1)P+p} = \{\tilde{x}[n] + \tilde{n}[p]\}, \qquad \text{式(3.2)}$$

最後，類似式(3.1)，我們將 $\tilde{y}[m]$ 經由 MFCC 擷取流程後半部轉換至倒頻譜域，如下式所示：

$$y[m] = f(\tilde{y}[m]), \qquad \text{式(3.3)}$$

此外，每個$y[m]$的權重值v_m則設定爲：

$$v_m\Big|_{m=(n-1)P+p} = \frac{w_n}{P}, \qquad \text{式(3.4)}$$

因此，$y[m]$之權重（即v_m）是其對應的乾淨語音碼字$x[n]$之權重w_n的$\frac{1}{P}$，其中P是純雜訊向量$\{\tilde{n}[p]\}$的個數。故$\{y[m], v_m \mid 1 \le m \le MP\}$便是代表此句雜訊語音在倒頻譜域上的碼簿及權重值。$\{x[n], w_n\}$與$\{y[m], v_m\}$這兩組分別代表乾淨訓練語音與雜訊測試語音的碼字，我們稱之爲虛擬雙通道碼簿。所謂虛擬的意思，是因爲雜訊語音的碼簿並不是直接由雜訊語音得到，而是經由乾淨語音碼簿與純雜訊估算值所間接得到的。

（二）碼簿式特徵參數正規化技術

這一節中，我們將介紹碼簿式特徵參數正規化技術。在前面曾提到，此類正規化技術，是同時針對乾淨語音與雜訊語音倒頻譜特徵參數作處理。而在這裡的碼簿式特徵參數正規化技術，是藉由在前一節中描述的虛擬雙通道碼簿，來建立特徵之統計量，進而對特徵做正規化。這三種特徵參數正規化技術分別爲：倒頻譜平均消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)、與倒頻譜統計圖等化法(HEQ)。對於CMS與CMVN而言，我們利用前一節所述之碼簿與權重$\{x[m], w_m\}$與$\{y[m], v_m\}$，計算出分別代表乾淨語音與雜訊語音特徵的近似統計值，如下式所示：

$$\mu_{X,i} \approx \sum_{n=1}^{N} w_n (x[n])_i, \quad \sigma_{X,i}^2 \approx \sum_{n=1}^{N} w_n \left(x[n]\right)_i^2 - \left(\mu_{X,i}\right)^2. \qquad \text{式(3.5)}$$

$$\mu_{Y,i} \approx \sum_{m=1}^{NP} v_m (y[m])_i, \quad \sigma_{Y,i}^2 \approx \sum_{m=1}^{NP} v_m \left(y[m]\right)_i^2 - \left(\mu_{Y,i}\right)^2. \qquad \text{式(3.6)}$$

其中$(u)_i$代表任意向量u之第i維，$\mu_{X,i}$與$\sigma_{X,i}^2$分別代表乾淨語音特徵向量x第i維的平均值與變異數；$\mu_{Y,i}$與$\sigma_{Y,i}^2$分別代表雜訊語音特徵向量y第i維的平均值與變異數，和之前文獻[5-7]中的方法明顯差異在於，此刻我們所用的統計值(平均值與變異數)是以加權平均(weighted average)的形式所測得，而非[5-7]中之均勻平均(uniform average)的形式。

碼簿式倒頻譜平均消去法(codebook-based cepstral mean subtraction, C-CMS)，是對倒頻譜特徵之平均值作正規化處理，其數學表示式如下：

$$(\overline{x})_i = (x)_i - \mu_{X,i}, \quad (\overline{y})_i = (y)_i - \mu_{y,i}. \qquad \text{式(3.7)}$$

其中$\overline{x}$與$\overline{y}$分別爲乾淨語音特徵x與雜訊語音特徵y在經過 C-CMS 處理後的新特徵值。

而碼簿式倒頻譜平均值與變異數正規化法(codebook-based cepstral mean and variance normalization, C-CMVN)，是針對倒頻譜特徵之平均值與變異數做正規化處理，其數學表示式如下：

$$(\overline{x})_i = \frac{(x)_i - \mu_{X,i}}{\sigma_{X,i}}, \quad (\overline{y})_i = \frac{(y)_i - \mu_{Y,i}}{\sigma_{Y,i}}. \qquad \text{式(3.8)}$$

其中$\overline{x}$與$\overline{y}$分別爲乾淨語音特徵x與雜訊語音特徵y經過 C-CMVN 處理後的新特徵值。

最後，碼簿式倒頻譜統計圖等化法(codebook-based cepsteral histogram equalization, C-HEQ)，其基本作法是利用$\{x[n], w_n\}$與$\{y[m], v_m\}$兩套碼簿分別計算出乾淨語音特徵與雜訊語音特徵之每一維之近似的機率分佈(probability distribution)，然後求一轉換函數，使二者之每一維特徵參數之機率分佈皆逼近於某一事先定義之參考機率分佈。具體

作法如下描述：

假設我們現在藉由碼簿$\{x[n], w_n\}$建立第i維乾淨語音特徵$(x)_i$的累積密度函數，由於碼簿本身意味著離散的形式，若我們假設第$(x)_i$本身對應之隨機變數爲X_i，則X_i的機率質量函數(probability mass function)可用下式表示：

$$P\left(X_i = \left(x[n]\right)_i\right) = w_n, \qquad \text{式(3.9)}$$

而X_i的機率密度函數(probability density function, pdf)，即可以下式表示：

$$f_{X_i}(x) = \sum_{n=1}^{M} w_n \delta\left(x - \left(x[n]\right)_i\right); \qquad \text{式(3.10)}$$

其中$\delta(\cdot)$爲單位脈衝(unit impulse)函數，故X_i之機率分佈，或稱爲累積機率密度函數(cumulative density function)，爲上式$f_{X_i}(x)$之積分，表示如下：

$$F_{X_i}(x) = P\left(X_i \le x\right) = \sum_{n=1}^{M} w_n u\left(x - \left(x[n]\right)_i\right); \qquad \text{式(3.11)}$$

其中$u(x)$爲單位步階函數(unit step function)，定義爲：

$$u(x) = \begin{cases} 1, & x \ge 0 \\ 0, & x < 0 \end{cases} \qquad \text{式(3.12)}$$

因此，第i維乾淨語音特徵$(x)_i$之機率分佈則可由式(3.11)的$F_{X_i}(x)$表示，同理，藉由碼簿$\{y[m], v_m\}$建立之第i維雜訊語音特徵$(y)_i$的機率分佈可由下式表示：

$$F_{Y_i}(y) = P\left(Y_i \le y\right) = \sum_{m=1}^{MP} v_m u\left(y - \left(y[m]\right)_i\right); \qquad \text{式(3.13)}$$

由上述作法得到$F_{X_i}(x)$與$F_{Y_i}(y)$之後，根據倒頻譜統計圖等化法(HEQ)的原理，我們利用下面兩式分別正規化第i維之訓練乾淨語音特徵$(x)_i$與測試雜訊語音特徵$(y)_i$：

$$(\bar{x})_i = F_N^{-1}\left(F_{X_i}\left((x)_i\right)\right), \qquad \text{式(3.14)}$$

$$(\bar{y})_i = F_N^{-1}\left(F_{Y_i}\left((y)_i\right)\right). \qquad \text{式(3.15)}$$

其中$F_N(\bullet)$爲一參考機率分佈(通常爲標準常態分佈)，$F_N^{-1}(\bullet)$爲$F_N(\bullet)$的反函數，$\bar{x}$與$\bar{y}$則爲經C-HEQ正規化後的新特徵值。

綜合以上所述，在過去的碼簿式特徵參數正規化技術中，所用的碼字是利用原始未分段之乾淨訊號特徵訓練而得，且每個碼字的比重皆相同，而在這裡所提出的改良式碼簿建立法上，我們應用語音偵測技術先將乾淨語音訊號中的語音區段與非語音區段區隔出來，然後利用語音區段的特徵訓練碼字。接著，根據不同的碼字所涵蓋的特徵數目賦予相對之權重(weight)，因此，這些碼字所計算出的語音特徵統計值或機率分佈，應當更爲精確而具代表性。在第四章的實驗結果中，將證明藉由此改良式碼簿所發展的碼簿式特徵參數正規化法，能獲得更好的辨識效果。

四、組合式特徵參數正規化技術

前一章提到，雖然碼簿式特徵參數正規化法之表現普遍比整段式的方法來的好，且

具備了即時運算的優點，但其可能的缺點在於純雜訊資訊不足，導致所得的雜訊語音碼簿不夠精準。因此，本章我們針對上述缺點，提出組合式的特徵參數正規化技術，簡單來說，我們在這些方法中，整合了之前所介紹的碼簿式與整段式方法所求取之特徵統計特性，希望得到更精確的統計值來執行各種特徵正規化法。這些方法，我們統稱爲組合式(associative)特徵參數正規化法。以下兩小節，我們便對組合式倒頻譜平均消去法(associative CMS, A-CMS)、組合式倒頻譜平均值與變異數正規化法(associative CMVN, A-CMVN)與組合式倒頻譜統計圖等化法(associative HEQ, A-HEQ)分別作介紹。

（一）組合式倒頻譜平均消去法(associative CMS, A-CMS)與組合式倒頻譜平均值與變異數正規化法(associative CMVN, A-CMVN)

這一節中將分別介紹 A-CMS 與 A-CMVN 兩種特徵參數正規化法。我們藉由一參數值α的調整，適當地整合碼簿與整段特徵之統計資訊，希望能達到較佳之辨識效果。就整段語句(utterance)的特徵而言，假設$X=\{X_1,X_2,...,X_N\}$爲一段訓練用或測試用語音在所擷取到的某一維倒頻譜特徵參數序列，則其整段式之特徵的平均值與變異數可由下兩式計算而得：

$$\mu_u=\frac{1}{N}\sum_{i=1}^{N}X_i, \qquad \text{式(4.1)}$$

$$\sigma_u^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_u)^2, \qquad \text{式(4.2)}$$

其中μ_u爲整段式之特徵平均值，σ_u^2爲整段式之特徵變異數，N爲整段語音的音框數。

而在碼簿上的特徵方面，假設$C=\{C_1,C_2,...,C_M\}$爲同一段語音對應到的各碼字(codewords)的某一維(與前一段所述之維值相同)之集合，則此段語音特徵之碼簿式的平均值與變異數可由下兩式計算而得：

$$\mu_c=\sum_{j=1}^{M}w_jC_j, \qquad \text{式(4.3)}$$

$$\sigma_c^2=\sum_{j=1}^{M}w_jC_j^2-\mu_c^2, \qquad \text{式(4.4)}$$

其中μ_c爲碼簿式之特徵平均值，$\sigma_c{}^2$爲碼簿式之特徵變異數，w_j爲每一碼字所對應到的權重，M爲碼字數目。

因此，組合式倒頻譜平均消去法(associative CMS, A-CMS)中，所使用的特徵參數之平均值μ_a，可由下式計算而得：

$$\mu_a=\alpha\cdot\mu_c+(1-\alpha)\cdot\mu_u, \qquad \text{式(4.5)}$$

其中μ_u與μ_c分別如式(4.1)與式(4.3)所示，而α爲一權重值，$0\le\alpha\le1$。
因此，A-CMS 處理後的新特徵參數，可表示爲：

$$\text{A-CMS:}\qquad \tilde{X}_i=X_i-\mu_a,\qquad 1\le i\le N. \qquad \text{式(4.6)}$$

而組合式倒頻譜平均值與變異數正規化法(associative CMVN, A-CMVN)中，所使用的特徵參數之平均值μ_a與變異數σ_a^2，可由下面兩式計算而得：

$$\mu_a=\alpha\cdot\mu_c+(1-\alpha)\cdot\mu_u, \qquad \text{式(4.7)}$$

$$\sigma_a^2 = \left(\alpha\left(\sigma_c^2 + \mu_c^2\right) + (1-\alpha)\left(\sigma_u^2 + \mu_u^2\right)\right) - \mu_a^2, \qquad \text{式(4.8)}$$

其中μ_u、μ_c、σ_u^2與σ_c^2分別如式(4.1)、式(4.3)、式(4.2)與式(4.4)所示，而α爲一權重值，$0 \le \alpha \le 1$。

A-CMVN 處理後的新特徵參數，可表示爲：

$$\text{A-CMVN:} \qquad \tilde{X}_i = \frac{X_i - \mu_a}{\sigma_a} \qquad \text{式(4.9)}$$

由式(4.5)、式(4.7)與式(4.8)可明顯看出，α的大小決定了組合式方法中，使用碼簿式統計量與整段式統計量的比例。當$\alpha = 1$時，A-CMS或A-CMVN即爲原始之碼簿式CMS(C-CMS)或碼簿式CMVN(C-CMVN)，相反地，當$\alpha = 0$時，A-CMS或A-CMVN即爲原始之整段式CMS(U-CMS)或整段式CMVN(U-CMVN)。

（二）組合式倒頻譜統計圖等化法(associative HEQ, A-HEQ)

在這一節中，我們將介紹組合式統計圖等化法(associative histogram equalization, A-HEQ)，類似之前的觀念，我們試著整合單一語句(utterance)特徵及其對應之碼字組合(codebook)兩方的統計資訊，然後建構出一代表此語句特徵的機率分佈$F_X(x) = P(X \le x)$，以作爲 HEQ 法等化特徵所用。以下，我們描述 A-HEQ 執行步驟：

假設某一待正規化的原語句之特定一維的特徵序列爲$\{X_1, X_2, ..., X_N\}$，其中N爲此序列之特徵總數，而其對應到之同一維的碼字，表示爲$\{C_1, C_2, ..., C_M\}$，權重爲$\{w_1, w_2, ..., w_M\}$，其中M爲碼字數目。首先，我們設定一參數β（$0 \le \beta \le \infty$），此參數代表了使用碼簿式資訊相對於使用整段式資訊的比例。接著，我們產生一組數目爲βN的新特徵$\{\tilde{C}_k\}$，此組新特徵是由碼字$\{C_m\}$根據其權重值$\{w_m\}$所建立的，新特徵$\{\tilde{C}_k\}$中有$[\beta N \times w_m]$個特徵的值和C_m完全相同，（$[\beta N \times w_m]$代表$\beta N \times w_m$取四捨五入後的值），換言之，新特徵$\{\tilde{C}_k\}$爲一組整合了權重值的新碼字，當原碼字C_m其權重值爲w_m時，它就會在新特徵$\{\tilde{C}_k\}$中出現$[\beta N \times w_m]$次，例如，假設原碼字集合爲$\{3,5,7\}$，對應之權重爲$\{0.2, 0.5, 0.3\}$，則當假設新特徵$\{\tilde{C}_k\}$的總數爲 20 時，$\{\tilde{C}_k\}$就包括了 4 個 3 $(20 \times 0.2 = 4)$，10 個 5 $(20 \times 0.5 = 10)$與 6 個 7 $(20 \times 0.3 = 6)$，因此，$\{\tilde{C}_k\}$即爲$\{\underbrace{3,3,3,3}_{4個}, \underbrace{5,5,5,...,5}_{10個}, \underbrace{7,7,7,...,7}_{6個}\}$（實際上，由於四捨五入的關係，最後得到的新特徵$\{\tilde{C}_k\}$其總數可能不會恰好是$\beta N$，即恰好爲原語句特徵數目$N$的$\beta$倍）。

接下來，我們就將原語句特徵$\{X_1, X_2, ..., X_N\}$與代表碼字的新特徵$\{\tilde{C}_1, \tilde{C}_2, ..., \tilde{C}_{\beta N}\}$串聯起來，共同決定一組代表此語句特徵的機率分佈：

$$F_X(x) = \frac{1}{(\beta+1)N}\left(\sum_{n=1}^{N} u(x - X_n) + \sum_{k=1}^{\beta N} u\left(x - \tilde{C}_k\right)\right), \qquad \text{式(4.10)}$$

最後，利用 HEQ 的原理，我們將原語句特徵正規化，如下式所示：

$$\text{A-HEQ:} \qquad \bar{x} = F_N^{-1}\left(F_X(x)\right) \qquad \text{式(4.11)}$$

其中F_N爲參考之機率分佈，x爲原始特徵參數(即前面提到的$\{X_1, X_2, ..., X_N\}$)，$\bar{x}$即爲 A-HEQ 法所得之新特徵參數。

由式(4.10)可看出，原語音特徵之機率分佈 $F_X(x)$ 由整句特徵 $\{X_n\}$ 與新碼字特徵 $\{\tilde{C}_k\}$ 共同決定，前者數目爲 N ，後者數目約爲 βN ，因此參數 β 大小決定了A-HEQ中，新碼字特徵 $\{\tilde{C}_k\}$ 對 $F_X(x)$ 的影響程度，當 $\beta = 0$ 時，相當於碼字方面的資訊完全被忽略，A-HEQ即變爲原先所介紹之整段式HEQ法(U-HEQ)，而當 β 很大($\beta \to \infty$)時，原先語句的特徵 $\{X_n\}$ 之資訊則幾乎被省略，則此時A-HEQ即趨近於原先所介紹之碼簿式HEQ (C-HEQ)。

在這一章中，我們介紹了組合式特徵正規化技術，這類技術同時整合了前兩章所介紹之整段式與碼簿式技術所用的特徵統計資訊，透過式(4.5)、式(4.7)、式(4.8)與式(4.10)中之參數 α 與 β 的調整，我們可以彈性地決定兩方所得之統計資訊的比例。在下一章的實驗結果，我們將看到這類組合式特徵正規化技術能帶來更好的語音辨識精確度。

五、辨識實驗結果與相關討論

本章開始是介紹在本論文上所使用的語音資料庫與系統效能的評估方式，而後的內容爲本論文所提及之各種強健性語音特徵參數技術之辨識實驗，其相關結果與討論。

（一）語音資料庫簡介

本論文使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute，ETSI)發行的AURORA 2語音資料庫[8]，內容是連續的英文數字字串，其中是以美國成年男女所錄製的乾淨環境連續數字語音，然後加上了八種不同的加成性雜訊與通道效應。這些加成性雜訊分別爲：地下鐵(subway)、人的嘈雜聲(babble)、汽車(car)、展覽會(exhibition)、餐廳(restaurant)、街道(street)、機場(airport)、火車站(train station)等環境雜訊共計八種，而通道效應有兩種，分別爲G712與MIRS[9]。

在AURORA 2資料庫裡有兩種不同的訓練環境及三種不同的測試環境，由於本論文只針對加成性雜訊做討論，因此在這裡，只使用到表一之一種訓練環境與兩種測試環境。

（二）實驗設定

本論文中所使用的特徵參數爲13維（第0維至第12維）的梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)，加上其一階和二階差量，總共爲39維的特徵參數。模型的訓練是使用隱藏式馬可夫模型工具(Hidden Markov Model Toolkit，HTK)[10]來訓練，產生11個數字模型(oh, zero, one, …., nine)與一個靜音模型，每個數字模型包含16個狀態，每個狀態包含20個高斯密度混合。

（三）各種強健性技術之辨識結果與討論

1. 改良之碼簿式特徵正規化法的辨識結果

在這一節中，我們將介紹本論文所提出之新的碼簿建立程序，分別應用於碼簿式倒頻譜平均消去法(C-CMS)、碼簿式倒頻譜平均值與變異數正規化法(C-CMVN)與碼簿式倒頻譜統計圖等化法(C-HEQ)的辨識結果。我們變動所運用的碼字數目 M ，分別設爲16、64與256，來觀測其效應。對於純雜訊的估測值 $\{\tilde{n}[p], 1 \le p \le P\}$ ，我們是以每一段測試語音的前10個音框作爲純雜訊音框的代表，即 $P = 10$ 。以下，表二、表三與表四分別爲新的碼簿建立程序所得之C-CMS、C-CMVN與C-HEQ在不同碼簿數 M 之下所得的平均辨識率（20dB、15dB、10dB、5dB與0dB五種訊雜比下的辨識率平均），AR與RR分別爲相較於基礎實驗結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。在這些表中加"*"標記者(C-CMS*或C-CMVN*)，則爲原始碼簿建立程序[5-7]所對應之C-CMS或C-CMVN法，而U-CMS、U-CMVN與U-HEQ分別爲整段式CMS、CMVN與HEQ。附帶一提的是，由於原始碼簿特徵正規化法的文獻[5-7]裡，只提及C-CMS與C-CMVN，並未介紹C-HEQ，因此在表四中，我們

只將新的C-HEQ與整段式HEQ(U-HEQ)的效能作比較。

表一、實驗所用之Aurora-2語料庫相關資訊

<table>
<tr><td colspan="3">AURORA2 語音資料庫</td></tr>
<tr><td>取樣頻率</td><td colspan="2">8kHz</td></tr>
<tr><td>語音內容</td><td colspan="2">英文數字 0～9(zero, one, two, three, four, five, six, seven, eight, nine, oh)，共 11 個音。</td></tr>
<tr><td>語音長度</td><td colspan="2">每一段語音包含不超過七個的英文數字</td></tr>
<tr><td>訓練語料</td><td colspan="2">句數：8440 句
摺積性雜訊：G712 通道；加成性雜訊：無加成性雜訊</td></tr>
<tr><td rowspan="2">測試語料</td><td>A 組雜訊環境</td><td>B 組雜訊環境</td></tr>
<tr><td>句數：28028 句
摺積性雜訊：G712 通道
加成性雜訊：
地下鐵雜訊(subway)
人的嘈雜聲雜訊(babble)
汽車雜訊(car)
展覽館雜訊(exhibition)
雜訊強度(signal-to-noise ratio, SNR)：clean、20dB、15dB、10dB、5dB、0dB</td><td>句數：28028 句
摺積性雜訊：G712 通道
加成性雜訊：
餐廳雜訊(restaurant)
街道雜訊(street)
機場雜訊(airport)
火車站雜訊(train station)
雜訊強度(signal-to-noise ratio, SNR)：clean、20dB、15dB、10dB、5dB、0dB</td></tr>
</table>

從這三個表格的結果，我們可觀察到下列幾點：

①就 CMS 法而言，原始之 C-CMS(C-CMS*)相對於基礎實驗結果進步較小(如在 N =256 下，在 Set A 下提升了 6.00%，在 Set B 下提升了 7.41%)，其效果甚至比整段式 CMS(U-CMS)來的差，然而，我們所提出的新 C-CMS，則帶來顯著的進步(如在 N =256 下，在 Set A 下提升了 9.54%，在 Set B 下提升了 13.70%)，由此證實，我們所用的新的碼簿建構程序確實能有效提升 C-CMS 的效果，而且其效果並不會隨著碼字數目的大小，而有明顯的變化。其效果在 Set A 下優於 U-CMS，在 Set B 下則略遜於 U-CMS，這可能原因在於，C-CMS 使用一段語音前幾個音框作雜訊估測，這在 Set B 此非穩定(non-stationary)雜訊環境中是比較不精確的。

②就 CMVN 法而言，原始之 C-CMVN(即 C-CMVN*)相對於基礎實驗結果雖已有了不錯的辨識率提升(如在 M =256 下，在 Set A 下提升了 14.75%，在 Set B 下提升了 18.46%)，但是相較於整段式 CMVN(U-CMVN)而言，在 M =16 與 M =64 下，其效果都比 U-CMVN 還要差，然而，我們所提出之新的 C-CMVN，則有明顯的進步，無論在 M =16、 M =64 或 M =256 下，其效果都比原始的 C-CMVN 還要好，且幾乎都優於 U-CMVN(僅在 M =16 時，Set B 之平均辨識率略遜於 U-CMVN)，由此證實，我們所用的新的碼簿建構程序確實能有效提升 C-CMVN 的效果，而且其效果並不會隨著碼字的大小，而有明顯的變化。

③就 HEQ 法而言，C-HEQ 同樣也能有效提昇辨識率，但無論在 A 組雜訊環境下或

B 組雜訊環境下，其平均辨識率都比 U-HEQ 來得差，我們推測其原因可能在於，C-HEQ 在純雜訊的估測上，是以每一段測試語音的前幾個音框作爲純雜訊音框的代表，因而造成純雜訊資訊不足，導致所得的雜訊語音碼簿不夠精準，最終造成 C-HEQ 辨識率比 U-HEQ 還要差的結果。

表二、U-CMS、原始C-CMS(C-CMS*)、與新C-CMS的平均辨識率(%)

Method	Set A	Set B	average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-CMS	79.37	82.47	80.92	11.07	36.71
C-CMS*(*M*=16)	74.21	70.81	72.51	2.65	8.81
C-CMS*(*M* =64)	74.03	70.74	72.39	2.53	8.39
C-CMS*(*M* =256)	77.92	75.20	76.56	6.71	22.24
C-CMS(*M* =16)	79.04	79.56	79.30	9.45	31.33
C-CMS(*M* =64)	80.79	80.19	80.49	10.64	35.28
C-CMS(*M*=256)	81.46	81.49	81.48	11.62	38.55

表三、U-CMVN、原始C-CMVN(C-CMVN*)、與新C-CMVN的平均辨識率

Method	Set A	Set B	average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-CMVN	85.03	85.56	85.30	15.44	51.22
C-CMVN*(*M* =16)	84.44	82.40	83.42	13.57	45.00
C-CMVN*(*M*=64)	84.13	81.53	82.83	12.98	43.04
C-CMVN*(*M*=256)	86.67	86.25	86.46	16.61	55.08
C-CMVN(*M*=16)	85.41	85.21	85.31	15.46	51.27
C-CMVN(*M*=64)	86.92	86.81	86.87	17.01	56.43
C-CMVN(*M*=256)	87.10	87.32	87.21	17.36	57.57

表四、U-HEQ與新C-HEQ的平均辨識率

Method	Set A	Set B	average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-HEQ	87.00	88.33	87.67	17.81	59.08
C- HEQ(*M*=16)	84.03	84.46	84.25	14.39	47.74
C- HEQ(*M*=64)	86.32	85.90	86.11	16.26	53.92
C-HEQ(*M*=256)	86.22	86.07	86.15	16.29	54.04

2.組合式特徵參數正規化法之辨識結果

在這一節中，我們將介紹本論文所提出之組合式(associative)特徵參數正規化技術之辨識結果，這三種技術分別爲組合式倒頻譜平均消去法(associative CMS, A-CMS)、組合式倒頻譜平均值與變異數正規化法(associative CMVN, A-CMVN)與組合式統計圖等化法(associative histogram equalization, A-HEQ)。在 A-CMS、A-CMVN 與 A-HEQ 三種正

規化技術中，由於在不同的碼字數目 N 下，產生最佳辨識率的 α 值(如式(4.5)、(4-7)與(4.8)中所示)或 β 值(如式(4.10)中所示)不盡相同，因此在以下的實驗辨識結果中，我們只呈現在不同的 N 值時，所產生最佳平均辨識率之 α 值或 β 值之結果。

首先，表五爲 A-CMS 在碼字數目 N 分別爲 16、64 與 256 下，所得到的最佳辨識結果，爲了比較起見，我們也將表二中的基本實驗、C-CMS(M =256)與 U-CMS 的平均辨識率列在表中。從此表中，我們可以觀察到以下幾種情形：

①組合式倒頻譜平均消去法(A-CMS)相較於基本實驗而言，無論在碼字數 M =16、64 與 256 下，其平均辨識率皆有大幅的進步，三者在 A 組雜訊環境下分別有 11.86%、11.30%與 10.98%的辨識率提升，在 B 組雜訊環境下分別有 17.76%、16.82%與 16.83%的辨識率提升，由此可看出 A-CMS 具有不錯之特徵強健化效果。

② A-CMS 在各種不同的碼字數 N 之下，其平均辨識率皆比 C-CMS 與 U-CMS 來得好，其中在 N =16 時能有最佳的效果，在 A 組雜訊環境與 B 組雜訊環境下之平均辨識率分別爲 83.78%與 85.55%，相較於 C-CMS 取 M =256 所得之最佳辨識率，A-CMS 在 A 組雜訊環境與 B 組雜訊環境下分別進步了 2.32%和 4.06%，這些進步都顯示了 A-CMS 優於 C-CMS。最後相較於 U-CMS，A-CMS 在 A 組雜訊環境與 B 組雜訊環境下其辨識率分別可以提升 4.41%和 3.08%。因此由實驗數據中可以證明，相對於 C-CMS 與 U-CMS 而言，A-CMS 都可以得到較好的辨識結果，這可能是因爲 A-CMS 同時整合了 C-CMS 與 U-CMS 所用的統計資訊，所以它更能有效改善語音在雜訊下的強健性。

表五、U-CMS、新C-CMS與A-CMS的平均辨識率

Method	Set A	Set B	average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-CMS	79.37	82.47	80.92	11.07	36.71
C-CMS(M=256)	81.46	81.49	81.48	11.62	38.55
A-CMS(M=16, α=0.5)	83.78	85.55	84.67	14.81	49.13
A-CMS(M=64, α=0.6)	83.22	84.61	83.92	14.06	46.64
A-CMS(M=256, α=0.6)	82.90	84.62	83.76	13.91	46.13

接著，表六爲A-CMVN在碼字數目 M 分別爲16、64與256下，所得到的最佳辨識結果，在表中，我們也列出原表三中的基本實驗、C-CMVN(M =256)與U-CMVN的平均辨識率以供比較。從此表中，我們可以觀察到以下幾種情形：

①組合式倒頻譜平均值與變異數正規化法(A-CMVN)在碼字數目 M =16、64 與 256 下，相較於基本實驗而言，其平均辨識率皆有大幅的改進，這三種 A-CMVN 在 A 組雜訊環境下分別有 16.19%、16.08%與 15.43%的辨識率提升，在 B 組雜訊環境下分別有 21.18%、20.77%與 20.26%的辨識率提升，由此可以發現 A-CMVN 確實能降低加成性雜訊對語音特徵的干擾，而提升辨識精確度。

② A-CMVN在各種碼字數N的情形下，其平均辨識率皆比C-CMVN、U-CMVN來得好，其中以 N =16時表現爲最佳，在A組雜訊環境與B組雜訊環境下之平均辨識率分別爲88.11%和88.97%，相較於C-CMVN取 M =256所得之最佳辨識率，A-CMVN在A組雜訊環境與B組雜訊環境則分別進步了1.01%與1.65%，這些進步都顯示了A-CMVN優於C-CMVN；而跟U-CMVN比較時，A-CMVN在A組雜訊環境與B組雜訊環境下，其辨識率分別可以提升3.08%和3.41%，其相對改善率分別爲20.55%與23.62%。類似之前的

A-CMS，A-CMVN同時整合了C-CMVN與U-CMVN所用的統計資訊，因此我們預期它具備了更佳的語音特徵強健化的效果，實驗數據也確實驗證了A-CMVN的表現明顯優於C-CMVN與U-CMVN。

表六、U-CMVN、新C-CMVN與A-CMVN的平均辨識率

Method	Set A	Set B	Average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-CMVN	85.03	85.56	85.30	15.44	51.22
C-CMVN(M=256)	87.10	87.32	87.21	17.36	57.57
A-CMVN(M=16, α=0.7)	88.11	88.97	88.54	18.69	61.98
A-CMVN(M=64, α=0.8)	88.00	88.56	88.28	18.43	61.12
A-CMVN(M=256, α=0.8)	87.35	88.05	87.70	17.85	59.20

最後，表七爲A-HEQ在碼字數目 M 分別爲16、64與256下，所得到的最佳辨識結果，爲了比較起見，我們也將表四中的基本實驗、C-HEQ(M=256)與U-HEQ的平均辨識率列在表中。從此表中，我們可以觀察到以下幾種情形：

①對於組合式統計圖等化法(A-HEQ)而言，無論在碼字數 M =16、64與256下，其平均辨識率相較於基本實驗而言，都有大幅的改進，三者在A組雜訊環境下分別有18.15%、17.28%與15.76%的辨識率提升，在B組雜訊環境下分別有23.08%、22.36%與21.10%的辨識率提升，顯示了A-HEQ在語音特徵強健性的效能，且相較於之前所述的兩種組合式特徵正規化法A-CMS與A-CMVN，A-HEQ的表現更爲優異。

②A-HEQ在各種碼字數M的情形下，其平均辨識率皆比C-HEQ與U-HEQ來得好，其中以M=16所得的平均辨識率爲最佳，在A組雜訊環境與B組雜訊環境下之辨識率分別爲90.07%和90.87%，相較於C-HEQ取 M =256所得之最佳辨識率，A-HEQ在A組雜訊環境與B組雜訊環境下其辨識率則分別進步了3.85%與4.80%，這些進步都顯示了A-HEQ優於C-HEQ；而跟U-HEQ比較時，A-HEQ在A組雜訊環境與B組雜訊環境下其辨識率分別提升了3.07%與2.54%，其相對改善率分別爲23.62%與21.76%。類似之前的結果，這裡我們再次驗證了組合式的方法優於碼簿式與整段式的方法，即A-HEQ比C-HEQ與U-HEQ更能提升雜訊環境下語音辨識的精確度。

表七、U-HEQ、新C-HEQ與A-HEQ的平均辨識率

Method	Set A	Set B	Average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-HEQ	87.00	88.33	87.67	17.81	59.08
C-HEQ(M=256)	86.22	86.07	86.15	16.29	54.04
A-HEQ(M=16, β=0.9)	90.07	90.87	90.47	20.62	68.39
A-HEQ(M=64, β=0.9)	89.20	90.15	89.68	19.82	65.75
A-HEQ(M=256, β=1)	87.68	88.89	88.29	18.43	61.14

六、結論與未來展望

在本論文中，我們主要討論的特徵參數正規化技術，分別爲倒頻譜平均消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)與倒頻譜統計圖等化法(HEQ)，這三種技術皆須使用到特徵的統計量。傳統上，這些統計量是經由一整段的語音特徵估測而得。因此，其對應的技術，我們統稱爲整段式(utterance-based)特徵參數正規化技術。在

近年來，本實驗室發展了碼簿式(codebook-based)特徵參數正規化技術，分別爲C-CMS與C-CMVN。顧名思義，在這些方法中，所使用的特徵統計量是由碼簿計算而得，實驗證實這些碼簿式特徵參數正規化技術其表現大致上皆優於整段式特徵參數正規化技術。然而我們發現，它們仍然有進一步的改善空間。因此，本論文中我們提出了一套改良式的碼簿建立程序，相對於原程序的不同之處，在於我們應用了語音偵測技術處理乾淨語音訊號，然後利用純語音區段的語音特徵來訓練碼字；此外，這些碼字根據其涵蓋的特徵數目賦予不同的權重(weight)，此改良法在第三章有詳細的說明。

除了提出上述改良式的碼簿建立程序之外，本論文另一重點在於，我們提出了一系列組合式(associative)特徵參數正規化技術，分別爲A-CMS、A-CMVN與A-HEQ，這些技術中，我們整合了整段式技術與碼簿式技術所用的特徵統計資訊，用此整合後之統計量來執行CMS，CMVN或HEQ，其詳述於第四章中，這樣的技術可以有效地補償碼簿式技術中，純雜訊資訊不足的缺點，第五章中的實驗結果證實，組合式的特徵參數正規化技術比整段式與碼簿式特徵參數正規化技術，均能更明顯地提升辨識精確度。

雖然組合式特徵參數正規化技術效果十分顯著，但其最佳表現有賴於某些自由參數(即式(4.5)、式(4.8)中的α及式(4.10)中的β)的手動調整來整合碼簿式與整段式之統計資訊，因此在未來的發展上，我們希望能自動地求取出最佳的α與β等參數值，來對兩方的統計資訊作更精確的整合，同時，在建構雜訊語音碼簿的程序上，我們也希望能參考許多雜訊估測的方法，更精確測得一段語音中純雜訊的統計特性，期待更有效地提升碼簿式特徵參數正規化技術的效能。

參考文獻

[1] Chung-fu Tai and Jeih-weih Hung, "Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments", 2006 International Conference on Spoken Language Processing (Interspeech 2006—ICSLP)

[2] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on Acoustics, Speech and Signal Processing, 1981

[3] S. Tiberewala and H. Hermansky, "Multiband and Adaptation Approaches to Robust Speech Recognition", 1997 European Conference on Speech Communication and Technology (Eurospeech 1997)

[4] A. Torre, J. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-Linear Transformations of the Feature Space for Robust Speech Recognition", 2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)

[5] Tsung-hsueh Hsieh, "Feature Statistics Compensation for Robust Speech Recognition in Additive Noise Environments" , M.S. thesis, National Chi Nan University, Taiwan, 2007

[6] Tsung-hsueh Hsieh and Jeih-weih Hung, "Speech Feature Compensation Based on Pseudo Stereo Codebooks for Robust Speech Recognition in Additive Noise Environments", 2007 European Conference on Speech Communication and Technology (Interspeech 2007—Eurospeech)

[7] Jeih-weih Hung, "Cepstral Statistics Compensation and Normalization Using Online Pseudo Stereo Codebooks for Robust Speech Recognition in Additive Noise Environments", IEICE Transactions on Information and Systems, 2008

[8] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," Proceedings of ISCA IIWR ASR2000, Paris, France, 2000

[9] ITU recommendation G.712, "Transmission Performance Characteristics of Pulse Code Modulation Channels," Nov. 1996

[10] http://htk.eng.cam.ac.uk/