International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而治後世聖人易之以書契百官以治萬民以察說文敘曰蓋文字者經藝之本宣教明化之始前人所以垂後後人所以識古故曰本立而道生知天下之至賾而不可亂也教化既萌文心雕龍則謂人之立言因字而生句積句而成章積章而成篇篇之彪炳

Vol.27 No.2 December 2022 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 27, no. 2, December 2022

Contents

Papers

The Association for Computational Linguistics and Chinese Language Processing

Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 1-30

Aligning Sentences in a Paragraph-Paraphrased Corpus with New Embedding-based Similarity Measures

Aleksandra Smolka[*], Hsin-Min Wang[+],

Jason S. Chang[#], and Keh-Yih Su[+]

Abstract

To better understand and utilize lexical and syntactic mapping between various language expressions, it is often first necessary to perform sentence alignment on the provided data. Up until now, the character trigram overlapping ratio was considered to be the best similarity measure on the text simplification corpus. In this paper, we aim to show that a newer embedding-based similarity metric will be preferable to the traditional SOTA metric on the paragraph-paraphrased corpus. We report a series of experiments designed to compare different alignment search strategies as well as various embedding- and non-embedding-based sentence similarity metrics in the paraphrased sentence alignment task. Additionally, we explore the problem of aligning and extracting sentences with imposed restrictions, such as controlling sentence complexity. For evaluation, we use paragraph pairs sampled from the Webis-CPC-11 corpus containing paraphrased paragraphs. Our results indicate that modern embedding-based metrics such as those utilizing SentenceBERT or BERTScore significantly outperform the character trigram overlapping ratio in the sentence alignment task in the paragraph-paraphrased corpus.

Keywords: Sentence Alignment, Sentence Similarity, Sentence Embedding, Paragraph-paraphrased Corpus

[*] Social Networks and Human Centered Computing, Taiwan International Graduate Program
Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: aleksandra.smolka@hotmail.com

[+] Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: {whm, kysu}@iis.sinica.edu.tw

[#] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
E-mail: jason@nlplab.cc

1. Introduction

Monolingual text matching is necessary for many downstream applications, such as Paraphrase Identification and Extraction (Qiu *et al.*, 2006), Question Answering (Weiss *et al.*, 2021), Natural Language Inference (MacCartney & Manning, 2008), and Text Generation (Barzilay & McKeown, 2005). Take the QA task as an example, identifying the text fragments that match the given question within the associated passage is often required for locating the desired answer.

However, modern neural network (NN) approaches to text matching often suffer from certain limitations when two sequences contain considerably different lexicons or diverse grammatical structures (McCoy *et al.*, 2019). For example, when the verb "*decide*" in the sentence "*They* ***decided*** *to go*" is nominalized to the noun "*decision*" in its paraphrase "*They* ***made a decision*** *to go*", the popular word embedding similarity approach might fail as the embedding-vectors of "*decide*" and "*decision*" are quite different[1]. Another example is a pair of sentences "*A cat is chasing a dog.*" and "*A dog is chasing a cat.*", which contain the same set of lexicons and syntactic structure but with opposite meanings.

Furthermore, the NN approaches frequently fail when the matching involves multi-word expressions, or when expressions require compositionality handling (Blevins *et al.*, 2018; Hupkes *et al.*, 2020; Zhou *et al.*, 2020). For example, it is difficult to match expressions "*put off*" and "*procrastinate*" using basic word embeddings, as the real meaning of the idiom "*put off*" is not the sum of the meanings of its tokens.

We found that the limitations of NN models in text matching could be greatly alleviated by utilizing lexico-syntactic paraphrasing patterns such as $[_{VP}[_{VBN}[see]_{NP}[X1]]] \rightarrow [_{S}[_{NP}[X1]_{VP}[_{VBD}[be]_{VP}[observe]]]$, which denotes the conversion from active to passive voice for the phrase pair "*see the lion*" and "*the lion is observed*". Since some key lexicons are involved in the pattern, it would be difficult to exhaustively list such patterns by a human. It is preferable to automatically extract them from a large paraphrase corpus.

To collect such lexico-syntactic patterns, a high-quality paraphrased sentence pair dataset is essential. Unfortunately, current *sentence-aligned* paraphrase datasets (e.g., MRPC (Dolan & Brockett, 2005), PPDB (Ganitkevitch *et al.*, 2013), and QQP (Aghaebrahimian, 2017)) are too trivial for this task, as they mainly contain lexical paraphrases that could be easily handled by a NN. On the other hand, some *paragraph-aligned* paraphrase corpora, containing different human translations from the same source text, fit our needs well. To utilize those paragraph-

[1] The nearest semantic associates of the verb *decide* based on the cosine similarity between the word2vec vectors (trained on English Wikipedia) are those verbs such as: *choose* (0.64), *opt* (0.62), *persuade* (0.61), *want* (0.58), *refuse* (0.57), *insist* (0.56). However, the noun *decision* only has a similarity score 0.512, which means that its similarity to the verb *decide* is even less than that between *decide* and its quasi-antonymous *refuse*.

aligned paraphrase corpora, monolingual sentence alignment is the first step in retrieving the desired patterns.

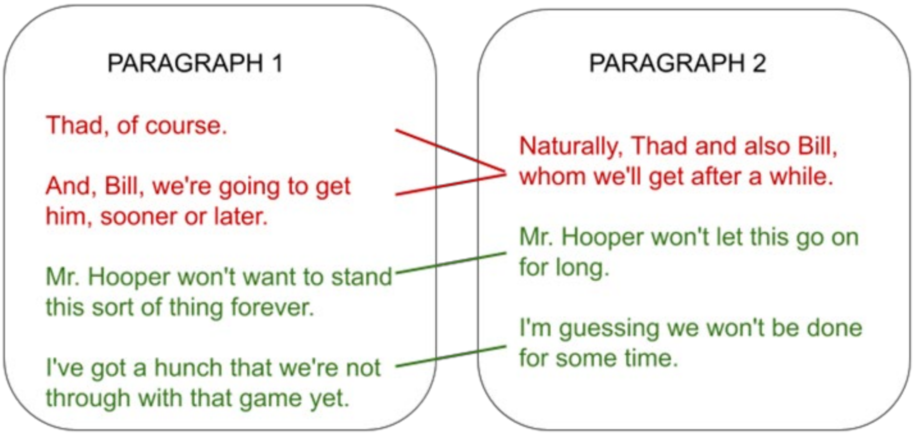

PARAGRAPH 1
Thad, of course.
And, Bill, we're going to get him, sooner or later.
Mr. Hooper won't want to stand this sort of thing forever.
I've got a hunch that we're not through with that game yet.
PARAGRAPH 2
Naturally, Thad and also Bill, whom we'll get after a while.
Mr. Hooper won't let this go on for long.
I'm guessing we won't be done for some time.

Figure 1. Sentence alignment for extracting paraphrased sentence pairs. Sentence pairs in green are those we want to extract; sentences in red are in multi-to-one relation and do not constitute sentential paraphrases. Figure adopted from Smolka et al. (2022).

Figure 1 shows how a correct sentence alignment could help extract paraphrased sentence pairs from longer paraphrased texts. Unless we correctly identify which sentences are in 1-to-1 relationships (green in the figure), we cannot correctly identify the desired paraphrased pattern.

Monolingual sentence alignment approaches could be classified into two categories: *model-based* approaches (e.g., Jiang *et al.*, 2020), which adopt specific models to encode the input sentences and perform alignment, and *model-agnostic* approaches (Štajner *et al.*, 2018), which can be directly applied to the selected dataset, without the necessity of training a neural model in advance. In our work, we focus on model-agnostic approaches, as they do not require additional labeled data to train the model.

The downside of previous model-agnostic approaches (Štajner *et al.*, 2017; 2018) is that they only test the early word2vec word embeddings, and do not explore those more advanced NN approaches such as Sentence-BERT (Reimers & Gurevych, 2019) and BERTScore (Zhang *et al.*, 2020). Also, they are mainly evaluated on Text Simplification (TS) datasets, which are different from our paraphrasing datasets.

In the TS dataset, the original and the simplified text often share a considerable number of keywords, which remain unchanged and are rarely substituted with synonyms. However, this property does not hold in our paraphrasing corpus, as its paraphrasing expressions usually possess diverse syntactic structures with many different lexical items.

Therefore, we suspect that the character trigram overlapping ratio, reported as the best for monolingual sentence alignment in previous works (Štajner *et al.*, 2017; 2018), would not perform best on our data. Since our paraphrasing corpus contains considerably different lexicons and word order, the string-based method such as character ngram similarity would lose its edge. Previously reported text similarity measures thus should be re-evaluated for our task, and more advanced NN approaches should be explored.

In this work, we not only compare various previously reported text similarity measures on a paraphrased paragraph corpus but also additionally test some new measures based on the most recent NN sentence embedding methods. We utilize those above measures with two sentence alignment approaches: simple greedy match (e.g., Štajner *et al.* 2018) and sequence match (Gale & Church, 1993; Barzilay & McKeown, 2001). We conduct the evaluation on a manually annotated sentence-aligned dataset with 400 paraphrased paragraph pairs randomly sampled from the multiple translation corpus Webis-CPC-11 (Burrows *et al.*, 2013).

Our contributions include:

(1) To the best of our knowledge, we present the first study on aligning sentences on a paragraph-paraphrased corpus;

(2) We show that character trigram similarity is not the best measure for aligning paraphrasing corpora. Instead, BERT-based embedding methods achieve significantly better results even without fine-tuning on the target dataset;

(3) We test several NN-related sentence similarity measures (other than word2vec) that have not been evaluated before for model-agnostic monolingual sentence alignment;

(4) We confirm and expand the observation of Choi *et al*., (2021), showing that [CLS] token representation is not necessarily superior to averaging individual word vectors for sentence representation while aligning paraphrased text under BERT.

(5) We compare the sentence alignment methods when an additional sentence length limitation is imposed on the data.

This publication is an extension of our previous conference paper on the same topic (Smolka *et al.*, 2022). In comparison to our conference publication, we have added a new data collection method for composing a dataset with sentence length limitation and introduced a new series of experiments performed on this new data (Section 3.4.2). We also extend the previous comparison of different methods of obtaining sentence representations using the BERT model (Section 3.4.3), and add a new discussion section to report our observations (Section 5). Finally, we extend some of the previously existing sections by additionally illustrating our search mechanisms (Section 2.1, Figure 2), showing an example of a non-paraphrased paragraph pair (Section 3.5, Figure 5), and a new error example in the error analysis (Section 4, Table 12).

2. Sentence Alignment Procedure

The proposed sentence alignment procedure is based on two basic elements, which we combine to test different experimental configurations. Those elements include: (1) search mechanism, which specifies the method used to search the sentence pairs that possess similar meaning; (2) similarity measure, which defines the method of calculating the similarity value among two given sentences (to be used during the search procedure).

2.1 Search Mechanisms

We implement two search mechanisms for aligning sentences among two paraphrased paragraphs: (1) *Directional Best Match*, which aligns each sentence in the paragraph separately (it also has two variations: *uni-directional*, which matches sentences from the first paragraph to the second one only, and *bi-directional*, which matches sentences in both directions), and (2) *Sequence Match*, which looks for the best alignment scheme for a paragraph as a whole. Figure 2 schematically illustrates the difference in how the sentence pairs are formed in the different search mechanism approaches, which we describe in the next two subsections.

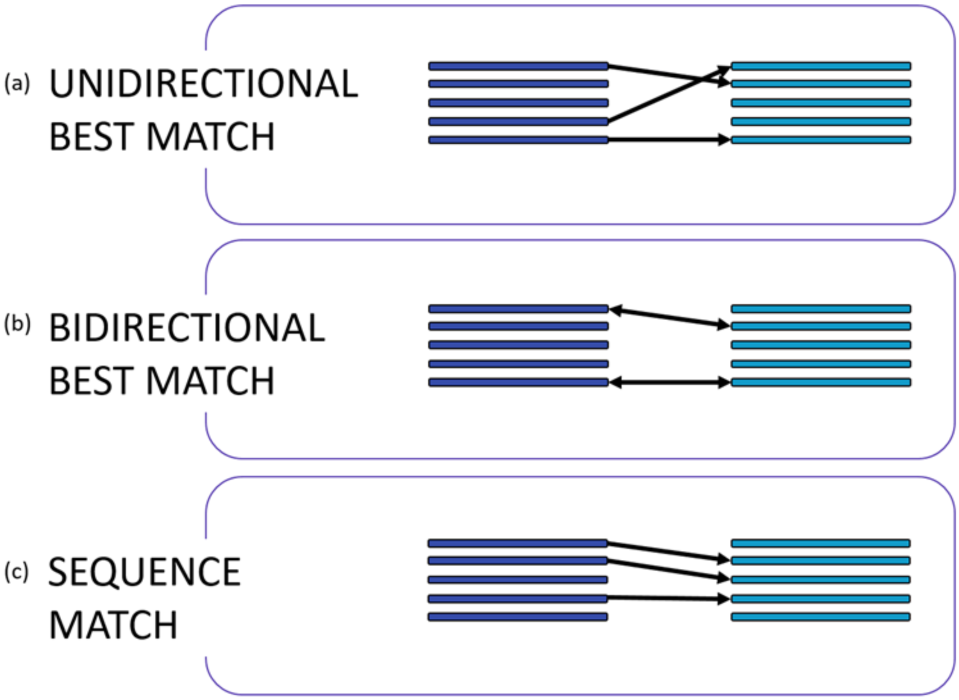


Figure 2. Schematic comparison of the different search mechanisms, illustrating the direction in which sentences are paired. (a) Uni-directional Best Match; (b) Bi-directional Best Match; (c) Sequence Match.

2.1.1 Directional Best Match

Directional Best Match is a simple greedy approach that relies on local judgments to create the alignments. This approach assumes that information such as adjacency and dependency information within sentences is negligible during matching. In our implementation, we follow the adopted SOTA approach (Štajner *et al.*, 2018). However, Štajner *et al.* (2018) only experimented with a uni-directional approach, which maps the original passages to the corresponding simplified passages. We believe that the bi-directional approach would be better applicable to our data since it is symmetric, unlike the dataset used by Štajner *et al.* (2018). Therefore, we additionally extend the Best Match method to a bi-directional approach.

Regardless of the above variation, we first calculate the associated similarity for each sentence pair that can be formed between the two given input paragraphs. Then, for each sentence in one paragraph, we select the sentence in another paragraph that has the highest similarity measure obtained above. For the uni-directional version, we directly take those pairs as the final alignments.

The bi-directional version follows the same steps, but we additionally repeat them in the opposite direction, i.e., matching sentences from the second paragraph to the first one. The final aligned pairs are obtained by taking the intersection of the two sets of aligned sentence pairs.

2.1.2 Sequence Match

Our sequence match adopts the dynamic programming searching algorithm to look for the best alignment path (among the two given paragraphs). Our implementation follows the common approach described in previous works (Gale & Church, 1993; Barzilay & McKeown, 2001). In this method, each alignment type (e.g., one-to-one and one-to-two) is associated with a different weight indicating the type probability estimated from the development set. The weights are then combined with the above similarity measures to find the best alignment path for the whole paragraph.

2.2 Similarity Measures

The text similarity measures adopted in our experiments fall into two main categories: (a) unit-overlap-based approaches, in which the similarity measure is based on the overlapping ratio of either ngrams or tokens between the sentences; (b) sentence-vector-based approaches, in which a neural model is first used to convert each sentence into its corresponding embedding-vector, and then the cosine similarity between these two sentence embedding-vectors is taken as the sentence similarity.

2.2.1 Unit-Overlap-Based Sentence Similarity

We adopt two different overlapping ratios: (1) *Character ngram*, which is reported as the state-of-art on the text simplification corpus (Štajner, 2018), and (2) *token*, which is commonly used in sentence alignment tasks (e.g., Barzilay & McKeown, 2001).

Character Ngram

We follow Štajner *et al.* (2018) to calculate the ngram similarity based on the *Character Ngram Similarity* model with tf-idf weighting (adapted from McNamee & Mayfield (2004)). We experiment with five different ngram sizes (1 to 5) and use NGRAM to refer to this measure. We add Laplace smoothing to account for those unseen ngrams in the test set. The final similarity is calculated by taking cosine similarity (Štajner *et al.*, 2018).

Token

For calculating token-based sentence similarity, we use the following token overlap formula:

$$similarity_{token} = \frac{|tokens_1 \cap tokens_2|}{|tokens_1| + |tokens_2|} \qquad (1)$$

where $tokens_1$ is the set of tokens in the first sentence, $tokens_2$ is the set of tokens in the second sentence, and the function | | specifies the cardinality of the token set. We consider two different normalization mechanisms for comparing two tokens: (1) converting the strings into their associated lemmas before comparison (abbreviated as TOKENstring); (2) also taking synonyms as exactly matched lemmas during comparison (abbreviated as TOKENsyn). Token lemmas for each sentence are retrieved using an automatic tokenizer and lemmatizer (Qi *et al.*, 2020). Synonymic relationships are taken from WordNet (Fellbaum, 1998).

2.2.2 Sentence-Vector- Based Sentence Similarity

This category includes similarity measures that utilize cosine vector similarity in some forms: (1) *word-embedding* based, where we first look up the word embedding-vector for every token in each sentence from a pretrained model, and then combine them into their associated sentence embedding-vector by vector averaging (Putra & Tokunaga, 2017). Afterward, we calculate the similarity between the two obtained sentence embedding vectors. (2) *sentence-embedding* based, where we use a model, such as BERT (Devlin *et al.*, 2019) or Sentence-BERT (Reimers & Gurevych, 2019), to directly embed a sentence into its associated sentence-embedding. We then calculate the similarity between these two sentence embedding vectors. (3) *BERTScore* (Zhang *et al.*, 2020), which uses BERT to directly generate the similarity value between two sentences.

Word-embedding Similarity

For directly retrieving the token-associated embedding vector from a pretrained embedding lookup table, we test both word2vec (Mikolov *et al.*, 2013) and Glove (Pennington *et al.*, 2014) embeddings. Additionally, we also test contextualized word embeddings retrieved from BERT (Devlin *et al.*, 2019).

Moreover, while it is common to use the [CLS] token yielded by the BERT encoder to represent the whole encoded sentence, recent works note that this might not be the best solution for different downstream tasks (Choi *et al.*, 2021). We therefore additionally test the following approach: generate the sentence embedding via averaging the contextual word embeddings retrieved from the BERT model.

Regardless of the way of selecting word embedding, we combine the associated embedding vectors into the corresponding sentence representation by taking an average over them (Putra & Tokunaga, 2017). The sentence similarity is then calculated as the cosine similarity between the two sentence embedding vectors.

Among various types of word embeddings, only word2vec is tested by Štajner *et al.* (2018). However, it was not reported as the best one in their experiments (the best one is the character trigram in their task).

Sentence-embedding Similarity

Another way to generate the sentence-embedding is to adopt BERT to transform all its associated token-embeddings into it. We test two methods of obtaining sentence representation via BERT. First, we take the [CLS] token from the BERT to represent the whole sentence. Alternatively, we use Sentence-BERT (Reimers & Gurevych, 2019), which is an alternative method of obtaining sentence representation from BERT-type models, suggested as a better alternative for directly adopting [CLS] token embedding. We use Sentence-BERT to separately obtain a single embedding for each sentence in the pair. The sentence similarity is then calculated between two obtained sentence embedding vectors.

BERTScore

Last, we can directly generate the desired similarity value among two sentences by adopting the BERTScore (Zhang *et al.*, 2020) approach, which is originally developed as an automatic evaluation metric for comparing various text generation systems. This approach first uses BERT to obtain the word embeddings of all input tokens. The pairwise similarity is then calculated for each possible token pair. Afterward, for each token from the first input sequence (i.e., the sentence from the "*original*" paragraph), BERTScore finds its matching token in the second sequence (i.e., the sentence from the "*paraphrased*" paragraph) via greedy search. Last, it calculates both precision and recall based on the matching result.

As BERTScore is designed to evaluate the similarity between the ground truth and the generated text, we thought it should be also suitable for measuring the sentence similarity for our task. Typically, BERTScore will report precision, recall, and F1-score at the same time. We take each of these values to represent a specific sentence pair similarity measure; and we refer to them as BERTprec, BERTrec, and BERTf1, respectively.

2.3 Similarity Score Thresholding

Regardless of the selected combination of search mechanism and similarity measure, we additionally impose a similarity score thresholding on the aligned sentences. In the final stage of the alignment procedure, we filter out sentence pairs that have similarity values below the experimentally selected threshold. This helps us further improve the overall test-set results and allows for a precision-recall trade-off if desired.

3. Experiments

Figure 3 shows the operation flow adopted in the experiments. We first take a pair of paraphrased paragraphs as input, clean the text in each paragraph, and split it into individual sentences. Then, we use the sentence alignment module with the selected search mechanism and similarity measure to generate the desired sentence alignments. Those one-to-one sentence alignments are then extracted and output as the answer.

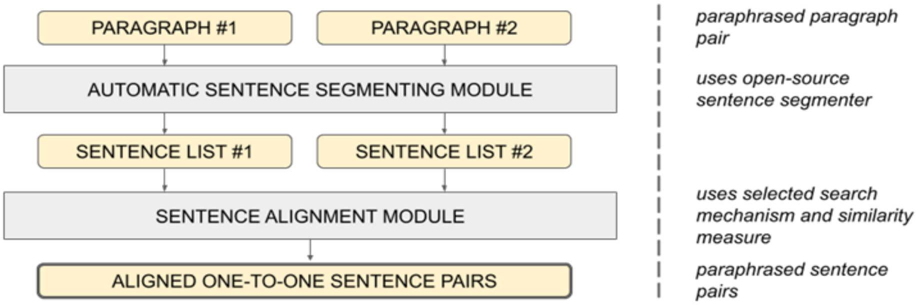

PARAGRAPH #1
PARAGRAPH #2
AUTOMATIC SENTENCE SEGMENTING MODULE
SENTENCE LIST #1
SENTENCE LIST #2
SENTENCE ALIGNMENT MODULE
ALIGNED ONE-TO-ONE SENTENCE PAIRS
paraphrased paragraph pair
uses open-source sentence segmenter
uses selected search mechanism and similarity measure
paraphrased sentence pairs

Figure 3. Operation flow for obtaining one-to-one sentence alignment within paraphrased paragraph pairs. Figure adopted from Smolka et al. (2022).

The following subsections give details of the experiment setting and results.

3.1 Dataset

We randomly sampled 400 paragraph pairs from the Webis-CPC-11 corpus (out of which 7 were found to be incorrectly marked as paraphrases, and removed from the evaluation data). The non-paraphrased pairs are excluded from the development and test data. However, we reserve them for additional experiments where we test methods for automatically detecting such undesired input from our data.

To evaluate the performance, we manually annotate the 400 paragraph pairs randomly sampled from the Webis-CPC-11 corpus. The annotation process consists of several stages: (1) Paragraph pre-processing, which is performed automatically and serves to clean the data and split each paragraph into its associated sentences; (2) Sentence alignment (marking both one-to-one and one-to-many alignment configurations), in which we manually match the sentences that have similar meanings.

After the paragraph pre-processing stage, the annotator receives two sets of sentences for each paragraph pair and is requested to align sentences between them (including both one-to-one and one-to-many mappings). The result of the manual annotation is a dataset in which each paraphrased paragraph pair is associated with the aligned sentence pairs between them. If the sample contains non-paraphrased paragraphs, the annotator is asked to simply mark them without adding alignment annotation.

As all tested similarity measures are model-agnostic, we do not require a training set. Therefore, we split all the aligned paragraph pairs (i.e., excluding those non-paraphrased pairs) into the development set and the test set with a 1:7 ratio. As a result, we end up with 48 paragraph pairs in the development set and 345 paragraph pairs in the test set. We use the development set for selecting hyper-parameters such as similarity cutting threshold and alignment type probabilities for the Gale-Church algorithm (Gale & Church, 1993).

Table 1. Dataset Statistics (without non-paraphrase cases). #Min-#Max specifies the range in paragraph range row. Also, 1-1 indicates the one-to-one mapping, 2-1 (1-2) indicates two-to-one and one-to-two mapping, and so on.

		all	dev	test
#input paragraphs		393	48	345
#input non-paraphrased pairs (dataset errors)		7	2	5
avg. paragraph length (#sentences)		2.3	2.4	2.3
avg. sentence length (#tokens)		20.9	19.3	21.1
paragraph range (# sentences)		1-7	1-6	1-7
% of alignment types	all	822 (100%)	87 (100%)	735 (100%)
	1-1 (ground truth)	**633 (77%)**	**67 (77%)**	**566 (77%)**
	2-1 (1-2)	132 (16%)	16 (18%)	118 (16%)
	2-2	8 (1%)	1 (1%)	7 (1%)
	Other (2-3,1-4,etc.)	49 (6%)	4 (4%)	45 (6%)

Table 1 gives the associated dataset statistics. Within them, 566 1-to-1 paraphrased sentence pairs (77% among all aligned passage pairs) exist in the test set. This set of 1-to-1 sentence pairs (i.e., sentential paraphrases) is the desired output in our task, and thus becomes the ground truth for our evaluation.

3.2 Pre-processing

Because the Webis-CPC dataset only contains un-segmented paragraphs, it must be first converted into a collection of sentences. We use an off-the-shelf sentence segmenter (Qi *et al.*, 2020) to split each paragraph into sentences. The output is thus two sets of sentences, one for each of the paragraphs.

3.3 Experimental Setting

For our baseline, we re-implement the SOTA approach proposed by Štajner *et al.* (2018), as there is no easily applicable code released by the authors. Therefore, we follow the descriptions in the original paper to implement the ngram character similarity. Our implementation has not been tested on the data adopted in the work of Štajner *et al.* (2018) because it lacks the annotations that are necessary for automatic evaluation. Furthermore, the original work introducing SOTA character trigram metrics used only human evaluation, which makes a direct comparison of our method with their results impossible.

When it comes to the pretrained models used for conducting the embedding-based similarity calculations, we select the models based on their open-source availability. For example, for getting the BERT word-averaging and [CLS]-token representation, we use the BERT-base model (Devlin *et al.*, 2019). When it comes to Sentence-BERT, three different pretrained models were tested, including *BERT-base* (Devlin *et al.*, 2019; abbreviated as SBERTbert), *ALBERT-mini* (Lan *et al.*, 2020; abbreviated as SBERTalbert), and *MiniLM* (Wang *et al.*, 2020; abbreviated as SBERTmini). The training data for those three SentenceBERT models varied and depended on the original open-source model released.[2] Among them, SBERTbert was trained with various Natural Language Inference data sets; SBERTalbert and SBERTmini were trained on various paraphrasing datasets.[3] Finally, the BERTScore open-source implementation uses ROBERTA-Large (Liu *et al.*, 2019).

3.4 Various Experiments

In our experiments, we test various combinations of the two alignment strategies with different similarity measures. We take precision, recall, and F1-score as the evaluation metrics. Moreover,

[2] https://huggingface.co/sentence-transformers

[3] The list of specific datasets used was not published by the open-source authors.

for each set of results, we apply the McNemar test (Dietterich, 1998) to check whether the performance improvement is statistically significant (with p≤0.05 as the significance test threshold).

In our experiments, we test similarity measures based on: **(1) Unit-Overlap-Ratio**, including character ngram overlap-ratio with *n* ranging from 1 to 5 (NGRAM), and token overlap-ratio calculated with either token strings (TOKENstring) or token synonyms (TOKENsyn); **(2) Sentence-Vector-Similarity**, including **(a) word-embedding-based** similarity measures calculated with word2vec (W2V), Glove (GLOVE) and BERTbase (BERTword) embeddings; **(b) sentence-embedding-based** similarity measures which consist of: (i) using [CLS] token yielded by BERTbase model (BERTcls), and (ii) Sentence-BERT embeddings with three different pretraining models (SBERTbert, SBERTalbert, and SBERTmini); **(c) BERTScore** with precision (BERTprec), recall (BERTrec), and F1-score (BERTf1).

3.4.1 Sentence Alignment Results on the Full Dataset

Tables 2-4 compare various similarity measures under the Best Match (Uni- and Bi-directional, separately) strategy and the Sequence Match strategy, respectively. For each measure, we only report the results with the best threshold value, which is selected on the development set based on the F1 value. The threshold for each specific similarity measure is different and is noted in the corresponding table. Measures that outperform the character trigram baseline in a significant manner are marked with the asterisk *.

Overall, comparing the best result of each approach, the sequence match approach (with the best F1-score equaling 88.8%) outperforms both best match approaches (the best F1-score of 85.1% is from the bi-directional mode). We conjecture that the sequence match performs the best as it additionally considers the adjacency and dependency information within sentences during matching.

Moreover, the Uni-directional Best Match approach performed the worst (only with 82.5% best F1) as expected. Since our data is symmetric, the matching results would be more reliable if the alignment is considered from both directions.

Furthermore, the best similarity measure varies under different search mechanisms. In the sequence match approach, three BERT-type measures (i.e., SBERTbert (88.8% F1), BERTrec (88.7% F1), and BERTf1 (88.7% F1)) significantly outperform the baseline. The SentenceBERT measure performs best, surpassing the character-trigram baseline method by 1.9% (88.8% vs. 86.9%) because it is trained to encode the overall sentence meaning, not the specific meaning of individual tokens, which fits our task well. Similarly, BERTScore also delivers good results because it is directly trained to measure the similarity between two sequences.

Table 2. Alignment results by adopting the uni-directional Best Match strategy on the full dataset. TH indicates the adopted threshold value. The asterisk * marks the measures that outperform NGRAM baseline (n=3) with p ≤ 0.05. Table adopted from Smolka et al. (2022).

measure	% on the test set			Best *TH*
	prec	rec	F1	
NGRAM(n=1)*	77.8	82.2	79.9	0.3
NGRAM(n=2)*	77.8	82.2	79.9	0.3
NGRAM(n=3)	79.9	72.5	76.1	0.3
NGRAM(n=4)*	77.8	82.2	79.9	0.3
NGRAM(n=5)*	77.8	82.2	79.9	0.3
TOKENstring*	83.7	73.1	78.1	0.2
TOKENsyn	77.1	71.5	74.2	0.1
W2V	79.7	74.5	77.0	0.8
GLOVE	73.5	81.2	77.1	0.95
BERTword*	78.5	**87.0**	**82.5**	0.75
BERTcls	81.9	67.9	74.3	0.9
SBERTbert	75.2	90.8	82.3	0.6
SBERTalbert	82.9	70.7	76.9	0.35
SBERTmini*	78.4	85.2	81.6	0.6
BERTprec*	86.5	72.9	79.1	0.9
BERTrec*	83.5	74.9	80.4	0.9
BERTf1*	**86.8**	74.9	80.4	0.9

On the other hand, in the bi-directional best match approach, the best result is again obtained by the Sentence-BERT measure (SBERTmini) with the best F1-score 85.1%, significantly outperforming the character ngram similarity measure at 82.7%. Also, both SBERTalbert and BERTf1 measures outperform the baseline with p<0.06. We believe that the above reasons given for the sequence match approach also apply here.

Table 3. Alignment results by adopting bi-directional Best Match strategy on full dataset. TH indicates the adopted threshold value. The asterisk * marks the measures that outperform NGRAM baseline (n=3) with p ≤ 0.05. Table adopted from Smolka et al. (2022).

measure	% on the test set			Best *TH*
	prec	rec	F1	
NGRAM(n=1)	80.5	81.8	81.1	0.3
NGRAM(n=2)	80.5	81.8	81.1	0.3
NGRAM(n=3)	78.9	87.0	82.7	0.1
NGRAM(n=4)	80.5	81.8	81.1	0.3
NGRAM(n=5)	80.5	81.8	81.1	0.3
TOKENstring	84.7	73.1	78.5	0.2
TOKENsyn	78.6	81.8	80.2	0.05
W2V	81.1	87.6	84.2	0.6
GLOVE	79.7	78.0	78.8	0.95
BERTword	82.3	86.4	84.3	0.75
BERTcls	**86.2**	66.5	75.1	0.9
SBERTbert	79.1	88.6	83.6	0.6
SBERTalbert	80.6	89.8	84.9	0.25
SBERTmini*	80.7	90.2	**85.1**	0.25
BERTprec	80.9	88.2	84.4	0.85
BERTrec	79.7	88.2	83.7	0.85
BERTf1	79.9	**90.8**	85.0	0.9

Last, in the uni-directional best match approach, several tested measures significantly outperform the baseline (76.1%), including BERTword (82.5%), SBERTbert (82.3%), SBERTmini (81.6%), BERTf1(80.4%), NGRAM with n≠3 (79.9%), BERTrec (79.7%), BERTprec (79.1%), and TOKENstring (78.1%). The measures that perform best in this search mechanism are again mostly those that encode the sentence as a whole, similar to other search mechanisms.

Table 4. Alignment results by adopting Sequence Match strategy on the full dataset. TH indicates the adopted threshold value. The asterisk * marks the measures that outperform NGRAM baseline (n=3) with p ≤ 0.05. Table adopted from Smolka et al. (2022).

measure	% on the test set			Best *TH*
	prec	rec	F1	
NGRAM(n=1)	89.1	83.4	86.1	0.2
NGRAM(n=2)	89.1	83.4	86.1	0.2
NGRAM(n=3)	89.7	84.2	86.9	0.1
NGRAM(n=4)	89.1	83.4	86.1	0.2
NGRAM(n=5)	89.1	83.4	86.1	0.2
TOKENstring	**92.7**	81.6	86.8	0.15
TOKENsyn	86.2	86.9	86.3	0
W2V	87.6	87.6	87.6	0.45
GLOVE	87.3	85.2	86.2	0.9
BERTword	91.5	82.2	86.6	0.75
BERTcls	92.3	81.4	86.5	0.85
SBERTbert*	89.8	**87.8**	**88.8**	0.6
SBERTalbert	91.1	85.8	88.3	0.25
SBERTmini	87.8	86.8	87.3	0.25
BERTprec	90.0	86.8	88.4	0.85
BERTrec*	89.9	87.6	88.7	0.85
BERTf1*	90.1	87.4	88.7	0.85

3.4.2 Alignment Results on Sentences with Limited Length

The above experiments are conducted without limiting the lengths of those input sentences. However, in our another study, we have found that it is difficult to extract appropriate lexico-syntactic patterns from sentences containing more than two clauses, as selecting the desired candidates will become much more confusing. As a result, the precision rate of extracting high-quality patterns would be lower. To ensure the quality of extracted templates, we thus conducted an additional set of experiments on those input sentences with limited length. Below, we first describe how to find out a reasonable way to filter out those sentences that would be too complicated/long for our purpose. Afterward, we repeat the above experiments on this new dataset to check if it would significantly change the alignment performance.

3.4.2.1 Finding the Appropriate Criterion to Filter out Long Sentences

To limit the degree of confusion in selecting the desired candidates, we would like to only use sentences with no more than two clauses to extract the desired templates. To automatically filter out those sentence pairs that might contain more than two clauses, we need to first find out a suitable criterion. For simplicity, we opt to use sentence length as the filtering criterion, because this value not only is highly correlated with the number of associated clauses but also could be easily measured.

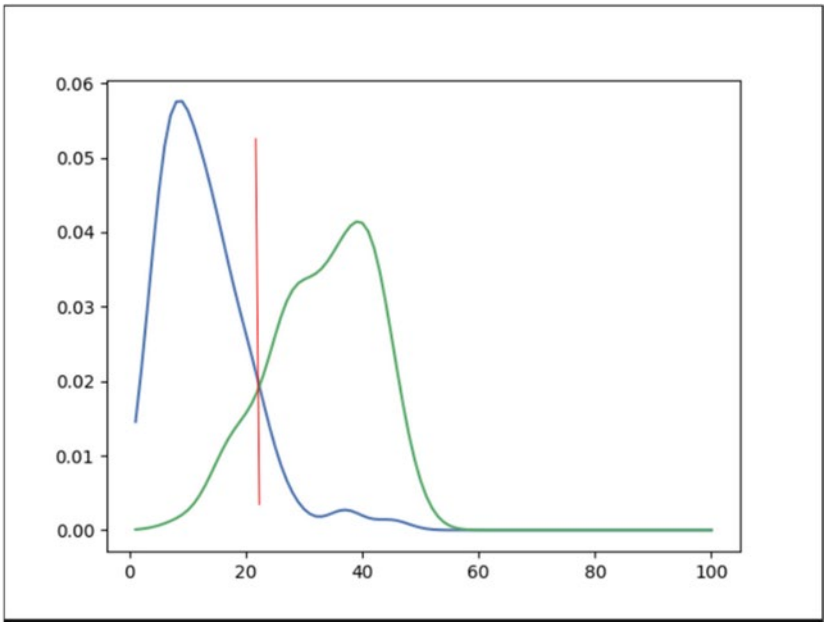


Figure 4. Finding the upper-limit sentence length from smoothed probability distributions (X-axis: sentence length in tokens). The blue curve is for the sentences with maximum two clauses, and the green curve is for the cases with more than two clauses. The red vertical line marks the intersection between the two distributions.

Figure 4 illustrates how we find the upper-limit sentence length. We first manually generate two smoothed probability distributions in Figure 4: The blue curve is for the sentences with two clauses at most (which we consider appropriate for our task), and the green curve is for the cases with more than two clauses (which we consider are too difficult). Those two smoothed probability distributions are constructed from 100 sentences in each group, which are randomly selected from the Webis-CPC-11 dataset and then manually checked to fit this target number. The smoothed probability distributions are calculated using kernel density estimation

(Rosenblatt, 1956). We then find the integer value that is closest to the intersection point between the two distributions (indicated by the red vertical line in Figure 4), which is 22. This value indicates that if the sentence has a length below it, it is more likely to belong to the "appropriate" category. On the contrary, a sentence is more likely to be too difficult for our purpose, if its length is above this value.

Table 5 gives the details of the newly constructed dataset. The main difference from the full dataset used in the previous experiments lies in the golden answers. In the new dataset, the benchmark consists of only 367 aligned sentence pairs that are shorter than 22 tokens (versus 633 sentence pairs in the original dataset, Table 1).

Table 5. Statistics for the dataset that considers sentence-length constraint. #Min-#Max specifies the range in "paragraph range" row. Also, 1-1 indicates the one-to-one mapping, 2-1 (1-2) indicates two-to-one and one-to-two mapping, and so on.

		all	dev	test
#input paragraphs		393	48	345
#input non-paraphrased pairs (dataset errors)		7	2	5
avg. paragraph length (#sentences)		2.3	2.4	2.3
avg. sentence length (#tokens)		20.9	19.3	21.1
paragraph range (# sentences)		1-7	1-6	1-7
% of alignment types	all	822 (100%)	87 (100%)	735 (100%)
	1-1 (all)	633 (77%)	67 (77%)	566 (77%)
	2-1 (1-2)	132 (16%)	16 (18%)	118 (16%)
	2-2	8 (1%)	1 (1%)	7 (1%)
	other (2-3,1-4,etc.)	49 (6%)	4 (4%)	45 (6%)
Evaluation Benchmark	**1-1 (<22 tokens, golden answers)**	**367 (45%)**	**50 (57%)**	**317 (43%)**

3.4.2.2 Experimental Results on Sentences with Limited Length

Tables 6-8 compare all similarity measures under the Best Match strategy (Uni- and Bi-directional, separately) and the Sequence Match strategy, respectively for the dataset containing only sentences shorter than 22 tokens. We follow the same scheme adopted in the previous experiments to report the new results.

Table 6. Alignment results on sentences shorter than 22 tokens for the uni-directional Best Match strategy. TH indicates the threshold value. The asterisk * marks the metrics that outperforms NGRAM baseline (n=3) with p ≤ 0.05.

measure	% on the test set			Best *TH*
	prec	rec	F1	
NGRAM(n=1)*	73.3	77.9	75.5	0.3
NGRAM(n=2)*	73.3	77.9	75.5	0.3
NGRAM(n=3)	74.4	65.9	69.6	0.3
NGRAM(n=4)*	73.3	77.9	75.5	0.3
NGRAM(n=5)*	73.3	77.9	75.5	0.3
TOKENstring*	77.7	70.3	73.8	0.2
TOKENsyn	71.3	65.9	68.5	0.1
W2V	74.6	65.9	70.0	0.8
GLOVE	65.7	74.4	69.8	0.95
BERTword*	73.9	83.9	78.6	0.75
BERTcls	78.2	66.9	72.1	0.9
SBERTbert*	70.3	**90.2**	79.0	0.6
SBERTalbert	76.9	70.3	73.5	0.35
SBERTmini*	74.2	85.2	**79.3**	0.35
BERTprec*	77.6	70.0	74.1	0.9
BERTrec*	81.5	73.8	77.5	0.9
BERTf1*	**80.9**	72.2	76.3	0.9

Overall, the performances (in terms of F1 scores) on those length-limited sentences are lower than that on the full dataset (Table 2-4). The drop in F1 score ranges from 2.4% (bi-directional Best Match; 85.1% vs. 82.7%) to 6.1% (Sequence Match; 88.8% vs. 82.7%). One reason for causing the drops is that it implicitly removes the simplest alignment cases after filtering out those longer sentences, where the whole paragraph just consists of one single sentence. Another reason is that shorter sentences are easier to be mistakenly linked because they have less distinctive tokens. Detailed explanation will be delayed to the discussion section (Section 5).

Table 7. Alignment results on sentences shorter than 22 tokens for the bi-directional Best Match strategy. TH indicates the threshold value. The asterisk * marks the metrics that outperforms NGRAM baseline (n=3) with p ≤ 0.05.

measure	% on the test set			Best *TH*
	prec	rec	F1	
NGRAM(n=1)	77.7	77.9	77.8	0.3
NGRAM(n=2)	77.7	77.9	77.8	0.3
NGRAM(n=3)	74.7	83.9	79.0	0.1
NGRAM(n=4)	77.7	77.9	77.8	0.3
NGRAM(n=5)	77.7	77.9	77.8	0.3
TOKENstring	79.4	70.3	74.6	0.2
TOKENsyn	73.6	76.3	74.9	0.05
W2V*	78.4	84.5	81.3	0.6
GLOVE	72.9	68.8	70.8	0.95
BERTword*	78.6	83.3	80.9	0.75
BERTcls	**83.7**	64.7	73.0	0.9
SBERTbert*	75.0	87.1	80.6	0.6
SBERTalbert*	77.1	**89.3**	**82.7**	0.25
SBERTmini*	76.0	89.0	82.0	0.25
BERTprec*	75.7	86.4	80.7	0.85
BERTrec*	76.2	82.0	81.3	0.85
BERTf1	81.8	72.2	76.7	0.9

Unlike in the experiments on the full dataset, two of the alignment strategies – Bi-directional Best Match and Sequence Match obtain the same F1 score (82.7%) with the SBERTalbert metric. This might indicate that the adjacency and dependency information used in Sequence Match (but not Best Match) is not as important for aligning sentences with limited length.

Table 8. Alignment results on sentences shorter than 22 tokens for the Sequence Match Best Match strategy. TH indicates the threshold value. The asterisk * marks the metrics that outperforms NGRAM baseline (n=3) with p ≤ 0.05.

measure	% on the test set			Best *TH*
	prec	rec	F1	
NGRAM(n=1)	76.5	82.3	79.3	0.2
NGRAM(n=2)	76.5	82.3	79.3	0.2
NGRAM(n=3)	74.7	83.9	79.0	0.1
NGRAM(n=4)	76.5	82.3	79.3	0.2
NGRAM(n=5)	76.5	82.3	79.3	0.2
TOKENstring	76.5	83.3	79.8	0.15
TOKENsyn	73.4	78.2	75.7	0
W2V*	78.2	84.9	81.4	0.45
GLOVE	72.8	72.6	72.7	0.9
BERTword*	**78.6**	83.3	80.9	0.75
BERTcls	77.5	78.2	75.7	0.85
SBERTbert*	75.0	87.1	80.6	0.6
SBERTalbert*	77.1	89.3	**82.7**	0.25
SBERTmini*	76.0	89.0	82.0	0.25
BERTprec*	75.7	86.4	80.7	0.85
BERTrec	74.9	84.5	79.4	0.85
BERTf1*	76.1	**90.5**	**82.7**	0.85

Furthermore, just as on the full dataset, the best similarity measure varies under different search mechanisms. In the sequence match approach, two BERT-type measures (i.e., all SentenceBERT variants with the best being BERTalbert (82.7% F1 score)), and two of BERTScore variants (i.e., BERTprec with 80.7% F1 score and BERTf1 with 82.7% F1 score) and word2vec metric (i.e., W2V, 81.4% F1 score) significantly outperform the baseline. The SentenceBERT and BERTScore measure performs best, surpassing the character-trigram baseline method by 3.7% (82.7% vs. 79.0%).

Similarly, in the bi-directional best match approach, the best result is again obtained by the SentenceBERT measure (i.e., SBERTalbert) with the best F1-score of 82.7%, significantly outperforming the character ngram similarity measure at 79.0%. This confirms the observation from previous experiments regarding the high suitability of sentence-embedding-based approaches in our task.

Last, in the uni-directional best match approach, several tested measures significantly outperform the baseline (69.6%), including SBERTmini (79.3%), SBERTbert (79.0%), BERTword (78.6%), BERTrec (77.5%), BERTf1(76.3%), NGRAM with n≠3 (75.5%), BERTprec (74.1%) and TOKENstring (73.8%). The measures that perform best in this search mechanism are again mostly those that encode the sentence as a whole, similar to other search mechanisms.

In comparison with the alignment results obtained from those sentences without length limitation, the F1-scores measures on length-limited sentences are lower (see the last item in Section 5). Although the performance of alignment of sentences with limited length is overall lower than on full data, we still prefer to impose the sentence length limitation, because it only slightly lowers the alignment performance but will offer considerable benefit while extracting the lexico-syntactic templates later.

3.4.3 Comparison of BERT Word-averaging and [CLS] Token Sentence Representation

Table 9. Comparison of results of BERT word-averaging and BERT [CLS] token-based similarity metrics on the full dataset. SM indicates Search Mechanism. The asterisk * indicates cases where the difference between two measures is statistically significant with p ≤ 0.05.

SM	measure	% on the test set		
		prec	rec	F1
Sequence Search	**BERTword**	91.5	82.2	**86.6**
	BERTcls	92.3	81.4	86.5
Best Match (uni)	**BERTword***	78.5	87.0	**82.5**
	BERTcls	81.9	67.9	74.3
Best Match (bi)	**BERTword***	82.3	86.4	**84.3**
	BERTcls	86.2	66.5	75.1

Comparing the performance of the methods using BERTword (i.e., word-averaging of BERT token embeddings) and the BERT [CLS] token, we observe that the BERTword achieves better performance regardless of the adopted search mechanism. Table 9 and Table 10 show how BERTword performs significantly better ($p<0.05$) than BERTcls regardless of the search mechanism for the dataset with sentence length constraint. The BERTword results are up to

6.5% higher, depending on the search mechanism (80.9% vs. 75.7% for sequence match; 80.9% vs. 73.0%, and 78.6% vs. 72.1% for bi- and uni-directional, respectively). For the full dataset, a noticeable difference can be observed for both versions of the Best Match approach with up to a 9.2% difference (84.3% vs. 75.1% and 82.5% vs. 74.3% for bi- and uni-directional, respectively). This is in line with the observation from Choi *et al.* (2021), who noted that interpreting the [CLS] token embedding as the sentence representation might be inferior to combining the individual sub-word embeddings obtained from BERT in some tasks.

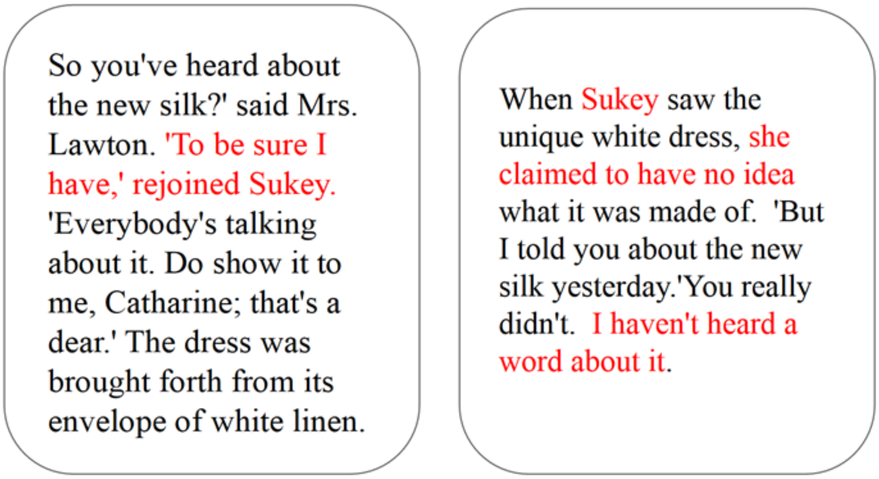
Table 10. Comparison of results of BERT word-averaging and BERT [CLS] token-based similarity metrics on sentences shorter than 22 tokens. SM indicates Search Mechanism. The asterisk * indicates cases where the difference between two measures is statistically significant with p ≤ 0.5.

SM	measure	% on the test set		
		prec	rec	F1
Sequence Search	**BERTword***	78.6	83.3	**78.6**
	BERTcls	77.5	78.2	77.5
Best Match (uni)	**BERTword***	73.9	83.9	**78.6**
	BERTcls	78.2	66.9	72.1
Best Match (bi)	**BERTword***	78.6	83.3	**80.9**
	BERTcls	83.7	64.7	73.0

3.5 Exploring Features for Non-paraphrased Paragraph-pair Detection

As shown in Table 1, we have found that some of the paragraph pairs we sampled from the Webis-CPC-11 were mislabeled as paraphrase-pairs, in which the meaning of the two paragraphs is not similar. Figure 5 shows an example of such a non-paraphrased pair, where the text fragments in red indicate two different meanings. In one paragraph the character “Sukey” is said to have heard about some issues, whereas in the other paragraph it is indicated she has no idea about them. In the 400 pairs with the positive labels that we sampled, 7 were not paraphrases.

Although we have excluded those outlier pairs from our previous experiments, they are manually detected, which would be too time-consuming to do so for a large corpus. Therefore, we would like to check whether it is possible to detect such incorrectly labeled data automatically. As the paragraph is just a longer passage in comparison with the sentence, we expect that the measures adopted to calculate the sentence similarity could be also applied to evaluate the paragraph similarity. We thus further test whether the measures adopted for sentence alignment are discriminative enough to filter out those incorrectly annotated paragraph pairs (i.e., non-paraphrased pairs found).



So you've heard about the new silk?' said Mrs. Lawton. 'To be sure I have,' rejoined Sukey. 'Everybody's talking about it. Do show it to me, Catharine; that's a dear.' The dress was brought forth from its envelope of white linen.

When Sukey saw the unique white dress, she claimed to have no idea what it was made of. 'But I told you about the new silk yesterday.'You really didn't. I haven't heard a word about it.

Figure 5. Example of a non-paraphrased paragraph pair (outlier) from the Webis-CPC-11 dataset. Red marks text fragments with opposite meanings.

To detect the outliers, we first calculate the paragraph similarity using the same similarity measures adopted in the previous experiments, but taking paragraphs, not sentences, as the input. We include the following similarity measures in the experiment: (1) based on the unit-overlap-ratio (including: NGRAM(n=3), TOKENstring, TOKENsyn); based on the sentence-vector-similarity (including SentenceBERT and BERTScore). We model the similarity values from all paraphrased paragraph pairs for each measure with a specific normal distribution and then calculate its 0.95 confidence interval to check whether the non-paraphrased paragraphs can be detected as outliers outside this interval.

Table 11 shows the percentage of non-paraphrased pairs that fall below the left boundary value of the 0.95 Confidence Interval for each of the adopted similarity measures. The best result is achieved using BERTprec, with which we can detect all outlier pairs. This leads to the conclusion that it is possible to automatically detect those non-paraphrased paragraph-pairs by using BERTScore as a similarity measure.

Table 11. Results of filtering out non-paraphrased paragraph pairs based on the 0.95 confidence interval. Mean is the mean similarity value for all (393) paraphrased paragraph pairs; L-CI is the left boundary of the Confidence Interval, and #pairs is the number of non-paraphrased pairs that fall outside the confidence interval (out of 7). Results with p ≤ 0.05 are marked with the asterisk *. Table adopted from Smolka et al. (2022).

measure	mean	L-CI (0.95)	% pairs
NGRAM(n=3)	0.547	0.530	71%
TOKENstring	0.221	0.214	57%
TOKENsyn	0.141	0.136	57%
SBERTbert	0.541	0.522	43%
SBERTalbert	0.411	0.391	43%
SBERTmini*	0.339	0.321	86%
BERTprec*	0.914	0.911	**100%**
BERTrec	0.917	0.914	71%
BERTf1*	0.915	0.913	71%

4. Error Analysis

We analyzed 50 errors generated by our best approach (i.e., Sequence Match with SBERTmini), and categorized them based on their associated error sources: (1) mistaking 1-n mapping for 1-1 (46%); (2) associated with incorrect sentence boundary (26%), in which the sentences are split incorrectly before conducting alignment (e.g., a sentence is incorrectly split into two sequences by the sentence segmenter); (3) paraphrased sentences take different sequence-orders within two given paragraphs (16%); (4) others (12%), of which it is difficult to attribute each error to a specific reason.

Table 12 shows an example of the first error category, which incorrectly marks a 1-n alignment as 1-1. The source of this error is likely due to the following two reasons. First, those proposed similarity measures are still incapable of truly reflecting the semantic similarity between two sentences when they are paraphrased in an abstract way; as a result, they might incorrectly convert a golden 1-n mapping into a 1-1 mapping. Second, because the alignment is selected based on the sentence similarity and the probability of each alignment type is estimated from the development set, the adopted model has a preference for extracting 1-1 alignments as they are most common in the dataset (cf. Table 1).

Table 12. An example which mis-interprets a one-to-many relationship as a 1-1 alignment. Gold sentence alignments (i.e., pairs "a", "b") are correctly extracted; "x" is incorrectly extracted and "0" is an annotated 1-n alignment which we do not want to extract.

		PARAGRAPH #1	PARAGRAPH #2
MODEL INPUT (FULL PARAGRAPHS)		**Thad, of course. And, Bill, we're going to get him, sooner or later.** Mr. Hooper won't want to stand this sort of thing forever. I've got a hunch that we're not through with that game yet.	**Naturally, Thad and also Bill, whom we'll get after a while.** Mr. Hooper won't let this go on for long. I'm guessing we won't be done for some time.
ALIGNED SENTENCES (GOLDEN ANSWER)	0	Thad, of course. And, Bill, we're going to get him, sooner or later.	Naturally, Thad and also Bill, whom we'll get after a while
	a	Mr. Hooper won't want to stand this sort of thing forever.	Mr. Hooper won't let this go on for long.
	b	I've got a hunch that we're not through with that game yet.	I'm guessing we won't be done for some time.
MODEL ANSWER	x	And, Bill, we're going to get him, sooner or later.	Naturally, Thad and also Bill, whom we'll get after a while.
	a	Mr. Hooper won't want to stand this sort of thing forever.	Mr. Hooper won't let this go on for long.
	b	I've got a hunch that we're not through with that game yet.	I'm guessing we won't be done for some time.

The second error category (i.e., with incorrect sentence boundary) occurs when the pre-processing module incorrectly split the sentences within one of the input paragraphs. Finally, the last type of error is caused by the sequence search mechanism, which assumes all paraphrased passage pairs follow the same relative order within each paragraph. If this assumption is violated in the given paragraph pair, it will always return an incorrect answer.

5. Discussion

Based on our results, we get the following observations:

- Among various sentence alignment strategies, Sequence Match tends to give the best and most consistent results across all our experiments. The advantage of Sequence Match is that it employs dynamic programming which makes it faster than the greedy approaches. It also performs well where the adjacency and dependency information between sentences is relevant to the matching. However, it will not perform well when the sentences are in a different order in the two paragraphs, in which case using the Best Match strategy would be preferable. Furthermore, the bi-directional Best Match shows much better performance than the uni-directional approach on both datasets we use, which can be explained by the symmetry in our data, as described earlier in the introduction section.

- In general, the measures that encode the sentence directly tend to perform better than those that are based on individual token representations (either unit-overlap or token-embedding-average). The only exception is the approach using the BERT [CLS] token. We believe it might be because the [CLS] token is not explicitly trained to condense a long text sequence into a vector, unlike SentenceBERT and BERTScore which are created specifically for doing so.
- The method using averaged word vectors from BERT outperforms the method using the [CLS] token in our task. The inferior performance of the method with [CLS] token representations might be due to that the [CLS] token is trained on a much smaller amount of data; in contrast, those individual token embeddings are trained from a much larger dataset.
- Noticeably, the best thresholds of those non-embedding methods tend to be much lower than those of the measures that utilize neural embeddings. We conjecture this is because the neural models estimate similarity based on soft/fuzzy matching (which would result lower thresholds), while string-based methods use hard/strict matching (which would result higher thresholds, as it cannot distinguish the soft matching case from the un-matched case).
- Finally, we have discovered that when the additional sentence length limitation is imposed, the performance drops across all approaches, with the biggest difference for the Sequence Matching approach. One possible explanation is that shorter sentences are easier to be mistakenly linked because they have less distinctive tokens (e.g., when comparing short sentences like "*John Walker went*." and "*John Walker came*.", the similarity between them will be always high because there is only one distinguishing token; however, it would be a less serious issue for the cases with longer sentences). Another reason might be that the sentence length limitation implicitly removes the trivial cases from the dataset, i.e., those cases where the whole paragraph only contains a single long sentence that will be automatically mapped to its corresponding paragraph (and forms a 1-1 mapping). Such cases are more likely to appear in the full dataset, which would make the overall result higher on this dataset.

6. Conclusions

We have presented the first comparison among various model-agnostic similarity measures used for aligning sentences among paraphrased paragraphs. For most cases, we find that embedding-based similarity measures outperform the string-based approaches (including the previous

SOTA character trigram approach tested on the TS dataset), and sentence-embedding-based methods are preferable to the word-embedding-based methods for most search mechanisms except the uni-directional greedy matching.

Additionally, our results have shown that in calculating the similarity for sentence alignment, word vector averaging is better than adopting the [CLS] token when retrieving a representation of a whole sentence from a BERT-based model.

References

Aghaebrahimian, A. (2017). Quora Question Answer Dataset. *TSD. Text, Speech, and Dialogue*, 66-73. https://doi.org/10.1007/978-3-319-64206-2_8

Barzilay, R., & McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01)*, 50-57. https://doi.org/10.3115/1073012.1073020

Barzilay, R., & McKeown, K. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, *31*(3), 297-328. https://doi.org/10.1162/089120105774321091

Blevins, T., Levy, O., & Zettlemoyer, L. (2018). Deep RNNs Encode Soft Hierarchical Syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 14-19. https://doi.org/10.18653/v1/P18-2003

Burrows, S., Potthast, M., & Stein, B. (2013). Paraphrase acquisition via crowdsourcing and machine learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*, *4*(3), 1-21. https://doi.org/10.1145/2483669.2483676

Choi, H., Kim, J., Joe, S., & Gwon, Y. (2021). Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR 2020)*, 5482-5487 https://doi.org/10.1109/ICPR48806.2021.9412102

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Dietterich, T.G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895-1923. https://doi.org/10.1162/089976698300017197

Dolan, W.B., & Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Fellbaum, Ch., (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gale, W.A., & Church, K.W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75-102.

Ganitkevitch, J., Durme, B.V., & Callison-Burch, C. (2013). PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 758-764.

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality Decomposed: How do Neural Networks Generalise? (Extended Abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 5065-5069. https://doi.org/10.24963/ijcai.2020/708

Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. (2020). Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics*, 7943-7960. http://dx.doi.org/10.18653/v1/2020.acl-main.709

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of 8th International Conference on Learning Representations (ICLR 2020)*. https://doi.org/10.48550/arXiv.1909.11942

Liu,Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Computing Research Repository, arXiv:1907.11692. Version 1. https://doi.org/10.48550/arXiv.1907.11692

MacCartney, B., & Manning, C.D. (2008). Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 521-528.

McNamee, P., & Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7, 73-97. https://doi.org/10.1023/B:INRT.0000009441.78971.be

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428-3448. https://doi.org/10.18653/v1/P19-1334

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of 1st International Conference on Learning Representations (ICLR 2013)*. https://doi.org/10.48550/arXiv.1301.3781

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Putra, J.W.G., & Tokunaga, T. (2017). Evaluating text coherence based on semantic similarity graph. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, 76-85. https://doi.org/10.18653/v1/W17-2410

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982-3992. https://doi.org/10.18653/v1/D19-1410

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27* (3). 832-837. https://doi.org/10.1214/aoms/1177728190

Smolka, A., Wang, H.M., Chang, J.S., & Su, K.Y. (2022). Is Character Trigram Overlapping Ratio Still the Best Similarity Measure for Aligning Sentences in a Paraphrased Corpus? In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, 49-60.

Štajner, S., Franco-Salvador, M., Rosso, P., & Ponzetto, S.P. (2018). CATS: A Tool for Customized Alignment of Text Simplification Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Ştajner, S., Franco-Salvador, M., Rosso, P., Ponzetto, S.P., & Stuckenschmidt, H. (2017). Sentence Alignment Methods for Improving Text Simplification Systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 97-102. https://doi.org/10.18653/v1/P17-2016

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C.D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101-108. https://doi.org/10.18653/v1/2020.acl-demos.14

Qiu, L., Kan, M., & Chua, T. (2006). Paraphrase Recognition via Dissimilarity Significance Classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 18-26.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, 5776-5788. https://dl.acm.org/doi/abs/10.5555/3495724.3496209

Weiss, D., Roit, P., Klein, A., Ernst, O., & Dagan, I. (2021). QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9879-9894. https://doi.org/10.18653/v1/2021.emnlp-main.778

Zhou, J., Zhang, Z., Zhao, H. & Zhang, S. (2020). LIMIT-BERT : Linguistics Informed Multi-Task BERT. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020*, 4450-4461. https://doi.org/10.18653/v1/2020.findings-emnlp.399

Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 31-46

探討語者驗證系統中特徵處理模組與注意力機制

Investigation of Feature Processing Modules and Attention Mechanisms in Speaker Verification System

陳廷威 [*]、林威廷[*]、陳嘉平[*]、呂仲理 [+]

詹博丞[+]、鄭羽涵[+]、莊向峰[+]、陳威妤[+]

Ting-Wei Chen, Wei-Ting Lin, Chia-Ping Chen, Chung-Li Lu,

Bo-Cheng Chan, Yu-Han Cheng, Hsiang-Feng Chuang, Wei-Yu Chen

摘要

本論文建構並替換不同的音訊特徵前處理模組與注意力機制來改進語者驗證系統。我們使用了基於 ECAPA-TDNN 所改進的模型作為基準模型，並透過替換與組合不同的前處理模組與注意力機制來進行比較，以選出最佳的組合作為論文提出的最終模型。訓練上我們使用了 VoxCeleb 2 資料集進行訓練，並使用多個測試集來測試模型的表現。最終模型在 VoxSRC2022 驗證集中對比基準模型有 16% 的進步幅度，成功在語者驗證系統上取得了更好的成效。

Abstract

In this paper, we use several combinations of feature front-end modules and attention mechanisms to improve the performance of our speaker verification system. An updated version of ECAPA-TDNN is chosen as a baseline. We replace and integrate different feature front-end and attention mechanism modules to compare and find

[*] 國立中山大學資訊工程學系
Department of Computer Science and Engineering, National Sun Yat-sen University
E-mail: {m103040017, m093040020}@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

[+] 中華電信研究院
Chunghwa Telecom Laboratories
E-mail: {chungli, cbc, henacheng, gotop, weiweichen}@cht.com.tw

the most effective model design, and this model would be our final system. We use VoxCeleb 2 dataset as our training set, and test the performance of our models on several test sets. With our final proposed model, we improved performance by 16% over baseline on VoxSRC2022 valudation set, achieving better results for our speaker verification system.

關鍵詞：語者驗證、前處理模組、注意力機制、時延神經網路
Keywords: Speaker Verification, Frontend Module, Attention Mechanism, Time Delay Neural Network

1. 緒論 (Introduction)

隨著資訊科技的日新月異，大量的數位化資訊充斥在我們的生活當中，透過各式各樣新穎的設備，任何事物、資料都可以被電子化的儲存，並隨時傳送到地球上的任何地方，這使得人們得以跳脫固有的時間與空間上限制，能以更為宏觀的視角來探索這個世界。然而，當每個人對這些通訊設備的依賴度越來越高時，個人資訊被不合法的洩漏、利用的情形也日漸增加，如何保護自身的資訊安全是一個非常迫切的議題。

語者辨識技術便是其中一項在近年來越來越受到重視的資訊防護方法，藉由這項技術，我們可以將語者的聲紋特徵轉換成具有語者特徵的嵌入向量，透過比對這個嵌入向量來對當前語者的身分進行確認，以防止個人資訊被偽造及竊取。

近年來，有許多過去在圖像領域發光發熱的模型結構被帶入到聲學領域當中，並為語者驗證技術帶來了極大的突破，像是以時延神經網路（Time Delay Neural Network, TDNN）作為主幹，並在其中引入了 Res2Net (Gao *et al.,* 2021) 多分支卷積結構與 SENet (Hu *et al.,* 2018) 注意力機制的 ECAPATDNN (Desplanques *et al.,* 2020) 與基於傳統二維卷積神經網路建構的 ResNet (He *et al.,* 2016a)，兩者都在近年的語者驗證競賽中取得亮眼的表現。而鑒於兩種截然不同架構都在競賽上取得優秀的成果，希望能夠集合兩種架構優點的新型架構被研究出來，也就是 ECAPA CNNTDNN (Thienpondt *et al.,* 2021)。在該模型中，ResNet 結構被設計為 ECAPA-TDNN 的前處理模組，用於降低輸入音檔特徵頻譜圖在頻率軸上的偏移，透過卷積操作重組與保留較重要之特徵訊息。該結構在實驗上進一步的提高模型的表現，並為語者驗證模型的變化性增加了更多的可能性。

在本篇論文中，我們使用基於 ECAPATDNN 架構進行改進的 Improving ECAPATDNN (Zhang *et al.,* 2021) 做為基底，透過修改部份結構以提出 IM ECAPA-TDNN 做為本次的基準模型，並將其依照 ECAPA CNNTDNN 的架構設計進行擴增。我們的實驗與分析集中在不同的前處理模組以及注意力機制上。首先，我們會將前處理模組替換為不同的結構進行訓練，除了原始的 CNN 結構外，我們另外實驗了預激活的 CNN 結構以及導入兩個維度注意力的 MFA 模組(Liu *et al.,* 2022)。之後我們會取這三組模型中表現較好的模型替換其中使用的注意力機制，將原有的 SE 模組分別替換成 CBAM 模組(Woo *et al.,* 2018) 以及 GC 模組(Cao *et al.,* 2019)。在我們的最終模型中，使用了預激活的 2D CNN 模組作為前處理模組以及 CBAM 模組作為模型的注意力機制，在

Voxceleb 1-O、Voxceleb 1-E、Voxceleb 1-H 及 VoxSRC2022 測試集上都實現了比起基準模型更好的表現。

本文主要分為五個部份，第一部份為緒論；第二部份為研究方法，會介紹使用到的資料前處理方法、模型架構、特徵前處理模組以及注意力機制；第三部份為實驗設置，說明實驗所使用到的資料集、參數設置以及評估準則；第四部份為實驗結果，會比較不同前處理模組與注意力機制的實驗數據，並根據實驗結果進行分析與討論；第五部份為結論。

2. 研究方法 (Research Methods)

在這個章節我們將會詳細的講解本次實驗所使用到的各種方法，包含對輸入音檔進行的處理、主幹模型架構的細節、不同前處理模組以及不同注意力機制的介紹。實驗上我們使用了 VoxSRC 官方所提供的訓練工具(Chung *et al.,* 2020) 進行訓練，並以 IM ECAPA-TDNN 做為基準模型，透過結合不同的前處理模組以及注意力機制觀察這些改動對模型效能所造成的影響。

2.1 資料前處理 (Data Preprocessing)

為了提高模型的強健性以及避免產生過度擬和（overfitting）的狀況，我們利用了資料增強的方法增加訓練資料的多樣性。透過對訓練音檔加入噪音跟迴響，能夠有效的提昇模型的泛化能力，使其在推論階段的表現更加優秀。而在將音檔轉換為特徵向量方面，在參考了近年競賽中各隊伍的作法後，我們選用梅爾頻譜作為主要聲學特徵。

2.1.1 資料增強 (Data Augmentation)

我們使用了兩種用於資料增強的資料集來對我們的訓練資料進行強化。首先是透過 MUSAN 資料集(Snyder *et al.,* 2015) 來為輸入音檔加入噪音，在 MUSAN 資料集中共分成了三個部份，分別為語音（speech）、音樂（music），以及噪音（noise），語音部份的內容全都是來自公共場合中的背景說話聲，包含朗讀書本章節以及美國政府部門聽證會等等，語音部份總共由 12 種語言組成，其中以英語的比例為最多；音樂部份的內容包含了多種不同時期、流派的音樂，比如有傳統流派的巴洛克、浪漫、古典音樂，也有流行流派的爵士、藍調、嘻哈音樂等等；噪音部份的內容則包含了科技性噪音（如撥號音、傳真機噪音等）以及環境聲音（如雷聲、雨聲、動物噪音等），有些檔案也會有包含模糊的人群噪音。另一個則是利用 RIR（Room Impulse Response，空間脈衝響應）資料集(Ko *et al.,* 2017) 將音檔加入迴響（Reverberate），在 RIR 資料集中有真實與模擬的聲音資料，我們只會使用模擬的空間音進行資料增強。

2.1.2 聲紋特徵擷取 (Acoustic Feature Extraction)

我們使用 80 維的梅爾頻譜（Mel-filter bankfeatures，FBank features）作為我們的主要聲學特徵，理由是相較於梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficients，MFCC）來說，梅爾頻譜因為沒有經過 DCT 變換，使得其保留了更多的聲音訊號資訊，能夠在分析語者特徵上取得更好的結果。

2.2 模型架構 (Model Architecture)

在 ECAPA-TDNN 推出之後，得益於優秀的多層聚合策略以及多尺度特徵卷積，該模型在各個語者驗證競賽中都取得優秀的表現，許多人也以其架構作為基底進行不同程度的改良。本篇論文我們以基於 ECAPA-TDNN 改進的 Improving ECAPA-TDNN 作為基底，配合後續實驗進行調整，降低了模型計算量並維持相近之模型表現。我們把修改後的模型命名為 IM ECAPA-TDNN，並將其作為本篇論文中的基準模型。

2.2.1 Improving ECAPA-TDNN

Improving ECAPA-TDNN 是基於 ECAPATDNN 所設計的一個改進版本。在該模型中，Zhang et al.使用了帶有 SE 注意力機制的 SCBlock (Liu *et al.,* 2020) 取代了原始架構主幹網路裡的 Res2Block，通過 SC-Block 所帶有的自校準計算及分割卷積來獲得更大的感受野（receptive field）及上下文的空間注意力，以此避免特徵中不必要的資訊，並在 SC-Block 後面接上 SE-Block，透過注意力機制使有效特徵圖（feature map）權重要大於低效的特徵圖。Zhang et al.還在每一層 SE-SC-Block 之間插入聚合（aggregation）層的結構，用來將不同分辨率的特徵串接整合並降採樣為下一層 SE-SC-Block 的輸入大小。這些聚合層會與原始 ECAPA-TDNN 的多層聚合方法結合，使模型成為一個階層式的聚合結構，也就是每一層 SE-SC-Block 的輸出都會作為之後每一層聚合層的輸入，而越接近模型尾端的聚合層就會融合越多不同分辨率的特徵，以提取更具語者資訊的嵌入向量。

2.2.2 IM ECAPA-TDNN

我們以 Improving ECAPA-TDNN 作為基底進行修改，最主要的改動便是我們減去了一層聚合層結構，與此同時也減去了一層的 SESC-Block，並將第一層 TDNN 結構的輸出也作為後面各聚合層的輸入，修改後的模型如圖 1 所示。我們想要透過聚合層來將保留更多特徵資訊的第一層 TDNN 輸出向量一併與後面每一層的 SE-SC-Block 的輸出向量進行特徵重組，以此來獲取更多的語者特徵訊息；而將 SE-SC-Block 及聚合層各減少一層的主要是考量到實驗彈性，由於首層 TDNN 的輸出會加入到每一層聚合層當中進行特徵重組，若是保留原有的四層結構，在替換前處理模組以及注意力機制的實驗上便會出現硬體限制的情況發生。基於以上原因，我們對原始的 Improving ECAPA-TDNN 進行了修改，並將修改後的模型命名為 IM ECAPA-TDNN。

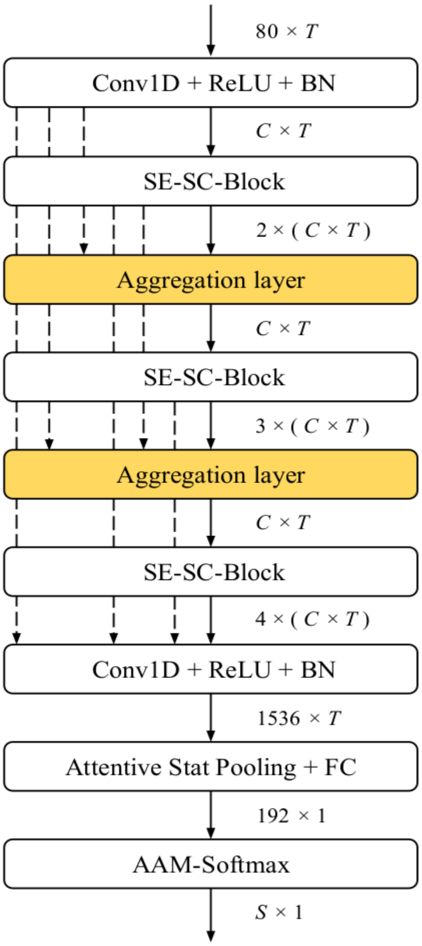


圖 1. 修改提出的 IM ECAPA-TDNN。其中 C 表示通道數，T 表示音框數，S 表示分類與者數量。
[Figure 1. The proposed IM ECAPA-TDNN. C denotes aschannels, T denotes as frames, S denotes as numbersof speaker.]

2.3 特徵前處理模組 (Feature Preprocessing Modules)

在 ECAPA CNN-TDNN 的研究成果中，通過將輸入音檔的特徵頻譜圖先傳入前處理模組中進行特徵重組，再將重組後的特徵圖在通道及頻率維度攤平（flatten），使其作為一般輸入傳入 ECAPA-TDNN 進行訓練能夠有效的提高模型表現，因此我們將這個設計加入基準模型當中。我們在 IM ECAPA-TDNN 前面實作了 3 種不同結構的前處理模組進行實驗，分別為原始論文中的 2D CNN 模組、經過預激活（pre-activation）修改的 2D CNN 模組，以及引入兩維度注意力 MFA 模組。

2.3.1 2D CNN 模組 (2D CNN Module)

為原始在 ECAPA CNN-TDNN 中所使用的前處理模組，通過一般的二維卷積與 ResNet 結構中的 ResBlock 進行組合而成，在實做上我們還有在 ResBlock 中加入 SE 模組，整體結構如圖 2 所示。由於實驗環境以及訓練時間等因素考量，我們將 residual block 的通道數下調為 64 以降低模型大小，同時參照原始模型設定將第一個及最後一個二維卷積的步幅（stride）設置為 2 來增加計算效率。

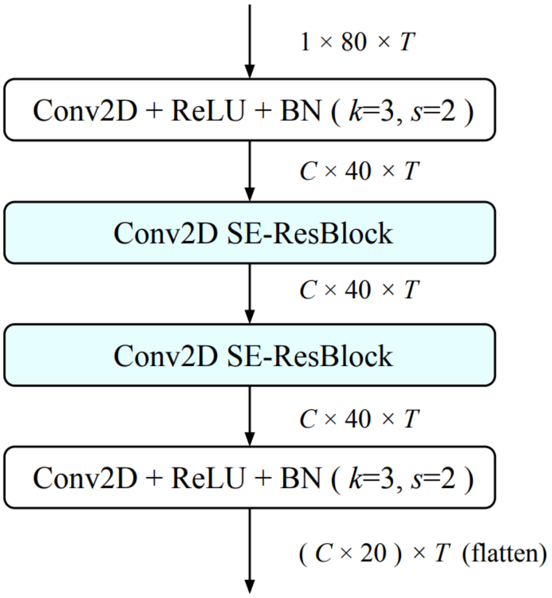
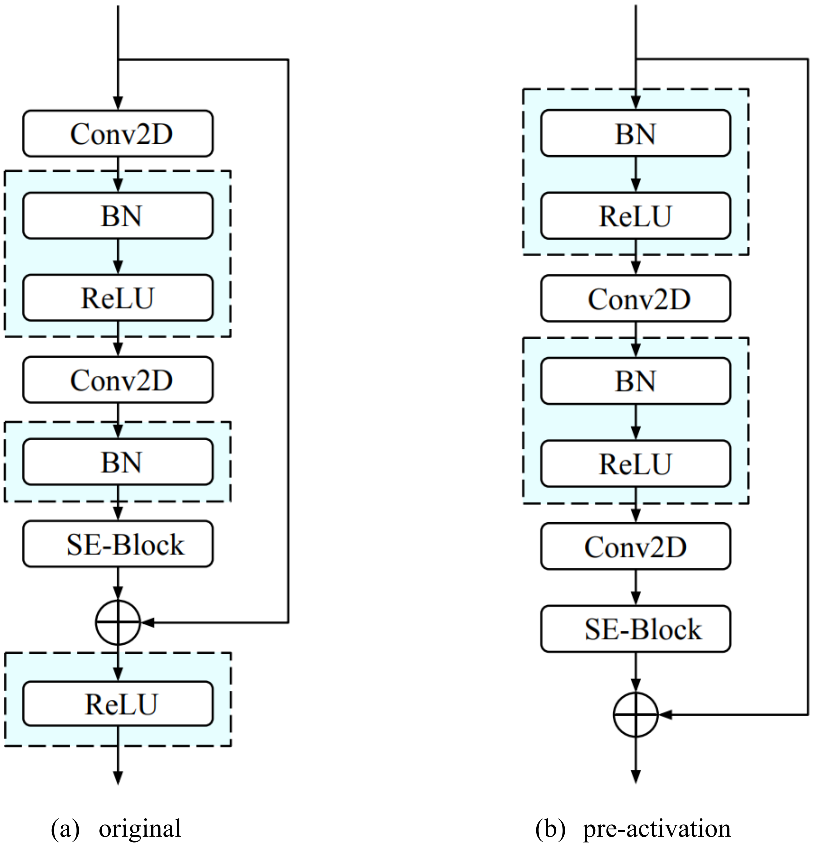


圖 2. 2D CNN 模組。其中 C 表示通道數，T 表示音框數。而卷積中的 k 與 s 表示卷積核大小及步伐長度。

[Figure 2. 2D CNN module. C denotes as channels, T denotes as frames. k and s in convolutions denote kernel size and stride.]

2.3.2 預激活的2D CNN 模組 (Pre-activated 2D CNN Module)

我們參考了(He *et al.*, 2016b) 中對殘差網路的研究結果，在該研究中表明當在 ResBlock 的捷徑連結（shortcut connection）上進行任何操作都會降低模型的表現；同時若是將模型中的激活函數從傳統的後激活（post-activation）改為預激活（pre-activation），能夠使模型更易於訓練，並有效的提高模型的泛化度。基於上述研究結果，我們將 2D CNN 模組中 ResBlock 的結構順序進行調整，新結構與舊結構比較如圖 3 所示。



(a) original (b) pre-activation

圖 3. 原始 SE-ResBlock 與預激活結構之比較。⊕ 表示元素對應相加。
[Figure 3. Comparison between original and pre-activation SE-ResBlock. ⊕ denotes aselement-wise addition.]

2.3.3 MFA 模組 (MFA Module)

MFA 模組是 Liu *et al*.在 MFA-TDNN 中設計用來取代 2D CNN 模組的新結構，其中使用了一個 Res2Block 變體來取代 ResBlock，這個變體是在傳統的 Res2Block 中改進了兩個新結構，也就是雙通道多尺度模組（dualpathway multi-scale module）以頻率及通道注意力模組（frequency-channel attention module），模組結構如圖 4 所示。雙通道多尺度模組的做法是在 Res2Block 中的每個分支卷積後額外再進行一個 TDNN 模組的卷積，並且這個模組的輸出會傳入到另一個分支當中，這就與 Res2Net 原有的卷積輸出形成了雙通道輸入到另一個分支中進行計算。頻率及通道注意力模組則是建構在前面提到的 TDNN 模組當中，結構如圖 5。其整體的概念其實與 SE 模組相似，不同的是特徵向量通過全局平均池化（Global average pooling，GAP）後是會留下頻率以及通道兩個維度的平面向量，接著將此向量攤平進行 SE 模組中激發（excitation）計算，最後再將激發後的向量重塑（reshape）回原來的平面向量並且作為權重值乘回原始的特徵向量。

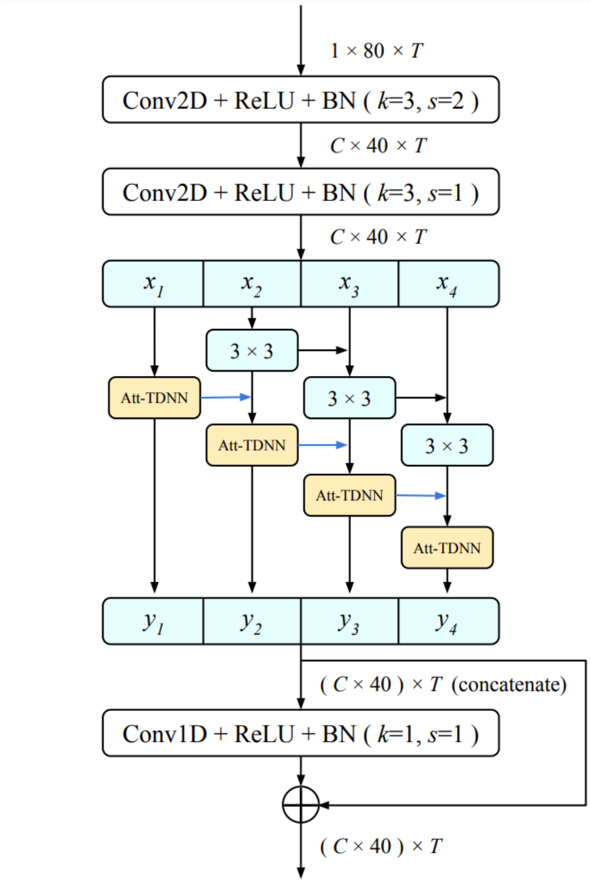

1 × 80 × T
Conv2D + ReLU + BN (k=3, s=2)
C × 40 × T
Conv2D + ReLU + BN (k=3, s=1)
C × 40 × T
x1
x2
x3
x4
3 × 3
Att-TDNN
3 × 3
Att-TDNN
3 × 3
Att-TDNN
Att-TDNN
y1
y2
y3
y4
(C × 40) × T (concatenate)
Conv1D + ReLU + BN (k=1, s=1)
(C × 40) × T

圖 4. MFA 模組。其中 C 表示通道數，T 表示音框數，卷積中的 k 與 s 表示卷積核大小及步伐長度，⊕ 表示元素對應相加。

[Figure 4. MFA module. C denotes as channels, T denotes as frames, k and s in convolutions denote kernel size and stride, ⊕ denotes as element-wise addition.]

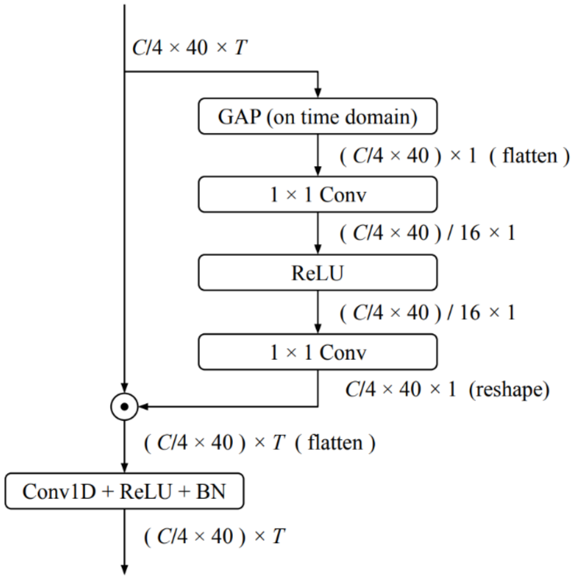


圖 5. MFA 模組中的 Att-TDNN 模組之結構。其中 C 表示通道數，T 表示音框數，⊙ 表示元素對應相乘。
[Figure 5. Att-TDNN module, which inside the MFA module. C denotes as channels, T denotes as frames, ⊙ denotes as element-wise product.]

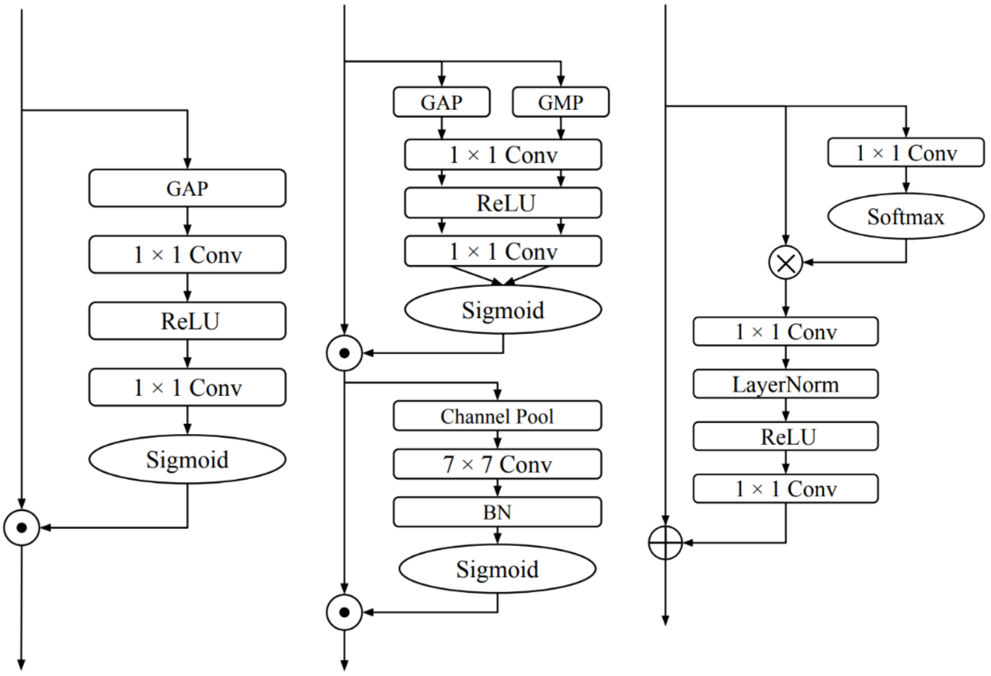
2.4 注意力機制 (Attention Mechanisms)

在原始的 ECAPA-TDNN 及後續的各個改進版本中，不論如何修改、擴增網路結構，其中都會引入注意力機制來提高模型整體的表現。就我們的基準模型以及 2D CNN 模組中使用到的 SE 模組來說，SE 模組會對特徵向量操作後取得特徵向量各通道不同的權重，透過權重，我們可以抑制特徵中不重要的資訊，並有效的將重要的特徵資訊給凸顯出來。而考慮到在 SE 模組問世至今，已有許多後起之秀在各大競賽中脫穎而出，藉由自身獨特的結構設計進一步增強注意力機制在模型上的影響，我們在此替換並比較包含 SE 模組在內，共計 3 種不同結構的注意力機制在本次語者驗證系統上的表現，要替換成的模組分別是 CBAM 模組以及 GC 模組。關於這些注意力模組的詳細結構請見圖 6。而由於 MFA 模組中自身較特殊的注意力設計，我們並不會替換 MFA 模組當中使用的注意力機制。

2.4.1 SE 模組 (SE Module)

SE（Squeeze and Excitation）模組為原始結構中所使用的注意力機制模組，模型結構如圖 6(a) 所示。其透過壓縮（squeeze）與激發（excitation）兩步驟來計算不同通道的權重。首先是壓縮，輸入特徵會對通道以外的維度進行全局平均池化計算，以取得各個通道的記述子（descriptor）；再來是激發，各通道的記述子會輸入兩層卷積層中進行降維升維

的操作，來學習不同通道記述子的重要程度，並透過 sigmoid 函數將其轉換成通道權重乘回原始特徵向量當中。



(a) SE 模組(SE module)　(b) CBAM 模組 (CBAM module　(c) GC 模組(GC module)

圖 6. 不同注意力機制模組之結構。⊙ 表示元素對應相乘，⊗ 表示矩陣相乘，⊕ 表示元素對應相加。

[Figure 6. Different attention mechanism architectures. ⊙ denotes as element-wise product, ⊗ denotes as matrix multiplication, ⊕ denotes as element-wise addition.]

2.4.2 CBAM 模組 (CBAM Module)

CBAM（Convolutional Block Attention Module）模組是基於 SE 模組的擴展，模型結構如圖 6(b) 所示。其在計算完通道權重之後，會接著計算空間權重以突顯更重要的空間特徵。同時在兩種權重的計算當中除了使用全局平均池化之外，還會使用全局最大池化（Global map pooling，GMP）來取得更多不同的資訊。

2.4.3 GC 模組 (GC Module)

GC（Global Context）模組是將 SE 模組與 Non-local 模組(Wang *et al.*, 2018) 進行結合而成，模型結構如圖 6(c) 所示。鑑於 Non-local 模組優秀的上下文建模（context modeling）能力與 SE 模組輕量的計算結構， Cao et al.通過簡化 Non-local 模組，然後將 Non-local

模組的特徵轉換層修改為類 SE 模組的結構以融合兩模組的優點。透過這樣的設計，GC 模組在各項電腦視覺領域的競賽當中皆有不俗的表現。

3. 實驗設置 (Experiments)

這個章節我們會介紹本論文實驗中所使用到的訓練資料集以及測試資料集，也會詳細描述模型在訓練中所設置的各項超參數，並說明最終用來評估模型表現的準則。

3.1 資料集 (Datasets)

我們使用 VoxCeleb 2 (Chung *et al.,* 2018) 中 dev 的部份作為我們的訓練資料集，並使用以 VoxCeleb 1 (Nagrani *et al.,* 2017) 資料集音檔所組成的 VoxCeleb 1-O/E/H 測試集以及 VoxSRC 2022 的驗證集作為本次模型的測試集。我們並沒有使用語音活性偵測（Voice activity detection，VAD）對實驗音檔進行調整。

3.2 參數設置 (Implementation details)

為了公平比較模型表現，所有模型皆套用了相同的訓練策略進行訓練：使用 Adam 優化器（optimizer）調整神經網路參數，初始學習率為 1e-03，每 10 個 epoch 會減少 25%。使用 AAM-Softmax 作為損失函數，其中 margin 設為 2，scale 設為 30。訓練期間應用權重衰減來防止模型過度擬合，將值設為 2e-05。訓練時的 batch size 設置為 256，並訓練 100 個 epoch 取其中最好的模型參數。主幹網路 IM ECAPA-TDNN 中的通道數量皆設置為 512，語者嵌入的輸出大小設置為 192；在前處理模組方面，2D CNN 模組不論是否為預激活其通道大小都設置為 64，而 MFA 模組基於模型大小則設為 32。

3.3 評估準則 (Evaluation Metrics)

我們以等錯誤率（Equal Error Rate, EER）以及最小檢測成本函數（Minimum Detection Cost Function, MinDCF）作為我們評估系統表現的準則。其中最小檢測成本函數依照 VoxSRC 2022 設定的標準，將參數設置為 C_{miss}=1、C_{fasle}=1、P_{target}=0.05。我們並沒有使用任何分數正規化方法對分數進行調整。

4. 實驗結果 (Experimental Results)

我們首先比對了原始 ECAPA-TDNN 與本次作為基準模型的 IM ECAPA-TDNN 在最簡單的 VoxCeleb1-O 及最困難的 VoxSRC2022 驗證集上的表現，其結果如表 1 所示。可以看到經過修改後的 IM ECAPA-TDNN 雖然在困難資料集上的表現與原始 ECAPA-TDNN 相差無多，但在簡單資料集上明顯是更為優秀的一方。

接著我們會分別討論不同的前處理模組以及不同的注意力機制對模型表現所造成的影響，並將表現最好的組合做為我們的最終模型。所有模型在各個測試集上的詳細結果如表 2 所示。

表 1. IM ECAPA-TDNN 與 ECAPA-TDNN 在最簡單及最困難的資料集上之表現比較
[Table 1. Comparisonthe performance beteen IM ECAPA-TDNNand ECAPA-TDNN on the easiest and the hardiesttest sets.]

Architecture	**VoxCeleb1-O**		**VoxSRC2022 val**	
	EER(%)	minDCF	EER(%)	minDCF
ECAPA-TDNN (Re-implemented)	1.3770	0.0931	**3.6735**	0.2479
IM ECAPA-TDNN	**1.2600**	**0.0849**	3.6824	**0.2462**

表 2. 不同模型在各測試集上的表現比較
[Table 2. Comparison the performance between different models on each test sets.]

Architecture	**VoxCeleb1-O**		**VoxCeleb1-E**		**VoxCeleb1-H**		**VoxSRC2022 val**	
	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
IM ECAPA-TDNN (baseline)	1.2600	0.0849	1.4733	0.0941	2.6891	0.1621	3.6824	0.2462
不同的前處理模組								
IM ECAPA CNN-TDNN	1.1218	0.0886	1.2763	0.0825	**2.3318**	0.1475	**3.2230**	0.2144
IM ECAPA CNN-TDNN (pre-act)	**1.0424**	**0.0739**	1.2646	0.0831	2.3518	**0.1415**	3.4471	0.2198
IM ECAPA MFA-TDNN	**1.0424**	0.0797	**1.2632**	**0.0813**	2.3526	0.1439	3.2535	**0.2118**
不同的注意力機制								
IM ECAPA CNN-TDNN (pre-act) with SE	**1.0424**	**0.0739**	1.2646	0.0831	2.3518	**0.1415**	3.4471	0.2198
IM ECAPA CNN-TDNN (pre-act) with CBAM	1.1484	0.0817	**1.2507**	**0.0821**	**2.3500**	0.1437	**3.1160**	**0.2053**
IM ECAPA CNN-TDNN (pre-act) with GC	1.2552	0.0992	1.3807	0.0926	2.5533	0.1551	3.4990	0.2282

4.1 前處理模組的比較(Comparison between Feature Prepocessing Module)

在加入了前處理模組之後，所有的模型相較於基準模型都有顯著的進步。相比於 2D CNN 模組在各個資料集上都有穩定的發揮，預激活的 2D CNN 模組雖然在相對簡單的 Voxceleb1-O 測試集上明顯優於原始的 2D CNN 模組，但是其在複雜度越高的測試集上表現卻較為差勁，我們認為主要是由於我們使用了輕量的 IM ECAPA-TDNN 作為主幹

網路，而在(He *et al.,* 2016b) 中表明了預激活的 ResBlock 要在深層的網路結構中才能發揮效果，所以才造成預激活 2D CNN 模組在複雜測試集上表現不佳的原因。而 MFA 模組得益於其多尺度多維度注意力的卷積結構，其在簡單的測試集上可以做到與使用預激活 2D CNN 模組一樣優異的表現，並在複雜的測試集上表現相對穩定。

4.2 注意力機制的比較 (Comparison between Attention Mechanisms)

考慮到 MFA 模組本身自帶的注意力機制無法輕易變動，我們在 2D CNN 模組中選擇了預激活的版本替換其注意力模組，來觀察各注意力機制對模型表現造成的影響。SE 模組在相對簡單的 Voxceleb1-O 測試集上依舊有著較佳的表現，但是 CBAM 模組在其他更為複雜資料集對比另外兩個注意力模組都有著更優秀的結果。會有這樣的差異我們認為是因為 CBAM 模型引入空間注意力能夠有效的將更多重要的語者特徵突顯出來，且相比 SE 模組只做了全局平均池化，CBAM 還加入了全局最大池化進行計算以取得不同方面的資訊，這些設計讓模型能夠在複雜的測試集上擷取更細微的特徵進行辨識，進而提高了辨識結果的表現；對比 CBAM 的優異表現，GC 模組反而在所有測試集的表現都不突出，會有這樣的問題我們認為是模組的設計與 TDNN 結構衝突，將模組結構硬是改寫為相容 TDNN 反而造成擷取特徵時產生冗餘的資訊，導致 GC 模組連 SE 模組的表現都達不到。

4.3 最終提出模型 (Final Proposed Model)

根據我們上述的實驗結果，我們將帶有預激活 2D CNN 前處理模組，並替換注意力機制為 CBAM 的 IM ECAPA-TDNN，即表 2 中的 IM ECAPA CNN-TDNN (pre-act) with CBAM 做為我們的最終提出模型。相比與基準模型，我們的最終模型在各測試集上都有明顯的進步，以最複雜的 VoxSRC2022 驗證集來說，最終模型在 EER 值與 minDCF 值上分別有 15.4% 以及 16.6% 的進步幅度。

5. 結論 (Conclusions)

本論文提出了基於 Improving ECAPA-TDNN 修改的 IM ECAPA-TDNN 結構作為我們的基準模型，並透過結合不同的前處理模組以及調整注意力機制來對模型表現進行進一步的強化。我們提出的最終模型通過結合預激活的 2D CNN 前處理模組與替換注意力機制為 CBAM 模組，在各項測試集上的表現對比基準模型都有著大幅提昇。未來我們將會以此為依據來修改其他更加複雜的主幹網路，希望能夠藉此來進一步的提昇我們語者驗證系統的效能。

參考文獻 (References)

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of 2019 IEEE/CVF International*

Conference on Computer Vision Workshop (ICCVW). https://doi.org/10.1109/ICCVW.2019.00246

Chung, J.S., Huh, J., Mun, S., Lee, M., Heo, H.-S., Choe, S., Ham, C., Jung, S., Lee, B.-J., Han, I. (2020) In Defence of Metric Learning for Speaker Recognition. In *Proc. Interspeech 2020*, 2977-2981, https://doi.org/10.21437/Interspeech.2020-1064

Chung, J.S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. CoRR, abs/1806.05622.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., & Torr, P. (2019). Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, *43*(2), 652-662. https://doi.org/10.1109/TPAMI.2019.2938758

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. In *Proceedings of ECCV 2016*, 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

Hu, J., Shen, Li, & Sun, G. (2018). Squeeze-andexcitation networks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5220-5224. https://doi.org/10.1109/ICASSP.2017.7953152

Liu, J.-J., Hou, Q., Cheng, M.-M., Wang, C., & Feng, J. (2020). Improving convolutional networks with self-calibrated convolutions. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10093-10102. https://doi.org/10.1109/CVPR42600.2020.01011

Liu, T., Das, R. K., Lee, K. A., & Li, H. (2022). Mfa: Tdnn with multi-scale frequency-channel attention for textindependent speaker verification with short utterances. In *Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7517-7521. https://doi.org/10.1109/ICASSP43922.2022.9747021

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of Interspeech 2017*, 2616-2620.

Snyder, D., Chen, G., & Povey, D. (2015). Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484

Thienpondt, J., Desplanques, B., & Demuynck, K. (2021). Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification. arXiv preprint arXiv:2104.02370

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7794-7803.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I.S. (2018). Cbam: Convolutional block attention module. arXiv preprint arXiv:1807.06521

Zhang, Y.-J., Wang, Y.-W., Chen, C.-P., Lu, C.-L., & Chan, B.-C. (2021). Improving Time Delay Neural Network Based Speaker Recognition with Convolutional Block and Feature Aggregation Methods. In *Proc. Interspeech 2021*, 76-80. https://doi.org10.21437/Interspeech.2021-356

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 47-56

中英文語碼轉換語音合成系統開發

Development of Mandarin-English Code-switching Speech Synthesis System

練欣柔 [*]、黃立宇[*]、陳嘉平[*]

Hsin-Jou Lien, Li-Yu Huang, and Chia-Ping Chen

摘要

本論文提出中英文語碼轉換語音合成系統。為了使系統可專注於學習不同語言間的內容，利用已統一語者風格的多語言人工資料集進行訓練。之後在合成器中加入語言向量，以增加系統對多語言的掌握。此外對輸入的中、英文分別進行不同的前處理，將中文進行斷詞且轉為漢語拼音，藉此增加語音的自然度，且減輕學習時的複雜度，也透過數字正規化判斷句子中的阿拉伯數字，是否需要加上數字單位。英文部份則對複雜的頭字語進行讀音判斷與轉換。

Abstract

In this paper, the Mandarin-English codeswitching speech synthesis system has been proposed. To focus on learning the content information between two languages, the training dataset is multilingual artificial dataset whose speaker style is unified. Adding language embedding into the system helps it be more adaptive to multilingual dataset. Besides, text preprocessing is applied and be used in different way which depends on the languages. Word segmentation and text-to-pinyin are the text preprocessing for Mandarin, which not only improves the fluency but also reduces the learning complexity. Number normalization decides whether the arabic numerals in sentence needs to add the digits. The preprocessing for English is acronym conversion which decides the pronunciation of acronym.

關鍵詞：語音合成、語碼轉換、資料前處理
Keywords: Speech Synthesize, Codeswitching, Text Preprocessing

[*] 國立中山大學資訊工程學系
Department of Computer Science and Engineering, National Sun Yat-sen University
E-mail: {m103040105, m093040070}@nsysu.edu.tw; cpchen@cse.nsysu.edu.tw

1. 緒論 (Introduction)

語碼轉換（Code-switching）是指在一句話或多句話裡，含有一種以上的語言被交替使用，這種情況在現今社會中十分常見，為因應這種趨勢，語音合成系統也朝著多語言（Multilingual）的方向發展。現今的語料多為單語言，較少有同一語者的多語言語料，這導致語碼轉換在訓練時會遇到許多問題，像是語者無法合成非母語的語句，亦或是語者隨著句子語言轉換而改變的狀況。為解決上述問題，我們參考任意語者風格中英文語音合成系統(Wang, 2021)，做為我們的資料生成模型，給予系統一個參考音檔，其可生成與參考音檔相同語者風格的聲音訊號，藉此系統統一多語言資料集的語者風格，以建置多語言語音合成系統。

在本文中，使用 FastSpeech2 (Ren *et al.*, 2020) 做為合成器， 將編碼器與解碼器改為(Gulati *et al.*, 2020) 所提出的Conformer架構，聲碼器使用 HiFi-GAN (Kong *et al.*, 2020)。此外為增加系統對於多語言的掌握，於合成器中加上語言向量（language embedding），並為句子依中、英文編上語言 ID（language ID），而語碼轉換的句子無法直接以單一語言 ID 表示，對此在實驗中進行了處理。

我們發現因中文數量龐大、字詞讀音多變，因此將中文字轉換為漢語拼音，降低系統學習時的複雜度。此外也發現交雜在中文句子中的阿拉伯數字，有需要數字單位與否的問題，於是對此進行正規化。而在英文中經常使用的頭字語，是指將一句話或較長的名詞，縮寫成連續大寫字母，其發音分為字母讀音或視為新單字，要系統完整學習所有的英文頭字語是較為困難的，我們創建頭字語字典，以便進行分類讀音方式，並進行轉換，我們透過對文本進行資料前處理，以降低複雜度，提升中文、英文語音合成之正確率。

論文之其餘章節安排如下，章節二：研究方法描述系統架構、改進方法及文字前處理；章節三：實驗設置描述資料集與模型參數設定；章節四：實驗結果對基礎架構與改進後的系統進行比較；章節五：總結我們系統的優點和未來的改進方向。

2. 研究方法 (Research Methods)

以 Conformer-FastSpeech2 加上語言向量作為模型架構，為了提升中、英文語音合成的品質，分別對輸入的中文和英文文本做不同的前處理。在中文方面，使用中文斷詞、文字轉拼音與數字正規化，英文則執行縮寫讀法判斷與轉換，並且針對語碼轉換做語言 ID 編碼，與因中、英文語速差異進行的調整。

2.1 多語言語音合成系統 (Multilingual Speech Synthesis System)

在本文中， 使用 FastSpeech2 (Ren *et al.*, 2020) 做為合成器， 聲碼器使用 HiFi-GAN (Kong *et al.*, 2020)。

FastSpeech2 是一個非自迴歸(Non-autoregressive)的模型，可用更短的時間合成出與自迴歸(Auto-regressive) 模型相同品質的語音。架構中的編碼器和解碼器使用

Transformer 架構，在我們系統中將 Transformer 改為 Conformer (Gulati *et al*., 2020)，並命名為 Conformer-FastSpeech2（CFS2）。Conformer 結合了 Transformer 和卷積模組 (Convolution module) 以增強效果，其網路包含前饋神經網路（Feed Forward Module）、多頭自注意力機制（Multi-Head Attention Module)、卷積模組、層正歸化（Layer Normalization)。系統架構如圖 1，而圖右半邊則為 Conformer 的架構。

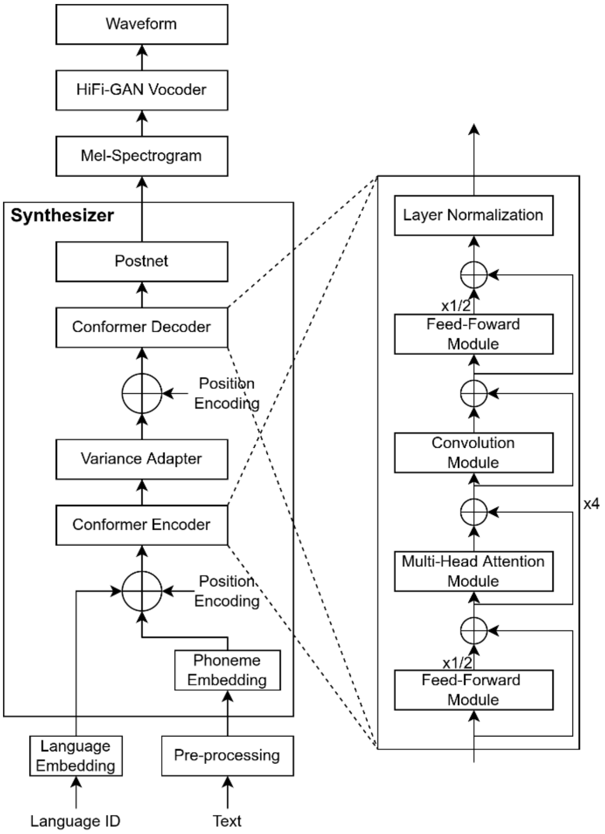


圖 1. Conformer-FastSpeech2 系統架構圖，基於 FastSpeech2 加入語言向量。右半邊為 Conformer 的架構，其基於 Transformer 架構再加上卷積模組以增強效果。
[Figure 1. The architecture of the Conformer-FastSpeech2 is based on the backbone of FastSpeech2 and combined with language embedding. The right handside of the figure is the architecture of Conformer which is associated with Transformer and convolution module to enhance feature extraction.]

在架構中加入語言向量， 並將其和 phoneme embedding 串接在一起做為編碼器的輸入。藉此提升系統合成多語言的表現，另外，依照資料的語言給予編號，稱為語言 ID，使用 0 和 1 為 language ID 分別表示英文與中文。

2.2 中文資料前處理 (Mandarin Data Preprocessing)

由於人們在交談時是有些微停頓的，為了讓系統學習語音這些細節， 我們首先使用了一個 python 工具名為 Jieba (Sun, 2012) 進行中文斷詞（Word segmentation），當中共有四種不同的斷詞模式，實驗中使用預設的精確模式，利用將符號置於斷詞處，以表示語句中的停頓，進而提升語音的自然度。此外 Jieba 工具可自行匯入符合使用者需求的字典，實驗中將 CKIP team (Ma & Chen, 2004) 的字典匯入，以提升斷詞的準確度，也將 CLMAD (Bai *et al*., 2018) 整理成另一份擴充字典，當系統應用於特殊領域時可匯入。

然而因為中文字本身數量龐大、字詞讀音多變，要系統學習所有的字詞是過於複雜的，因此不可直接將其作為輸入。於是我們利用 pypinyin 將中文字轉換成漢語拼音，其為一個 grapheme-to-phoneme（G2P）的 python 工具，以英文表示拼音，並用數字表示聲調（Tone），藉由此拼音組合的轉換簡單化中文的表示，使系統可以用較簡單的方式學習中文的發音。表 1 為中文斷詞及文字轉拼音的範例。

表 1. 中文資料前處理。先對文本進行中文斷詞，再將其轉換為漢語拼音。
[Table 1. Mandarin Data Preprocessing. Word segmentation would be applied before text-to-pypinyin.]

前處理方法	文本狀態
原始文本	明天不會下雨
中文斷詞	明天* 不會* 下雨
文字轉拼音	ming2 tian1 * bu4 hui4 * xia4 yu3*。.

此外我們還發現，當阿拉伯數字若在中文句子中時，會有是否需要唸出數字單位的差異，數字單位是指個、十、百、千、萬等。因此我們參考 [1]Chinese Text Normalization 作為基礎概念，其做法為將數字的常用情況進行分類，並以 Regular Expression 對數字找出相對應的模式，再判斷是否需加上數字單位，然而我們對模式內容進行修改，使其更貼近我們所需，共有五大種模式，表 2 為各模式的例子及正規化後的結果。

2.3 英文資料前處理 (English Data Preprocessing)

英文的頭字語可細分為 acronym 和 initialism，兩者的差異是縮寫後的單字該如何發音。acronym 指將縮寫後的單字讀為一個新的詞，例如：NASA 會讀做“na-suh”，FOMO 讀做“fow-mow”，而 initialism 則是指在發音上只念字母的讀音，而非視為一個新的詞，像是 FBI、NBA、BBC 等。然而由於頭字語為 acronym 或是 initialism，較難單純以文字進行分類，這導致系統難以學習，因此我們收集大量的頭字語，自行建立了一個頭字語字典，當輸入的文本含有全大寫的英文時，搜尋字典確認此輸入是否為 initialism，若是，則將字母轉換為相似讀音，以增加合成的正確性，若非則不做更改，舉例來說，當 BBC 經確認是 initialism ，會轉換為“bee bee ci”，FBI 則會轉換為“ef bee I”。

[1] https://github.com/speechio/chinese_text_normalization

表 2. 輸入的文本以 Regular Expression 找出相對應的模式，判斷是否要加上數字單位或其他處理。

[Table 2. Use Regular Expression to check the pattern of the text, and decides whether it requires additional number units.]

模式名稱	範例文本	正規化結果
Date	1986 年 8 月 18 日	一九八六年八月十八日
	1997/9/15	一九九七年九月十五日
Money	19588 元	一萬九千五百八十八元
Phone 手機	0919114115	零九一九一一四一一五
Phone 市話	02-2720-8889	零二二七二零八八八九
percentage	62%	百分之六十二
cardinal 量詞	1999 個蘋果	一千九百九十九個蘋果
	130 顆球	一百三十顆球
	124000 瓶水	十二萬四千瓶水
cardinal 編號	cardinal 編號學號是 103040100	學號是一零三零四零一零零
cardinal 純數	175.5 公分	一百七十五點五公分

2.4 針對語碼轉換之處理 (Process for Code-switching)

在訓練階段，使用英文和中文兩種語言 ID 進行語言向量。然而在合成階段，若輸入為語碼轉換的文本，無法單純以中文或英文予以編號。為此設立編定語言 ID 於語碼轉換文本之方法，如圖 3 所示，首先依語言分段輸入的文本，計算各分段的字元長度，藉由相對位置予以對應的語言 ID 且進行語言向量。分段後的文本分別進行資料前處理，再進行音素向量（phoneme embedding）作為編碼器的輸入。最後將編碼器輸出的隱藏特徵序列（hidden state sequence），和語言向量的輸出相加，獲得新的隱藏特徵序列進行後續的訓練。

由於中、英文資料集的語速差異，導致系統在合成語碼轉換之句子時，會有英文部份語速較快而感到不自然的問題。FastSpeech2 架構中的 Length Regulator，有一參數 α 可調整 duration predictor 輸出的時長（duration）大小，藉此改變梅爾頻譜圖的隱藏特徵序列長度，α 預設為 1。若 α= 1.5，表示將時長序列乘上 1.5 倍，進而使隱藏特徵序列拉長 1.5 倍，即為放慢速度。搭配語言 ID，即可透過相對位置單獨調整英文的速度。兩者差異如圖 2。左圖為無搭配語言 ID，針對整個序列進行調整。右圖則為單獨對英文進行調整，將英文部份的時長與 α 相乘，四捨五入，獲得新的時長序列。

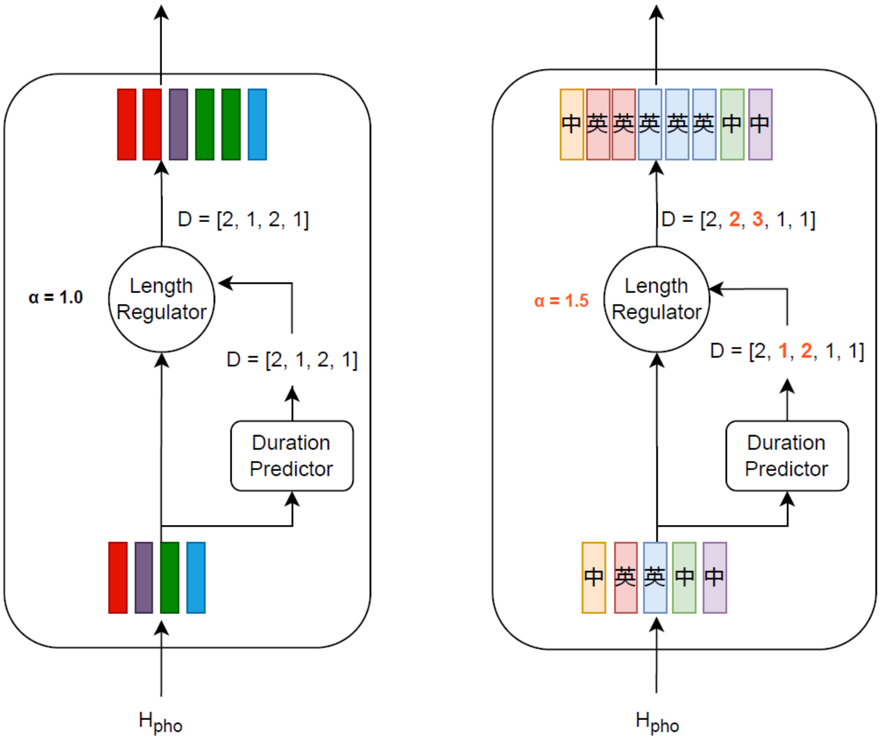


圖 2. Length Regulator 的作法。以 Length Regulator 中之參數α 調整隱藏特徵序列長度架構圖。D 表示時長（duration），H_{pho} 表示 phoneme 的隱藏特徵，H_{mel} 為梅爾頻譜圖的隱藏特徵。右側為依語言 ID 選取要調整的時長元素，再將元素乘上α 後四捨五入，得到新的時長以調整序列長度，左側則為對全部序列進行調整。

[Figure 2. Length Regulator. Use the parameter α in Length Regulator to adjust the length of the hidden state sequence. D denotes duration. H_{pho} denotes the phoneme hidden state. H_{mel} denotes the mel-spectrogram hidden state. The right handside of the figure shows that the specific duration is decided by the language ID. The duration elements multiply α and round it to get a new duration sequence. The left hand side of the figure adjusts all sequence.]

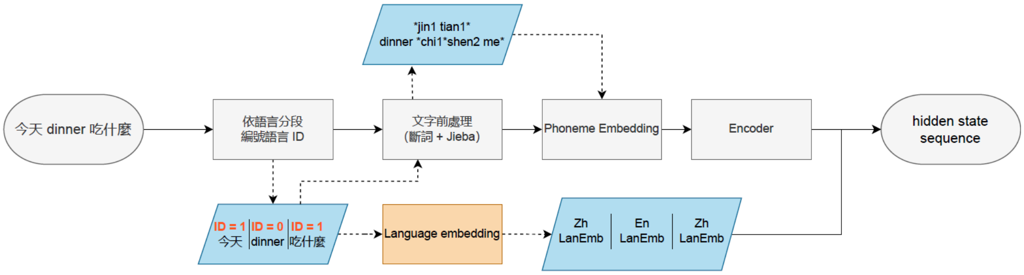


圖 3. 語碼轉換語言向量流程圖，LanEmb 表示語言向量。將輸入文本依語言分段並編號語言 ID，每段依序進行資料前處理、音素向量，將結果做為編碼器的輸入。對語言 ID 進行語言向量，將輸出與編碼器的輸出相加。

[Figure 3. Flow chart of the code-switching language embedding. LanEmb denotes language embedding. The input would be categorized by its language and the corresponding language ID would be given. After that, the result of the data preprocessing and phoneme embedding would be the input of the encoder. Finally, the output of the encoder would merge with the result of language embedding.]

3. 實驗設置 (Experiments)

在實驗中的資料集分為原始資料集，以及利用資料生成系統所生成的人工資料集，此外使用 ESPnet2 (Watanabe *et al*., 2018) 做為開發工具協助開發。

3.1 資料集 (Datasets)

- 原始資料集：使用的資料集包含中文語料 AISHELL3 (Shi *et al*., 2020) 及英文語料 VCTK (Yamagishi *et al*., 2019) ，在實驗中發現無需使用全部的資料，即可訓練出一個品質相當的系統，減少資料量亦可減少整體訓練時間，因此各選取了 30 名語者的資料做為我們實驗用的資料集，時長約為整體資料集的四分之一，並命名為 AISHELL3-thirty 和 VCTK-thirty，資料集的詳細資訊如表 3 所示。
- 人工資料集: 參考任一語者風格中英文語音合成系統(Wang, 2021)，作為我們的資料生成系統。選用 AISHELL3 資料集中的一個音檔作為參考音檔，並使用 AISHELL3-thirty 和 VCTK-thirty 的文本作為生成資料時的文本，藉此生成與參考音檔相同語者風格的多語言資料集，將其稱為 Generated-multi，共 25,362 筆音檔，共 15.6 小時，如表 3 所示。

3.2 訓練設定 (Implementation details)

本文使用 ESPnet2 (Watanabe *et al*., 2018)作為開發的工具。CFS2 的訓練集為多語言的 Generated-multi，架構中的 Conformer 編碼器和解碼器 kernel size 分別為 7 及 31，padding

為 3 與 15，優化器（Optimizer）使用 Adam (Kingma & Ba, 2014)，學習率（Learning rate）設定為 1。因為我們的系統為多語言的語音合成系統，為了使聲碼器可將多語言的梅爾頻譜圖轉為聲音訊號，HiFi-GAN 聲碼器利用 AISHELL3-thirty 和 VCTKthirty 資料集進行訓練， 批量大小(Batch size)設定為 32，使用 Adam 作為優化器，學習率設定為 0.0002。

表 3. 資料集詳細資訊。包含選取三十位語者的 VCTK-thirty 和 AISHLLE3-thirty，及生成資料集 Generate-multi。

[Table 3. The details of the dataset contain VCTK-thirty, AISHELL3-thirty and Generate-multi which is the generated dataset.]

資料集	音檔數量	總時長（小時
VCTK-thirty	11, 654	22.5
AISHELL3-thirty	13, 708	19
Generate-multi	25, 362	15.6

4. 實驗結果與分析 (Results and Analysis)

本實驗採用平均意見分數(Mean Opinion Score, MOS) 作為評估機制，分數區間為 0(低）～5（高），針對語音的整體品質進行評分，包含了流暢度、人聲相似度和有無雜訊等。隨機選取各實驗所需要的文本進行合成，由我們研究室中的 11 位研究人員參與聆聽，並對各合成語音進行評分，最後將所有分數平均做為結果。

4.1 生成資料集的品質 (Quality of the Generated Dataset)

對 3.1 生成資料集 Generated-multi 與原始資料集 AISHELL3-thirty 和 VCTK-thirty 進行比較，以確保此生成資料集的品質，由表 4 所示，可得生成資料集的分數皆在 4 分以上，表示使用生成的方式依然可獲得不錯的聲音訊號，以此資料集訓練合成器是可行的。

表 4. 資料集的 MOS。基於 VCTK-thirty 和 AISHELL3-thirty 的文本做為生成 Generatedmulti 時的文本。比較生成資料集的語音品質，生成的資料集分數在 4 分以上。

[Table 4. The MOS of the dataset. The text of the Generated dataset is based on the text of VCTK-thirty and AISHELL3-thirty. The MOS score of generated dataset is higher than 4.]

資料集	**MOS**	
	英文	中文
VCTK-thirty	4.46 ± 0.22	-
AISHELL3-thirty	-	4.73 ± 0.17
Generated-multi	4.28 ± 0.31	4.09 ± 0.62

4.2 資料前處理結果 (Results of Data Preprocessing)

由於兩種語言是分開進行資料前處理，因此在 MOS 評分，將中、英文前處理的效果分開進行比較。中文文本在訓練時皆轉為漢語拼音，於是用於評分的文本皆有經過文字轉拼音，以便系統合成，由表 5 可知，文本進行中文斷詞後，MOS 有些微的增加，另外，進行評估數字正規化的文本，為突顯正規化的效果，文本皆選用含有阿拉伯數字的中文句子，正規化後 MOS 分數由 4.02 提高到了 4.45，分數大幅的提升了，由此可知，數字正規化對於文本的重要。在英文結果的部份，選用在英文句中含有連續大寫的文本，用以評估處理頭字語的效果，然而在加入頭字語處理後，MOS 分數由 3.69 增加至 3.99，由結果可知透過前處理能提升合成之品質。

表 5. 有無進行前處理的 MOS。比較有無前處理的語音訊號品質，處理後的品質，皆有所提升。

[Table 5. The MOS of the speech which is with preprocessing or not. The quality of the speech is better after processing the text.]

語言	資料前處理流程	**w/o**	**MOS**
中	中文斷詞	w/o	4.50 ± 0.11
		w/	4.52 ± 0.18
	數字正規化	w/o	4.02 ± 0.40
		w/	4.45 ± 0.20
英	頭字語處理	w/o	3.69 ± 0.20
		w/	3.99 ± 0.15

5. 結論 (Conclusions)

我們建立的中英文語碼轉換語音合成系統，其有相當不錯的表現，透過中、英文的資料前處理大幅提升語音的品質，尤其是中文的數字正規化與英文的頭字語處理，分別由 4.02 上升至 4.45，及 3.69 至 3.99，不過整體系統依舊有進步的空間，因此，未來也將持續改進語碼轉換中，中英文的語音流暢度，以及以建立一個可分離語者資訊，單純學習文本資訊的編碼器為目標，無需再使用生成模型生成的資料集進行訓練，依然可合成多語言語碼轉換的句子。

參考文獻 (References)

Bai, Y., Tao, J., Yi, J., Wen, Z., & Fan, C. (2018). Clmad: A chinese language model adaptation dataset. In *Proceedings of 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)*, 275-279. https://doi.org/10.1109/ISCSLP.2018.8706600

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networksfor efficient and high fidelity speech synthesis. In *Proceedings of Advances in Neural Information Processing Systems*, 33, 17022-17033.

Ma, W.-Y. & Chen, K.-J. (2004). Design of ckip chinese word segmentation system. *International Journal of Asian Language Processing*, *14*(3),235-249.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558

Shi, Y., Bu, H., Xu, X., Zhang, S., & Li, M. (2020). Aishell-3: A multi-speaker mandarin tts corpus and the baselines. arXiv preprint arXiv:2010.11567

J Sun. 2012. Jieba chinese word segmentation tool.

Wang, Y.-W. (2021). Integrating hidden speaker and style information to multi-lingual and codeswitching speech synthesis. (Master's thesis). Retrieved from https://hdl.handle.net/11296/du785x

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E.Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., & Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015.

Yamagishi, J., Veaux, C., & MacDonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92), [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2645..

Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 57-72

Analyzing Discourse Functions with Acoustic Features and Phone Embeddings: Non-lexical Items in Taiwan Mandarin

Pin-Er Chen[*], Yu-Hsiang Tseng[*], Chi-Wei Wang[*],

Fang-Chi Yeh[+], and Shu-Kai Hsieh[*]

Abstract

Non-lexical items are expressive devices used in conversations that are not words but are nevertheless meaningful. These items play crucial roles, such as signaling, turn-taking, or marking stances in interactions. However, as the non-lexical items do not stably correspond to written or phonological forms, past studies tend to focus on studying their acoustic properties, such as pitches and durations. In this paper, we investigate the discourse functions of non-lexical items through their acoustic properties and the phone embeddings extracted from a deep learning model. Firstly, we create a non-lexical item dataset based on the interpellation video clips from Taiwan's Legislative Yuan. Then, we manually identify the non-lexical items and their discourse functions in the videos. Next, we analyze the acoustic properties of those items through statistical modeling and building classifiers based on phone embeddings extracted from a phone recognition model. We show that (1) the discourse functions have significant effects on the acoustic features; and (2) the classifiers built on phone embeddings perform better than the ones on conventional acoustic properties. These results suggest that phone embeddings may reflect the phonetic variations crucial in differentiating the discourse functions of non-lexical items.

Keywords: Non-lexical Item, Discourse Function, Acoustic Property, Acoustic Representation, Pragmatics

[*] National Taiwan University
E-mail: {cckk2913, seantyh}@gmail.com; {r09142007; shukaihsieh}@ntu.edu.tw

[+] National Tsing Hua University
E-mail: fangchiyeh2000@gmail.com

1. Introduction

People's everyday interactions include sounds that are not verbal words in the traditional sense. These sounds, such as sighs, sniffs, and grunts, are used in indexing the turn-taking in dialogues, marking stance, showing affections, and expressing roles and meanings in conversations (Dingemanse, 2020). Examples of these *non-lexical items* are *un-huh* in English as a marker showing understanding and attentiveness, while the single syllable *uh* and *um* act as fillers and disfluency markers (Ward, 2006; Buschmeier *et al.*, 2011).

While these non-lexical items are important linguistically, they pose an interesting challenge to linguistic inquiry. Non-lexical items do not belong to a major word class, and some do not conform to the language's phonological requirements (Keevallik & Ogden, 2020). Moreover, while the phonetic properties of non-lexical items could be generally described, they are nevertheless "phonetically underspecified" (Keating, 1988). For example, in the study of "moan" in board game interactions, Hofstetter (2020) found "moans" involve phonetic properties related to open vowels, irrespective of their frontness, backness, or roundedness. The study suggests that a non-lexical item can not be represented as a single phonetic symbol; instead, it may refer to the vowel space for which we do not have a general phonetic symbol. Some studies, therefore, analyze these items in terms of their acoustic properties: the components' sound (Ward, 2006), the fundamental frequencies, durations, and intensities. (Shan, 2021; Ballier & Chlébowski, 2021).

In contrast to the conventional acoustic property analysis, an alternative approach to analyzing non-lexical items is through the acoustic representations learned by data-driven methods. These methods include deep learning models mapping the audio segments to the latent embedding space from acoustic data in a (self-)supervised fashion (Li *et al.*, 2020; Xu *et al.*, 2021; Baevski *et al.*, 2020). Although the models are not explicitly trained to represent the similarities among phonetic features, studies nonetheless find the audio segments with similar linguistic properties are closer together in the embedding space (Ma *et al.*, 2021; Cormac English *et al.*, 2022; Silfverberg *et al.*, 2021). Therefore, these phonetic representations may already encode the phonetic variability of non-lexical items to reflect their different discourse functions.

This study thus aims to investigate how the acoustic properties contribute to the non-lexical items' discourse functions and how the phone embeddings extracted from the deep learning model help differentiate those functions. The rest of the paper is organized as follows. We first review related works on discourse markers and how they are analyzed with acoustic properties (Sec. 2). Next, we describe our dataset on non-lexical items (Sec. 3) in Taiwan Mandarin, in which we manually identify the items and annotate their discourse functions in interpellation video clips of Taiwan's Legislative Yuan. Finally, based on the dataset, we conduct the acoustic property analysis (Sec. 4) and build classifiers based on the phone embeddings extracted from a deep learning model (Sec. 5). Finally, Section 6 concludes the paper.

2. Related Works

2.1 Discourse Marker

Discourse markers (hereafter, DMs) have received increasing attention since Schiffrin (1987, p. 31) initially defined them as "sequentially dependent elements which bracket units of talk." However, little consensus has been not only on the terminology[1] of DMs but on the classification frameworks. Schiffrin (1987) has proposed that DMs form a category composed of phrases, conjunctions, and interjections, and that they have a part in discourse coherence considering different planes of talk.[2] Additionally, DMs can also serve as identifiers of participation status, speaker's assumptions, or hearer's knowledgement (Schiffrin, 1987; Schwenter, 1996; Fraser, 1999).

Despite that earlier research considered DMs as text-connective items bonding to syntactic structures, Fischer (2006, p. 9) de fined DMs as devices involved in "turn-taking, interpersonal management, topic structure, and participation frameworks." Subsequently, Diewald (2006, 2013) suggested that DMs demonstrate pragmatic functions, manage discourse in a syntactically-independent way, and present their polyfunctionality in discourse (c.f. Fraser, 2009; Hansen, 2006; Németh, 2022).

Although numerous analyses were conducted on the pragmatic functions of DMs, they focused mostly on the associations with semantic senses and syntactic structures (e.g., Aijmer, 2011; Crible, 2017; Ford & Thompson, 1996). That is, studies of the connections between the discourse functions and the phonological information of DMs are relatively few.

2.2 Acoustic Property

The previous works which interwove DMs and their acoustic properties were mainly on the pragmatic-prosodic interface. Shan (2021) and Zhao and Wang (2019) investigated the Mandarin Chinese DMs, 你知道 *ni zhidao* 'you know' and 你不知道 *ni bu zhidao* 'you don't know', respectively. While Shan (2021) analyzed on duration, tempo, intensity, and fundamental frequencies (i.e., pitch, hereinafter F_0), Zhao and Wang (2019) examined the

[1] For instance, discourse marker (Jucker & Ziv, 1998; Schiffrin, 1987); discourse particles (Aijmer, 2002; Fischer, 2006); pragmatic marker (Brinton, 1996); among others

[2] Schiffrin has suggested the five planes of talk: the Exchange structure (ES), Action structure (AS), Ideational structure (IdS), Participation framework (PF), and Information state (InS). More details can be seen in Schiffrin (2005), Maschler and Schiffrin (2015), and Hamilton *et al*. (2015).

speech tempo, mean F_0 frequencies, and pitch accents of the DMs. In general, they have found correlations between the discourse functions and the acoustic properties. Moreover, Tseng *et al.* (2006) have suggested that connectors are predictable from speech prosody; most 'redundant prosodic fillers' are duration-triggered and manifested through narrowed F_0 ranges, whereas 'obligatory discourse markers' are syntax-triggered and manifested through widened F_0 ranges and resets.

The acoustic properties and their relevance to the pragmatic functions of DMs have also been analyzed cross-linguistically (e.g., Cabarrão *et al.*, 2018; Raso & Vieira, 2016; Gonen *et al.*, 2015; Beňuš, 2014). Referring to Wu *et al.* (2021), the phonetic variations of DMs in French are likely to appear in spontaneous speech and undergo phonetic reduction, considering their shorter mean phone duration and a rather centralized vowel space. Additionally, Schubotz *et al.* (2015) investigates the common English construction *you know* in terms of its duration, which is likely to be affected by the residuals of speech rate.

In addition to acoustic properties, past studies also examined the phonetic representations learned with data-driven methods. For example, Silfverberg *et al.* (2021) studied phonological alternations of Finnish consonant gradation with vector representations retrieved from RNN models. Other studies also tried to learn dense vector representations purely from text using grapheme-to-phoneme mappings with CBOW and SkipGram models (O'Neill & Carson-Berndsen, 2019). Notably, recent studies found transformer-based speech processing models (Baevski *et al.*, 2020; Hsu *et al.*, 2021), while not explicitly modeling phonetic properties, encoded the phonetic categorization information in the model representations, such as vowels and consonants, or fricatives and stops (Ma *et al.*, 2021; Cormac English *et al.*, 2022).

Tracing back to the former sections, previous literature on DMs mostly concentrated on their status at the semantic-pragmatic interface. The reviewed acoustic-related research, however, focused on those construction-wise DMs, and not to mention that the analyzed acoustic properties were limited to suprasegmental features, such as pitch and duration. In this case, the potential phonetic-pragmatic interrelationship of non-lexical items is yet to be elaborated.

3. Non-lexical Items Dataset

First, we used four interpellation video clips from Taiwan's Legislative Yuan.[3] Audio tracks were then extracted from the clips, converted into 16 bit WAV format, and resampled with 22kHz sampling rates. The overall data comprise separate interpellation of two male and two female legislators, each ranging 6-8 minutes. The equal number of genders was to balance

[3] The clips were downloaded from the Parliament TV website (https://www.parliamentarytv.org.tw/) and encoded as AAC, H.264

potential gender differences in the utterances.

Secondly, the audio segments of non-lexical items (e.g., *uh*, *em*, and *ho*) were annotated by three native speakers via Praat 6.2.03 (Boersma & Weenink, 2021). Each non-lexical item acquired two tags, one for functional *Role* and one for pragmatic *Meaning*. Referring to Ward (2006), we defined the six candidates of *Role* as follows:

- BACKCHANNEL, which occurs repetitively and shows the agreement of the hearer; it often overlaps the main channel[4] of the utterance.
- CFT (Clause-final token), which occurs in the sentence-final position and ends certain turn of talk.
- DISFLUENCY, which refers to the onset or coda of a word that can hardly be recognized due to its discoursal incompleteness.
- FILLER, which serves as a connector between two sentences or a sentence-initial particle of the speaker.
- RESPONSE, which occurs in the main channel and often indicates a flippant attitude.
- OTHER, which represents the non-lexical item not belonging to the above types.

Similarly, we summarized the following eight candidates for *Meaning*. It is noted that certain non-lexical items may carry multiple pragmatic meanings, and that the candidates below are not mutually exclusive. Thus, one non-lexical item is allowed to be annotated with multiple *Meaning* tags.

- authority. The speaker demonstrates his profession, personal experience, or intention in the speech.
- control. The speaker is in control of knowing exactly what to say or do next.
- concern. The speaker lacks confidence in his own words or tries to show respect to the audience.
- thought. The speaker takes the words (from himself or the other participant) as involving or meriting thought.
- dissatisfaction. The speaker is unsatisfied with his own words, the conversation, or the other participant.
- new information. The speaker wants to express that he has received new information; the

[4] see also Heinz (2003), Li *et al*. (2010), and McNely (2009) among others.

speaker successfully lets the other participant understand the topic of the speech.

- old ground. The speaker is expecting to move on to the next topic since he has already acknowledged the current one.
- neutral.

In sum, a total of 143 non-lexical items produced by the legislators were manually annotated. We then moved on to extract the acoustic properties for the dataset.

4. Acoustic Property Analysis

With the assumption that the discourse functions may encode phonological variations, we illustrated our data collection and the annotation for non-lexical items in Sec. 3. The following sections (4.1 and 4.2) then present the analyses and results of acoustic properties.

4.1 Property Extraction

For each non-lexical item, we retrieved six conventional acoustic properties: mean pitch, duration, F1, F2, F3, and nasality, via customized Praat scripts (Styler, 2017). As formant frequencies construct the vowel space, F1 is determined by the vowel height, F2 is determined by the vowel backness, and F3 is determined by the vowel roundness.[5]

In terms of nasality, it can be quantified by ***a1-p1*** (for high vowels such as [i, u, y]) or ***a1-p0*** values (for non-high vowels such as [a, o, ə, e]). Since most of the annotated non-lexical items are realized and transcribed with non-high vowels, only the ***a1-p0*** values were considered. While ***a1*** stands for the amplitudes (in *dB*) of F1, ***p0*** stands for the amplitude of the nasal peak below F1 (Chen, 1997; Cho *et al*., 2017; Chiu & Lu, 2021).

Subsequently, to build up the most comprehensive acoustic properties, the values of F1, F2, F3 frequencies and ***a1-p0*** amplitude for each annotated non-lexical item were measured at 5 different time-points (i.e., the 10%, 30%, 50%, 70%, 90% time-points within each item interval). The retrieved acoustic data for 715 tokens[6] were processed and modified into machine-readable forms using the pandas package (The Pandas Development Team, 2020) in Python 3.8.9 (Python Core Team, 2021).

The statistical analysis was performed via the lmerTest package (Kuznetsova *et al*., 2017) in R 4.2.1 (R Core Team, 2022). Some factors contain rare categories were therefore re-coded. Specifically in the candidates of *Role*, DISFLUENCY and RESPONSE in were merged into

[5] The higher the F1, the lower the vowel; the higher the F2, the more anterior the vowel; the lower the F3, the rounder the vowel (Flanagan, 1955; Lindblom & Studdert-Kennedy, 1967).

[6] Each 143 annotated non-lexical items were measured at 5 different points, resulting in 715 tokens.

OTHER, considering their extremely few occurrences. As for the candidates of *Meaning*, the items with multiple candidate tags were recoded as complex. The OTHER and complex were set as references in *Role* and *Meaning* factors, respectively. Finally, Box- Cox transformations (Box & Cox, 1964) were applied to each response variable to reduce the non-normalities in the distributions.

4.2 Evaluations

To explore the effect of discourse functions on the acoustic properties, we conduct statistical analyses with linear mixed-effects models and classification tasks with SVM.

Statistical Modeling.

Apart from the two discourse functions (*Role* and *Meaning*), we also take *Transcriptions* into consideration. As *Transcriptions*, annotated for segment identification, reflects the annotators' perception for each non-lexical item, it is likely a control variable that poses significant effects on the properties. Thus, for the evaluation of each acoustic property, we actually compare two models: one full linear mixed-effects model (composed of *Role*, *Meaning*, and *Transcriptions*) as well as one counterpart baseline model (composed of only *Transcriptions*).

Table 1. Model comparisons of linear mixed-effects in different response variables. The comparisons are between the base model, which only contains transcription and random intercepts, and the full model, which additionally includes discourse function predictors. For brevity, only comparison statistics are shown. * p < 0.05, ** p < 0.01, *** p < 0.001.

	Chiq	Df	*p*-value
Duration	83.79	9	<.001 ***
Pitch	124.66	9	<.001 ***
F1	10.12	9	.341
F2	20.32	9	.016 *
F3	7.62	9	.573
Nasality	15.29	9	.083

Table 1 illustrates the sequential (Type I) ANOVA results for the linear mixed-effects models, in which one specific acoustic property is used as the dependent variable. Specifically, the acoustic properties that reach statistical significance among the model comparisons are Duration, Pitch, and F2, suggesting that certain types of roles and meanings present additional effects on acoustic properties, after controlled for the transcriptions. These results imply that the discourse functions show additional effects on the Duration, Pitch, and F2.

Table 2. Parameter estimates of discourse functions in the linear-mixed effect models. The variables of transcriptions are included in all models, but their estimates are not shown in the table for brevity. Response variables are Box-Cox transformed, the parameters are therefore in the transformed scale. * p < 0.05, ** p < 0.01, *** p < 0.001.

	Duration	Pitch	F1	F2	F3	Nasality
(transcriptions)			--			
CFT	0.034	12.04^{***}	35.68	6.28	10169.4	4.03
FILLER	0.042	14.92^{***}	2.67	1.22	10 913.4	5.67
authority	−0.016	3.87^{**}	3.98	2.29^{***}	−6832.3	2.52
control	−0.013	0.16	3.49	7.87	2345.1	0.18
dissatisfaction	−0.052	-10.07^{***}	45.70	3.16^{**}	−9942.1	4.08
neutral	−0.016	0.05	58.17	1.58^{*}	1948.2	0.30
new information	-0.267^{**}	10.17^{***}	−40.21	1.65	−5134.1	−2.71
old ground	−0.003	0.82	−4.51	1.31	3383.3	0.13
thought	-0.288^{***}	−2.36	97.46	1.55	2643.0	2.75

To further examine such possibility, Table 2 compiles the fixed-effect results of the full linear mixed-effects models for the acoustic properties, where the discourse functions[7] are the predictors. We find that both types of discourse functions show the most significance on Pitch, which corresponds to the reviewed works in Sec. 2.2. Yet, only certain types of *Meaning* show correlation with Duration and F2; not to mention the other three acoustic properties (i.e., F1, F3, and Nasality) which do not show any statistical significance.

Support Vector Machines.

Support Vector Machines (SVM) model is implemented for the classification tasks, in which the acoustic properties are used in prediction of discourse functions. As we assume that the discourse functions may reflect in the phonological variations of the non-lexical items, linear models such as SVM are applicable.

We use random 70-30 splits for training and testing data. While the training data comprise 500 tokens, the testing data comprise 215 tokens. A random guessing model, serving as a *the-most-frequent baseline*, is also implemented for comparison. It calculates the frequency

[7] Notice that the aforementioned BACKCHANNEL (as Role) and concern (as Meaning) only exist in the supplementary annotation for those non-lexical items produced by the administrative officers in opposition to the legislators. Data are reserved for the future studies.

distributions of all discourse functions, and then it invariably predicts the most frequent class. We use the accuracy, precision, recall, and F1-score to evaluate the performance of the two models.

Table 3. Evaluation of acoustic models

Role	Acc	Pr	Rc	F1
acoustics	.76	.15	.20	.17
acoustics-base	.76	.15	.20	.17
Meaning	Acc	Pr	Rc	F1
acoustics	.48	.09	.14	.11
acoustics-base	.38	.04	.10	.06

Table 3 shows that both models, based on the acoustic properties, find it harder to predict *Meaning* than *Role*. Specifically, the acoustics achieved slightly better accuracy (.48) and precision (.09) than the baseline (.38 and .04). In the prediction of *Role*, however, the performance of the models was very similar. It implies that the acoustics in fact does not acquire much advantage in predicting discourse functions. This observation is consistent with the results of the previous liner mixed-effects model, in which we found few correlations between the acoustic properties and the discourse functions. Therefore, we attempt to find other presentations of phonological variations that may better capture the candidates of discourse functions with higher accuracy.

5. Phone Embeddings

As the conventional acoustic properties did not show promising results of capturing the discourse functions, we reached out to phonetic vector representations, in which the phonological variations of non-lexical items might be encoded.

Instead of the common end-to-end models trained on waveforms and language-specific transcriptions in ASR tasks, we chose the *Allosaurus* model by Li *et al*. (2020)[8] for retrieving the phone embeddings. Specifically, the Allosaurus is a universal phone recognizer integrating an ASR encoder with an allophone layer, in which language-independent phone distributions are directly recognized and mapped into language-dependent phoneme distributions.

We first examine the phone embeddings learned by the phone recognition model. In the video clips collected in Section 3, the model automatically identifies 29,218 phones in the conversations. To investigate the phone organizations in the embedding space, we then extract

[8] https://github.com/xinjli/allosaurus

the bi-LSTM representations[9] with which model predicts the phones as phone embeddings. Next, we average these embeddings by their predicted phones and obtain 34 phone centroids in the embedding space. We follow the literature (Cormac English *et al.*, 2022) and conduct hierarchical clustering with Ward linkage based on the Euclidean distances between the centroids. The clustering results are shown in Figure 1a and Figure 1b. We not only observe clear clusters of vowels and consonants but observe that the fricatives and stops tend to be close to each other with similar phonetic properties. The patterns suggest that the phone embeddings might reflect the phonetic variations in our conversation data.

Moreover, we inspect the clustering structure of recognized phones that occurred in the non-lexical items. Figure 1c shows the two-dimensional t-SNE (Pedregosa *et al.*, 2011) visualization of the 640-dimension phone embeddings obtained from Allosaurus. The same phones tend to form distinct clusters, and the general distinction between vowels and consonants is still observed in the figure. It indicates that the embeddings may represent their corresponding phonetic properties. As Li *et al.* (2020) have shown in their studies, Allosaurus has the advantage of multilingual phone recognition and involves more phonological knowledge. It is thus appropriate for us to leverage these phone embeddings, by which the discourse functions of non-lexical items may be encoded.

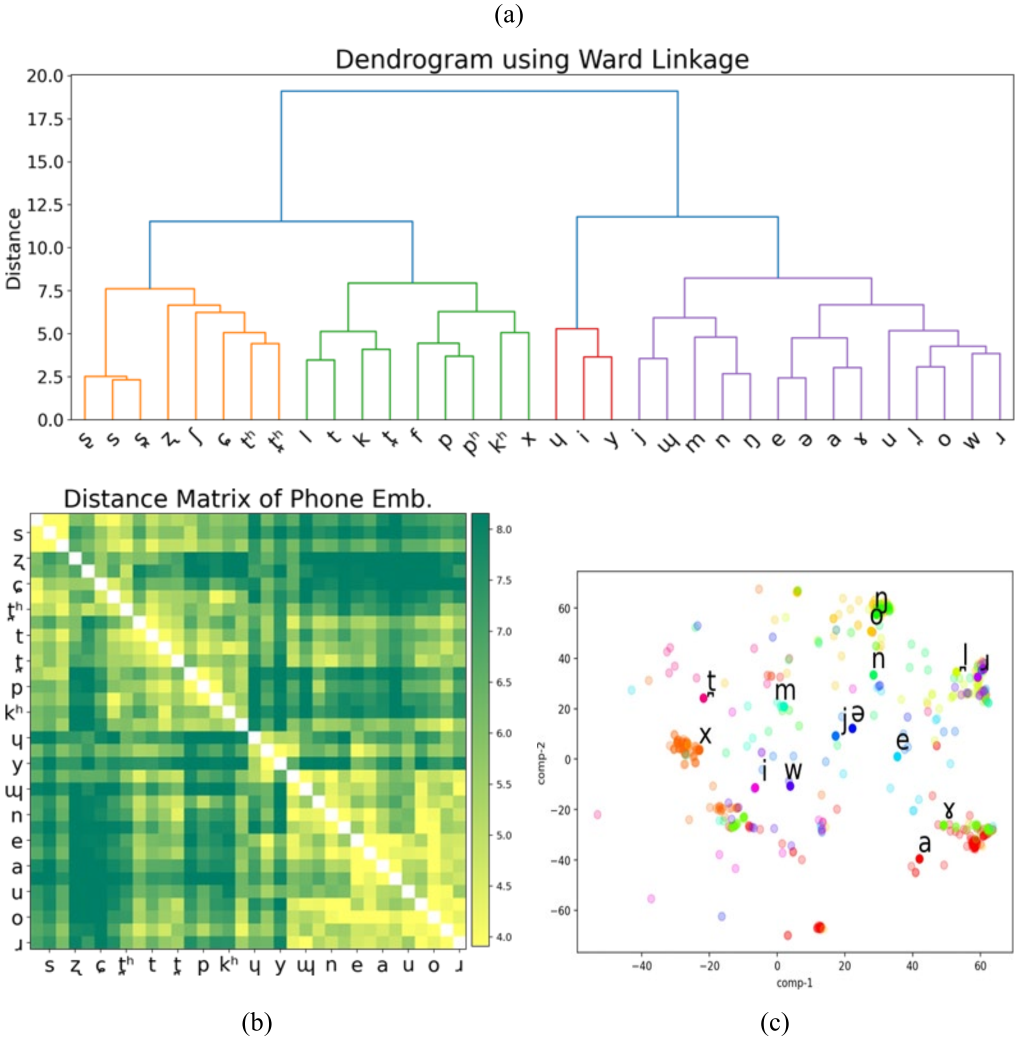
5.1 Classification Task

The output data by Allosaurus (i.e., the phone embeddings and phoneme transcriptions) are aligned with our annotations of discourse functions for non-lexical items. It is noted that only the phoneme, whose timestamp matches the 715 tokens of non-lexical items, are kept for the classification tasks. The data is split randomly 70-30 into training and testing datasets as in Section **4.2**.

We also implement a linear SVM model and a random guessing model serving as a *the-most-frequent baseline* for the classification tasks.[10] The only difference here is that we replace use the acoustic properties with the phone embedding vectors to predict the candidates of the discourse functions.

[9] Referring to the comments from the reviewers, the bi-LSTM representations are used as the phone embeddings considering their better performance than the other representations (i.e., the 40-dimension MFCCs and the phone logits) generated by Allosaurus.

[10] Regarding the comments from the reviewers, the linear SVM model and the model baseline are adopted to not only display the data distributions but highlight the results of Allosaurus, as we mainly focus on whether the phone representations really help us explore non-lexical items. Based on the results, we did find the the model using phonetic realizations performs better in predicting the discourse functions, and we expect future research to develop better representations and state-of-the-art models that allow us to describe non-lexical items more appropriately.



(a)

(b)

(c)

Figure 1. (a) The dendrogram of the hierarchical clustering with Ward linkage. The links are color-coded for visual references. Generally, the top left and right branches loosely correspond to consonants and (semi-)vowels. The leftmost branch (orange) are mostly fricatives (e.g., s, ʂ, ɕ); the one on the right (green) includes stops (e.g., k, t, p). (b) The distance matrix shows a consistent pattern with the one in the dendrogram. (c) The t-SNE projection of the phones in non-lexical items. Only the most-frequent 15 phones are shown for clarity. IPA symbols mark the median points of each category.

5.2 Evaluation Results

As shown in the upper part of Table 4, phone emb. stands out with the highest accuracy (.92) and precision (.96) in prediction of *Role*. While baseline presents the accuracy of .78, the acoustic models (see Table 3) show even lower accuracies (.76) and precision (.15). As for predicting *Meaning*, phone emb. Significantly outperforms its baseline and remains the highest in accuracy (.77) and precision (.84) among all models. In general, phone emb. presents superior performance than the other models in prediction of both discourse functions.

Table 4. Evaluation of classifiers based on phone embeddings

Role	Acc	Pr	Rc	F1
phone emb.	.92	.96	.87	.91
baseline	.78	.16	.20	.18
Meaning	Acc	Pr	Rc	F1
phone emb.	.77	.84	.68	.72
baseline	.42	.05	.11	.07

Moreover, both models (i.e., acoustics and phone emb.) are better at predicting *Role* than *Meaning*, likely due to the fact that *Meaning* comprises more types of candidates and internally more equal distribution. In this case, the gap between the accuracies of phone emb. (i.e., between .92 and .77) is still the smallest among the models. This suggests that our model is better at capturing the discourse functions by using the phone embeddings, the phonetic realizations, than the statistical acoustic properties.

6. Conclusions

This paper focuses on the phonetic-pragmatic interrelationship of non-lexical discourse markers in Taiwan Mandarin. As we assume that the discourse functions may be captured by the phonological variations, we firstly analyzed on the common acoustic properties (i.e., duration, nasality, mean pitch, F1, F2, and F3), followed by the classification tasks considering the 640d-phone embeddings. In comparison with the conventional acoustic properties, the model using phonetic realizations performs better in prediction of the functional *Role* and pragmatic *Meaning* of the non-lexical items. The result is consistent with our hypotheses that the phonetic realizations, embeddings via deep learning, encode certain phonological variations of non-lexical items and correlate with their discourse functions.

Acknowledgments

We sincerely thank the three native-speakers for the annotation work and the two anonymous reviewers whose comments and suggestions have helped us clarify the technical details in the

paper.

References

Aijmer, K. (2002). *English discourse particles: evidence from a corpus*. Number 10 in Studies in corpus linguistics. Benjamins. https://doi.org/10.1075/scl.10

Aijmer, L. (2011). Well i'm not sure i think…the use of well by non-native speakers. *International Journal of Corpus Linguistics*, 16, 231-254. https://doi.org/10.1075/ijcl.16.2.04aij

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449-12460.

Ballier, N., & Chlébowski, A. (2021).“see what i mean, huh?”evaluating visual inspection of f0 tracking in nasal grunts. In *Proceedings of Interspeech 2021*, 376-380.

Beňuš, S. (2014). Conversational entrainment in the use of discourse markers. In *Recent Advances of Neural Network Models and Applications*, 345-352. https://doi.org/10.1007/978-3-319-04129-2_34

Boersma, P. & Weenink, D. (2021). Praat: Doing phonetics by computer [computer program] version 6.2.03, retrieved 23 august 2022 from http://www.praat.org/.

Box, G. E. P. & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211-243.

Brinton, L.J. (1996). *Pragmatic markers in English: grammaticalization and discourse functions*. Number 19 in Topics in English linguistics. Mouton de Gruyter.

Buschmeier, H., Malisz, Z., Włodarczak, M., Kopp, S., & Wagner, P. (2011). Are you sure you're paying attention?'-uh-huh'communicating understanding as a marker of attentiveness. In *Proceedings of Interspeech 2021*, 2057-2060.

Cabarrão, V., Moniz, H., Batista, F., Ferreira, J., Trancoso, I., & Mata, A.I. (2018). Cross-domain analysis of discourse markers in european portuguese. *Dialogue & Discourse, 9*(1), 79-106. https://doi.org/10.5087/dad.2018.103

Chen, M.Y. (1997). Acoustic correlates of english and french nasalized vowels. *The Journal of the Acoustical Society of America*, *102*(4), 2360-2370. https://doi.org/10.1121/1.419620

Chiu, C., & Lu, Y.-A. (2021). Articulatory evidence for the syllable-final nasal merging in taiwan mandarin. *Language and Speech*, *64*(4), 771-789. https://doi.org/10.1177/0023830920948084

Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in english. *Journal of Phonetics*, 64, 71-89. https://doi.org/10.1016/j.wocn.2016.12.003

Cormac English, P., Kelleher, J.D., & Carson-Berndsen, J. (2022). Domain-informed probing of wav2vec 2.0 embeddings for phonetic features. In *Proceedings of the 19th*

SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, 83-91. https://doi.org/10.18653/v1/2022.sigmorphon-1.9

Crible, L. (2017). Discourse Markers and (Dis) fluency across Registers. Ph.D. thesis, Université de Berne.

Diewald, G. (2006). Discourse particles and modal particles as grammatical elements. *Approaches to Discourse Particles*, 403-425.

Diewald, G. (2013). "same same but different"- modal particles, discourse markers and the art (and purpose) of categorization. *Discourse Markers and Modal Particles.Categorization and Description*, 19-46.

Dingemanse, M. (2020). Between sound and speech: Liminal signs in interaction. *Research on Language and Social Interaction*, *53*(1), 188-196. https://doi.org/10.1080/08351813.2020.1712967

Fischer, K. (2006). Towards an understand- ing of the spectrum of approaches to discourse particles: introduction. In Fischer, editor, *Approaches to discourse particles, number 1 in Studies in pragmatics*, 1-20. Elsevier.

Flanagan, J.L. (1955). A difference limen for vowel formant frequency. *The journal of the Acoustical Society of America*, *27*(3), 613-617. https://doi.org/10.1121/1.1907979

Ford, C. & Thompson, S. (1996). Interactional units in conversation: Syntactic, intonational and pragmatic resources. *Interaction and grammar*, (pp.134-184). Cambridge University Press.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, *31*(7), 931-952. http://dx.doi.org/10.1016/S0378-2166(98)00101-5

Fraser, B. (2009). An account of discourse markers. *International Review of Pragmatics*, 1, 293-320.

Gonen, E., Livnat, Z., & Amir, N. (2015). The discourse marker axshav ('now') in spontaneous spoken hebrew: Discursive and prosodic features. *Journal of Pragmatics*, 89, 69-84. https://doi.org/10.1016/j.pragma.2015.09.005

Hamilton, H.E., Tannen, D., & Schiffrin, D. (2015). *The handbook of discourse analysis*. John Wiley & Sons.

Hansen, M.-B. M. (2006). *A dynamic polysemy approach to the lexical semantics of discourse markers: (with an exemplary analysis of French toujours,)* number 1 in Studies in pragmatics, (pp. 21-41). Elsevier.

Heinz, B. 2(003). Backchannel responses as strategic responses in bilingual speakers'conversations. *Journal of pragmatics*, *35*(7), 1113- 1142. https://doi.org/10.1016/S0378-2166(02)00190-X

Hofstetter, E. (2020). Nonlexical "moans" : Response cries in board game interactions. *Research on Language and Social Interaction*, *53*(1), 42-65. https://doi.org/10.1080/08351813.2020.1712964

Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A.. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451-3460. https://doi.org/10.1109/TASLP.2021.3122291

Jucker, A.H. & Ziv, Y. (1998). *Discourse Markers: Descriptions and Theory*. Pragmatics & Beyond New series, 57. Lightning Source Incorporated.

Keating, P.A. (1988). Underspecification in phonetics. *Phonology*, *5*(2), 275-292.

Keevallik, L. & Ogden, R. (2020). Sounds on the margins of language at the heart of interaction. *Research on Language and Social Interaction*, *53*(1), 1-18. https://doi.org/10.1080/08351813.2020.1712961

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest pack- age: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Li, H. Z., Cui, Y., & Wang, Z. (2010). Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. *International journal of psychological studies*, *2*(1), 25-37. https://doi.org/10.5539/ijps.v2n1p25

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D., Neubig, G., Black, A., & Metze, F. (2020). Universal phone recognition with a multilingual allophone system. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8249-8253.

Lindblom, B.E.F. & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical society of America*, *42*(4), 830-843. https://doi.org/10.1121/1.1910655

Ma, D., Ryant, N., & Liberman, M. (2021). Probing acoustic representations for phonetic properties. In *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, 311-315.

Maschler, Y. & Schiffrin, D. (2015). Discourse markers language, meaning, and context. *The handbook of discourse analysis*, 189-221.

McNely, B. (2009). Backchannel persistence and collaborative meaning-making. In *Proceedings of the 27th ACM international conference on Design of communication*, 297-304. https://doi.org/10.1145/1621995.1622053

Németh, Z. (2022). The conversation- organising role of the non-lexical sound öö in hungarian. *Journal of Pragmatics*, 194, 23-35. https://doi.org/10.1016/j.pragma.2022.04.002.

O'Neill, E. & Carson-Berndsen, J. (2019). The effect of phoneme distribution on perceptual similarity in english. In *Proceedings of The 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*, 1941-1945.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Python Core Team. (2021). *Python: A dynamic, open source programming language*. Python Software Foundation. Python version 3.8.9.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raso, T. & Vieira, M. (2016). A description of dialogic units/discourse markers in spontaneous speech corpora based on phonetic parameters. *CHIMERA Romance corpora and linguistic studies*, 3, 221-249.

Schiffrin, D. (1987). *Discourse Markers*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Schiffrin, D. (2005). Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 54-5.

Schubotz, L., Oostdijk, N., & Ernestus, M. (2015). Y'know vs. you know: What phonetic reduction can tell us about pragmatic function. In *S. Lestrade, P. de Swart & L. Hogeweg (Eds.). Addenda. Artikelen voor Ad Foolen.*, (pp 261-280). Nijmegen: Radboud Universiteit Nijmegen.

Schwenter, S. A. (1996). Some reflections on o sea: A discourse marker in spanish. *Journal of Pragmatics*, *25*(6), 855-874. https://doi.org/10.1016/0378-2166(95)00023-2

Shan, Y. (2021). Investigating the interaction between prosody and pragmatics quantitatively: A case study of the chinese discourse marker ni zhidao ("you know"). *Frontiers in psychology*, 12. https://doi.org/10.3389/fpsyg.2021.716791

Silfverberg, M., Tyers, F., Nicolai, G., & Hulden, M. (2021). Do RNN states encode abstract phonological alternations? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 5501-5513. https://doi.org/10.18653/v1/2021.naacl-main.435

Styler, W. (2017). On the acoustical features of vowel nasality in english and french. *The Journal of the Acoustical Society of America*, *142*(4), 2469-2482. https://doi.org/10.1121/1.5008854.

The Pandas Development Team. (2020). Pandas-dev/pandas: Pandas.

Tseng, C.-y., Su, Z.-y., Chang, C.-H., & Tai, C.-h. (2006). Prosodic fillers and discourse markers–discourse prosody and text prediction. In *Proceedings of 2nd International Symposium on Tonal Aspects of Languages (TAL 2006)*, 108-113.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, *14*(1), 129-182. https://doi.org/10.1075/pc.14.1.08war

Wu, Y., Hutin, M., Vasilescu, I., Lamel, L., Adda-Decker, M., Degand, L. (2021). Fine phonetic details for discourse marker disambiguation: a corpus-based investigation. In *The 10th Workshop on Disfluency in Spontaneous Speech (DiSS 2021)*, 25-27.

Xu, Q., Baevski, A., & Auli, M. (2021). Simple and effective zero-shot cross-lingual phoneme recognition. arXiv preprint arXiv:2109.11680.

Zhao, B. & Wang, G. (2019). The prosodic features of the mandarin discourse marker nibuzhidao under different functions. In *Proceedings of the 3rd International Conference on Art Design, Language, and Humanities*, 203-209.

Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 73-88

A Chinese Dimensional Valence-Arousal-Irony Detection on Sentence-level and Context-level Using Deep Learning Model

Jheng-Long Wu*, Sheng-Wei Huang*, Wei-Yi Chung*,

Yu-Hsuan Wu* and Chen-Chia Yu+

Abstract

Chinese multi-dimensional sentiment detection task is a big challenge with a great influence on semantic understanding. Irony is one of the sentiment analysis and the datasets established in the previous studies usually determine whether a sentence belongs to irony and its intensity. However, the lack of other sentimental features makes this kind of datasets very limited in many applications. Irony has a humorous effect in dialogues, useful sentimental features should be considered while constructing the dataset. Ironic sentences can be defined as sentences in which the true meaning is the opposite of the literal meaning. To understand the true meaning of a ironic sentence, the contextual information is needed. In summary, a dataset that includes dimensional sentiment intensities and context of ironic sentences allows researchers to better understand ironic sentences. The paper creates an extended NTU irony corpus, which includes valence, arousal and irony intensities on the sentence-level; and valence and arousal intensities on the context-level, which called the Chinese Dimensional Valence-Arousal-Irony (CDVAI) dataset. The paper analyzes the difference of CDVAI annotation results between annotators, and uses a lot of deep learning models to evaluate the prediction performances of CDVAI dataset.

Keywords: Irony Annotation, Dimensional Valence-Arousal-Irony, Sentiment Analysis, Deep Learning.

* Department of Data Science, Soochow University
E-mail: jlwu@gm.scu.edu.tw; {iwihwung11,hwork0511,ikaroskasane}@gmail.com
The author for correspondence is Jheng-Long Wu.

+ The University of Edinburgh School of PPLS
E-mail: alisonyu119@gmail.com

1. Introduction

There are billions of posts on various kinds of forums and social media every day, which shows the exchange of opinions online are high in action and frequency. Human conversations are complex behaviors, because opinions by the people may use direct or indirect presentation sentences. Therefore, the semantic understanding of online opinions is more complicated. In addition, metaphors, irony, sarcasm, etc. also widely appear on online social media. These kinds of expressions cause challenges for natural language understanding (NLU) and natural language processing (NLP). Joshi *et al*. (2018) has reviewed the irony detection problem. Although most of the literature lacks a clear and consistent definition of irony, they found that the most common feature of ironic sentences is the inversion of the literal meaning and true meaning. For example: "Great, it's raining, but I didn't bring an umbrella....", the literal meaning is that raining without an umbrella is a great situation. However, the context "it's raining, but I didn't bring an umbrella..." shows a negative emotion, contrasted with "Great" which is a positive emotion. This emotional contrast caused the semantic turn from negative to positive, which enables the expression of irony. In Chinese irony, the contrast between positive and negative emotions is often used to indicate the difference between sentences and contexts. This emotional contrast is often used to achieve ironic expressions (Veale & Hao, 2010). According to the grammatical structure mentioned above, this study argues that context must be considered to match the characteristics of ironic sentences to improve the performance on irony detection task. The work in sentiment analysis of irony has turned to the study of ironic language features (Colston, 2019). With the development of machine learning, some studies have begun to use machine learning methods to predict the intensity of irony (Chia *et al*., 2021; Dimovska *et al*., 2018). However, most of them still predict irony using whole sentences instead of considering context as mentioned above.

To improve machine learning performance of detecting ironic sentences, some studies proposed to annotate grammatical structural features or use feature selection to screen important irony spans in the English language (Kumar & Harish, 2019). Long *et al*. (2019) proposed the usage of capitalized words as a hint of irony in English. However, capitalization does not exist in Chinese so the capitalization is not suitable for use. In conclusion, while the grammatical structure of irony has been thoroughly studied in English, it is not appropriate to apply it directly to Chinese. Although some studies summarized Chinese irony grammatical structures (Jia *et al*., 2019), there are few datasets annotated based on these rules. Since irony has a humorous effect in the conversation processes, the paper considers irony detection as a sentiment detection task. Therefore, considering the multi-dimensional Valence-Arousal-Irony (VAI) intensity for irony sentences and context is more possible to identify the true meaning of ironic sentences and the emotional state of the social media user.

Based on Tang's (Tang & Chen, 2014) open data on irony sentences, the paper proposes to extend sentence-level intensity of valence, arousal and irony, and context-level intensity of valence and arousal. This annotation method provides a way to judge the difference in context and semantics in irony sentences. By quantifying emotional indicators, the pattern of sentiment while using ironic sentence can be more easily understood. This augmented CDVAI dataset is the first dataset to do sentiment annotations for irony context.

Furthermore, this paper proposes deep learning models based on the pretrained Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2018) model to learn the dimensional VAI on the ironic sentences and dimensional VA on ironic contexts. The paper uses pre-trained BERT to extract hidden features of sentence or context, respectively. Then there are three methods to combine hidden features and predict VAI scores of sentence-level or VA scores of contexts-level: (1) using a linear layer to predict VAI and VA, respectively; (2) summing two hidden features from two encoders of sentence and context. (3) Concatenating two hidden features from two encoders of sentence and context. Furthermore, the paper constructs a token classification model to automatically predict the position of context. Then the predicted positions of context are used to replace the origin positions of context, and predict VAI scores of sentence-level or VA scores of contexts-level.

2. Related Works

Because of different research perspectives, the definition of irony is often adjusted. However, previous studies summarized a basic consensus in the process of exploring ironic sentences. "Irony is an expression in which the true meaning is the opposite of the literal meaning" (Li & Huang, 2020). Based on the above, the most common feature of irony is metaphor, which can make the literal meaning opposite to the true meaning of the sentence that the commenter wants to express. The form of ironic sentence can be expressed as using keywords of exaggeration with positive emotions to describe context with negative emotions. This emotional contrast makes the sentences have an ironic effect (Veale & Hao, 2010). Li *et al.* (2020) proposed an irony identification program (IIP) based on the grammatical structure of ironic sentences, which supports future studies to identify whether a sentence is ironic. The above research provides support for the definition of irony in the paper.

Irony sentences are not usually used in official documents. Thanks to the prevalence of social media, many ironic sentences have been posted online which has led researchers to collect and analyze ironic sentences on social media platforms (Lestari, 2019). Among the studies related to irony detection or sentiment detection. There are very few corpuses including VAI indicators. The possible reasons are that irony detection is not traditionally attributed to the domain of sentiment detection. However, irony has a humorous effect in conversation, which can result in specific emotional patterns for the writer and reader. Therefore, the paper considers irony detection as an emotion detection task. But most of the existing corpus are only included valence and arousal (VA) or only include irony (I) indicator.

Recent studies have collected data on social media to build corpus. Preoţiuc-Pietro *et al.* (2016) used the Likert nine-point scale to annotate VA indicators for Facebook posts. They found that there is high correlation between VA. Bosco *et al.* (2013) annotate irony and emotional expression for Twitter tweets to establish the Senti-TUT corpus. Their corpus includes positive, negative emotions and irony, which considers the concept of valence. Ghosh *et al.* (2015) annotate figurative language such as irony, satire, and metaphor on a 11-point scale at SemEval-2015 Task 11. In addition, there are many constructions of VA or I corpus, but there are very few studies that comprehensively considers VAI.

The necessity of considering VAI indicators simultaneously is that there are correlations among the three indicators. Effects of irony on human emotions in conversation was found in the study of Pfeifer (Pfeifer & Lai, 2021). People who use irony are in a less negative and less excited state of mind. Existing VAI corpus was constructed by Xie *et al.* (2021). They found that stronger irony expressions may have lower valence (more negative) and higher arousal levels, respectively. However, since context is important information to construct ironic sentences which their study didn't consider. The biggest difference between Xie *et al.* (2022) and the paper is that the context is considered and annotated with VA score. To conclude, the above study proves that it is necessary to consider VAI together because of the correlations in these three indicators.

Irony Corpus built in Chinese such as Xiang *et al.* (2020) proposed Ciron dataset. Their dataset contains 8.7K Weibo posts. However, they annotated the intensity of ironic sentences in the corpus without considering context and other sentiment indicators. Existing corpus that include irony sentences and context is NTU Irony Corpus (Tang & Chen, 2014), but their corpus without other sentiment indicators. Lack of consideration of sentiment indicators is impossible to understand clearly on the emotional transitions and semantic changes in the sentences. Therefore, the corpus provided in this paper has a greater advantage in understanding the structure of ironic sentences and sentiment patterns.

In terms of the irony detection model, Rangwani *et al.* (2018) considered emojis on Twitter as a feature when annotating ironic sentences. They use Convolutional Neural Network (CNN) to pre-train the emoji and connect to a XGBoost model for classification. Naseem *et al.* (2020) proposed a T-Dice model based on the frame of the Transformer to detect valence and irony, then connected to Bi-directional Long Short-Term Memory (Bi-LSTM) to classify emotions. The accuracy of their model's prediction results exceeded the state-of-the-art methods of the time. Xiang *et al.* (2020) found that the performance of BERT is better than GRU in their experimental results on the Ciron dataset they built. Lu *et al.* (2020) improved the Bi-GRU model based on BERT in the Chinese sentiment analysis task to achieve the best results. To sum up, in recent years, no matter in sentiment or irony detection tasks. Models that can connect the information of the entire sentence have achieved better results. Furthermore, models with attention mechanisms such as BERT or based on Transformers frame can make the model achieve better results. In summary, this paper will base on BERT to detect the VAI score of sentences and the VA score of the contexts.

3. CDVAI Dataset

The paper proposes to extend the NTU irony corpus to a Chinese dimensional valence-arousal-irony called CDVAI. The NTU irony corpus is the only Chinese corpus that includes ironic sentences and contexts. Therefore, the paper proposes to annotate the VAI intensity of the sentence-level and the VA intensity of the context-level, respectively. Li and Huang (2020) analyzed the sentence structure of Chinese irony based on the existing corpus. They summarized that context is an important information for detecting irony. Based on the sentence structure in the NTU irony corpus and their findings, the paper defines irony as "irony is an expression in which the true meaning is the opposite of the literal meaning." Context is the true meaning of the sentence (usually a negative description), while ironic keywords (usually positive descriptions) can make the literal meaning contrary to the context.

3.1 Dimensional VAI annotation

The paper annotated irony sentences with VAI intensity, and irony contexts with VA intensity. Every indicator was rated from 1 to 5 points. The detailed annotation judgement as follow:

- Valence: Lower valence scores indicate more negative emotions (1-2 points), whereas higher valence scores indicate more positive emotions (4-5 points), and 3 indicate neutral emotions, or inability to judge.
- Arousal: A score of 1 indicates the sentence is close to an objective description, or difficult to judge whether the sentence expresses excitement. A score of 2 indicates that the annotator can feel the low excitement expressed in the sentence, but there is no emotion word such as sad, annoyed, lost, happy, etc. in the sentence. A score of 3 and above indicate the annotator can feel the medium excitement expressed in the sentence, or with explicit emotional words or phrases to clearly describe the emotional state. A score of 4 indicates that the annotator can clearly feel strong excitement expressed in the sentence, such as madness, rage, etc. Furthermore, the sentence may contain violent words, such as aggressive language. A score of 5 indicates in addition to strong excitement, words with discrimination, hated, or words with obvious manic emotions. For example: “Great, the class report is going to be in the same group with that pathetic nerd!”.
- Irony: The annotator reads a sentence and judges whether the true meaning is the opposite of the literal meaning. Most of the sentences in NTU irony corpus use negative

descriptions as the context, and positive descriptions as the keywords to express irony. Irony intensity will be determined according to the gap between the positive intensity of irony keywords and the negative intensity of context. In this paper, the positive intensity of various ironic keywords appearing in the corpus is summarized as: wonderful > great > very good > good. A special case is "it's fine to get worse!", the true meaning in this case is the situation is already bad but the commenter doesn't want the situation to get worse, the ironic keywords "it's fine to" makes the literal meaning opposite to the true meaning. However, this case means the situation is already bad so the gap between positive intensity of irony keywords and the negative intensity of context is small. The larger the gap between the positive intensity of the ironic keyword and the negative intensity of the context, the higher the score of irony, and vice versa. A score of 1 indicates that the gap is very small, or the context is close to an objective description, which leads to hard judgement. For example: "Good, it's raining.". A score of 2 indicates that there is a small gap between ironic keywords and context. A score of 3 indicates that there is a moderate gap between the ironic keywords and the context. A score of 4 indicates that there is a big gap between the ironic keywords and the context. A score of 5 indicates that there is a great gap between ironic keywords and context. The sentence may contain discriminatory or morally unacceptable metaphors, such as sexual innuendo.

3.2 Annotated result analysis

There are 1004 sentences in NTU Irony Corpus, and 843 sentences with an ironic context. Each sentence was annotated by three annotators. The annotators consist of postgraduate students and an undergraduate student, all of them are native Chinese speakers and ages between 20 and 25. Due to the subjective judgement bias of different annotators, the paper uses the average of 3 annotators as the gold standard. The paper using scores to annotate VAI is more reasonable. Human perception of emotional intensity is closer to continuous scores than classification. The meaning of the annotating criterion in the paper is to concretize the definition of intensity of VAI and set the standard score line. Continued from above, the traditional method which is used to evaluate the agreement between annotators such as Cohen's kappa doesn't conform to the hypothesis of the paper. So, the paper uses mean absolute error (MAE) to evaluate the annotation consistency. At the sentence level, the MAE of the three annotators ranged from 0.05 to 0.31 in valence, 0.25 to 0.41 in arousal, 0.22 to 0.56 in irony. At the context level, the MAE of the three annotators ranged from 0.07 to 0.4 in valence, 0.15 to 0.65 in arousal. From the above, the MAE difference between of the three annotators is very small, which proves that the annotating is effective.

● **For example:**

Score of a sentence: valence: 1, arousal: 5, irony: 4

Score of a context: valence: 1, arousal: 5

Sentence: *"很好 (applause)工廠的廠務小姐已經來上班好多好多年了,跟我說她不會用 outlook 發會議通知!! ㄍㄋㄋ勒!!妳的薪水也給我我就幫你發通知!!"* (*"Very good (applause) The factory manager of the factory has been coming to work for many years. She told me that she doesn't know how to use Outlook to send meeting notices!! mother fucker!! Give me your salary and I will send the notices for you!!"*)

Context: *"工廠的廠務小姐已經來上班好多好多年了,跟我說她不會用 outlook 發會議通知!!"* (*"The factory manager of the factory has been coming to work for many years. She told me that she doesn't know how to use Outlook to send meeting notices!!"*)

Judgement: First, in terms of judging the score of valences, there are extremely negative emotions in this sentence such as "mother fucker!! Give me your salary and I will send a notice for you!!". Clearly, the emotions expressed by the swear words and complaints in the sentence are highly negative. Thus, valence is given a score of 1. In terms of judging the score of arousals, we can notice the abuse language and feel the emotion of manic. Thus, arousal is given a score of 5 points. In terms of judging the score of irony, the irony keyword "very good" is a weak positive description. However, according to the description of the sentence, the incident described in the context caused serious discomfort and negative emotions to the commenter. As we can see, there is a big gap between positive irony keyword and negative describe of context. Besides that, the sentence also contains sarcasm spans, such as "Give me your salary and I will send a notice for you.", so it is given a high score of 4 points in irony.

3.3 Statistics of Annotated Result

Table 1 shows the annotated result of CDVAI dataset in different levels and sentiment. Since the dataset is mainly ironic sentences, which results in valence scores that are all low (negative emotion) at sentence-level. While few valence scores of contexts are neutral at context-level. The sentences corresponding to these kinds of contexts often show low scores in valence and irony. There are many sentences containing emotions, which can be observed in the arousal scores centered on points 2, 3 and 4. The score of arousals at context-level is distributed to a lower score than sentence-level. The reason is that irony keywords usually have exaggerated expressions, resulting in a higher arousal. The distribution of the score is like arousal. Gap between positive irony keyword and negative context are usually small, which can be observed in the irony scores centered on points 1, 2 and 3.

Table 1. Score frequency of all sentiments.

Level	Sentiment	0	1	2	3	4	5
Sentence	Valence	0	380	624	0	0	0
	Arousal	0	60	406	369	150	46
	Irony	0	181	428	310	75	20
Context	Valence	0	302	516	25	0	0
	Arousal	56	279	264	161	76	26

4. Model Performance Evaluation

To validate the annotation consistency and the validity of the proposed CDVAI dataset in the paper, the paper constructs deep learning models to predict the VAI score of sentence-level and VA scores of context-level. Table 2 shows the general statistics of CDVAI dataset. The paper uses stratified sampling to split the dataset into training, validation, and the testing set. The ratio of training set and testing set is 7:3, and validation set is split from training set which ratio is 9:1.

Table 2. Statistics of the proposed CDVAI dataset.

dataset	Sentence-level	Context-level
Training set	632	531
Validation set	71	59
Testing set	301	253

4.1 Prediction Model

This paper uses pre-trained BERT models as an encoder to extract hidden state of sentences and contexts, then connected to a linear layer to perform score prediction. There are three methods to obtain final hidden features such as (1) M1: After input sentence and context into the encoder, the hidden features of the sentence are used to predict sentence VAI score through a linear layer. The hidden features of context are used to predict context VA score. (2) M2: The position of the context in the sentence has been located. After input sentence and context into the encoder, the hidden features at the context position are summed, then predict sentence VAI and context VA scores. (3) M3: After input sentence and context into the encoder, concatenate two hidden features of sentence and context then predict sentence VAI and context VA scores. Above processes are the first part of the experiment in this paper. The second part of the experiment, the paper attempted to create a model to predict context span automatically. The paper uses the pre-trained BERT models as encoders, and then the output of encoder with linear layer to predict the span of context in the sentence. Finally, the predicting context will replace the origin context in the first part of the experiment, with the predicted context of the proposed model, then repeat the process of the first experiment.

The paper compares BERT models pre-trained on a Chinese corpus to find the best results. The pre-trained models are as follow:

- PM1: *hfl/chinese-macbert-base* uses Wikipedia simplified and traditional Chinese as the corpus to train the model. (Cui *et al.*, 2020)
- PM2: *shibing624/macbert4csc-base-chinese* using the SIGHAN typo correction corpus to train the model. (Cui *et al.*, 2020)
- PM3: *uer/chinese_roberta_L-4_H-256* uses UER toolkit and CLUECorpus2020 to train the model. (Turc *et al.*, 2019)
- PM4: *IDEA-CCNL/Erlangshen-Ubert-110M-Chinese* uses datasets from a variety of tasks for open-source UBERT. (Wang *et al.*, 2022)
- PM5: *IDEA-CCNL/Erlangshen-Ubert-330M-Chinese* uses datasets from a variety of tasks for open-source UBERT.
- PM6: *IDEA-CCNL/Erlangshen-UniMC-RoBERTa-110M-Chinese* uses 13 supervised datasets to train the model. (Yang *et al.*, 2022)

4.2 Experimental Settings

The proposed CDVAI dataset includes the annotation of irony context to allow the model to understand contextual emotional changes during the training process. The paper uses a variety of modified pre-trained BERT models as the experimental encoder. The parameters are shown in Table 3. Each pre-trained model uses the same parameters, except the learning rate. Since context contains less information than sentences, a smaller learning rate should be tried. The context span prediction model in the second part of the experiment were tried smaller learning rate due to the difficulty to learn the span of context in the sentence.

Table 3. Parameter settings of BERT models.

Parameter	Value
Optimizer	Adam
Learning rate - sentence-level	4e-4, 4e-5, 4e-6
Learning rate - context-level	4e-5, 4e-6, 4e-7
Learning rate – span prediction	43e-6, 45e-6
Number of epochs	50

4.3 First Part of Experimental Results

The prediction performance of dimensional VAI score on sentence-level is shown in Table 4. First, the performance of valence prediction is quite good. All MAEs are about 0.4, no matter what approach in this paper. However, the paper can still discover that M1 got the greatest performance, which indicates more complex hidden features don't get better result in valence. The reason is detecting the score of valences is relatively easy in the task, so more complex hidden features cause worse results. The performance of arousal prediction is a bit worse than valence, which indicates arousal is relatively difficult to learn. All MAEs are about 0.6, however M1 does not have the greatest approach on all models. M2 and M3 make the performance progress at PM3. PM4, MP5 and PM6 improve performance while using M2 or M3. Finally, the performance of irony prediction is a bit better than arousal. All MAEs are about 0.5 to 0.6, which means our annotated method to judge irony is effective. M2 and M3 are more helpful to improve the performance of PM2, PM4, PM5 and PM6, which indicate these models can deal with complex hidden features better. Overall, the result of sentence-level VAI is quite well, but M2 and M3 doesn't show significantly helpful while predict VAI scores.

Table 4. Prediction performance of dimensional VAI score on sentence-level.

Model	Valence			Arousal			Irony		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
PM1	0.346	0.421	0.390	0.521	0.649	0.639	0.521	0.577	0.603
PM2	0.380	0.421	0.390	0.619	0.649	0.639	0.601	0.577	0.603
PM3	0.371	0.410	0.371	0.643	0.603	0.596	0.538	0.570	0.566
PM4	0.380	0.412	0.390	0.619	0.614	0.639	0.601	0.572	0.603
PM5	0.376	0.381	0.420	0.615	0.616	0.610	0.575	0.591	0.559
PM6	0.380	0.412	0.390	0.619	0.614	0.639	0.601	0.577	0.603

The prediction performance of dimensional VA score on context-level is shown in Table 5. The context-level valence also performs quite well. Overall, MAE is around 0.4. However, M2 and M3 improve the performance significantly. Among them, M3 provides an even better effect. This shows our approaches are more effective on context-level. The reason may be the complex relation of sentence and context, which shows that the true sentiment pattern of ironic sentences requires a judgment of the context first then combined with the whole sentence to understand. This result can also be seen in arousal. However, M2 showed more effective help in predicting arousal scores. The inference is the information of context itself is more important than the whole sentence while predicting arousal scores, and this effect significantly.

Table 5. Prediction performance of dimensional VA score on context-level.

Model	**Valence**			**Arousal**		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**
PM1	0.431	0.408	0.389	0.819	0.649	0.787
PM2	0.413	0.408	0.389	0.796	0.649	0.787
PM3	0.431	0.468	0.427	0.798	0.603	0.829
PM4	0.413	0.415	0.389	0.796	0.614	0.787
PM5	0.426	0.463	0.428	0.815	0.616	0.834
PM6	0.413	0.415	0.389	0.796	0.614	0.787

Analysis of the above shows that M2 and M3 can improve the performance on context-level significantly. However, they don't seem quite helpful on sentence-level. In summary, depending on the choice of pre-trained model, context information can improve performance while predicting VAI score in sentence. Results on context-level shows that understanding the true sentiment pattern of ironic sentences requires to combine sentence and context information.

4.4 Second Part of Experimental Results

Due to the lack of context annotation in previous study. The paper proposes a model to predict the irony context span automatically. The paper proposes to fine-tuning PM1 to PM6 to compare prediction performances. But the performances of the model are hard to accept. So, the paper adds a new pre-trained model to solve this problem, which is PM7: *IDEA-CCNL/Erlangshen-DeBERTa-v2-97M-Chinese* (He *et al.*, 2020) to improve the model. The results show in Table 6.

Table 6. Prediction performance of context span predict in ironic sentences.

Model	**Indicators**		
	Precision	**Recall**	**F1**
PM1	0.373	0.302	0.333
PM2	0.349	0.274	0.307
PM3	0.329	0.218	0.262
PM4	0.373	0.309	0.331
PM5	0.436	0.352	0.390
PM6	0.373	0.298	0.338
PM7	0.438	0.377	0.405

The paper uses PM7 to predict context span, then replace origin context span with the predicted context span and execute the same process of above experiment to evaluate the availability. The span predict model results on sentence-level are shown in Table 7. Compared with the first part of the experiment, the MAE of valence on sentence-level on M2 is making progress. The reason may be that although the predicted context spans are not correct, they contain more emotion information words accidently. The MAE of arousal on sentence-level becomes larger on M2, however the MAE reduces on M3. The reason may be the noise of the context improves the performance. The same situation occurs with irony. This discovery is quite surprising that the information of whole context may not be the important one to improve the prediction performance but the critical part or words in the context.

Table 7. Prediction performance of dimensional VAI score on sentence-level in part 2 experiment.

Model	Valence		Arousal		Irony	
	M2	**M3**	**M2**	**M3**	**M2**	**M3**
PM1	0.337	0.403	0.631	0.585	0.575	0.638
PM2	0.337	0.403	0.631	0.585	0.575	0.638
PM3	0.384	0.376	0.618	0.648	0.606	0.613
PM4	0.392	0.403	0.677	0.585	0.594	0.638
PM5	0.383	0.364	0.607	0.609	0.561	0.574
PM6	0.392	0.403	0.677	0.585	0.594	0.638

Finally, the span predict model results on context-level are shown in Table 8. Since the context span of the model predictions cannot be fully correct. Therefore, the main purpose of this part of the experiment is to examine the effect of VA score prediction with a biased context span. Compared with the first part of the experiment, the MAE of valence on context-level decreases a little. The MAE of arousal decreases quite a lot. This proves that the correction of context span matters.

Table 8. Prediction performance of dimensional VAI score on context-level in second part of experiment.

Model	Valence		Arousal	
	M2	**M3**	**M2**	**M3**
PM1	0.419	0.447	0.819	0.824
PM2	0.419	0.447	0.819	0.824
PM3	0.471	0.436	0.867	0.810
PM4	0.435	0.447	0.844	0.824
PM5	0.442	0.451	0.801	0.844
PM6	0.435	0.447	0.844	0.824

4.5 Error Analysis

Based on the performance of the model, the PM3 model has well performance in experiments. The paper presents an incorrect prediction case, as follows:

Sentence: "*很好...連喇叭都壞了 X-(" ("Very good.... even the speakers are broken X-(")*

Context: "*連喇叭都壞了" ("even the speakers are broken*")

Judgement: The prediction results are shown in Table 9. The reason why the model judges the valence to be 1.71 on sentence-level, may be that it judges "連", "壞了" ("*even, broken*") as negative words. However, the post only indicated that the speakers are broken, which is usually not perceived as highly negative. The lack of common sense may have led to the failure to detect its valence correctly. In terms of irony, the prediction score is relatively large. It is speculated that because the judgment of valence is relatively negative and the term "很好" ("*very good*") is positive, there is a large emotional gap. The model therefore yields a higher irony score. However, the sentence has no other span that emphasizes irony, so the annotated score is lower.

Table 9. Prediction results of the example

	Sentence-level			**Context-level**	
	V	**A**	**I**	**V**	**A**
Annotated	2	3	1	2	3
Predicted	1.71	3.45	1.94	1.63	1.97

5. Conclusion

This paper established the CDVAI dataset which extended from NTU irony corpus. The CDVAI dataset contains multi-dimensional sentiment annotation and irony context sentiment annotation, which is helpful for developing Chinese irony detection methods that allow the model to learn sentimental patterns in ironic sentence and context. The experimental results showed that the annotation of CDVAI dataset provides a learning direction for the BERT based model to understand the irony structure and sentiment contrast between sentence-level and context-level. Using M3 can improve performance significantly. The paper has summarized our experiment results below. First, M2 and M3 don't show significantly helpful while predicting VAI scores. However, in the second part of the experiment that the information of the whole context may not be important to improve the prediction performance but the critical part or words in the context. Second, M2 and M3 show significant improvement in predicting score of context-level, which proves the sentiment pattern of ironic context needs to combine sentence information. Finally, the sentiment in ironic contexts is harder to learn for the model, which needs correct spans of context to improve the performance.

The weakness of the CDVAI dataset is that the corpus is relatively small and excludes the whole ironic grammatical structure. Nevertheless, the paper is suitable to use as guide data to obtain more samples or as a template for annotation guidelines. Furthermore, the proposed CDVAI dataset could be combined with other ironic corpora to extend the training sample size. Furthermore, the model can be improved in the future.

Acknowledgments

This research was partially supported by the Ministry of Science and Technology (MOST), Taiwan (Grant numbers: MOST 110-2221-E-031-004, and MOST 111-2221-E-031-004MY3).

References

Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, *28*(2), 55-63.

Colston, H. L. (2019). Irony as indirectness cross-linguistically: On the scope of generic mechanisms. In *Indirect Reports and Pragmatics in the World Languages* (pp. 109-131). Springer.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Version 2.

Dimovska, J., Angelovska, M., Gjorgjevikj, D., & Madjarov, G. (2018). Sarcasm and irony detection in English tweets. In *Proceedings of the International Conference on Telecommunications*, 120-131. https://doi.org/I:10.1007/978-3-030-00825-3_11

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 470-478.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Jia, X., Deng, Z., Min, F., & Liu, D. (2019). Three-way decisions based feature fusion for Chinese irony detection. *International Journal of Approximate Reasoning*, 113, 324-335. https://doi.org/10.1016/j.ijar.2019.07.010

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2018). Automatic sarcasm detection: A survey. *ACM Computing Surveys*, *50*(5), 1-22. https://doi.org/10.1145/3124420

Kumar, H., & Harish, B. (2019). Automatic irony detection using feature fusion and ensemble classifier. International Journal of Interactive Multimedia & Artificial Intelligence, 5(7). 70-79. https://doi.org/10.9781/ijimai.2019.07.002

Lestari, W. (2019). Irony analysis of Memes on Instagram social media. *PIONEER: Journal of Language and Literature*, *10*(2), 114-123. https://doi.org/10.36841/pioneer.v10i2.192

Li, A.-R., & Huang, C.-R. (2020). A method of modern Chinese irony detection. *From Minimal Contrast to Meaning Construct* (pp. 273-288). Springer.

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., & Li, M. (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE transactions on affective computing*, *12*(4), 900-912. https://doi.org/10.1109/TAFFC.2019.2903056

Lu, Q., Zhu, Z., Xu, F., Zhang, D., Wu, W., & Guo, Q. (2020). Bi-GRU sentiment classification for Chinese based on grammar rules and BERT. International Journal of Computational Intelligence Systems, 13(1), 538-548. https://doi.org/10.2991/ijcis.d.200423.001

Naseem, U., Razzak, I., Eklund, P., & Musial, K. (2020). Towards improved deep contextual embedding for the identification of irony and sarcasm. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, 1-7. https://doi.org/10.1109/IJCNN48605.2020.9207237

Pfeifer, V. A., & Lai, V. T. (2021). The comprehension of irony in high and low emotional contexts. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *75*(2), 120-125. https://doi.org/10.1037/cep0000250

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 9-15. https://doi.org/10.18653/v1/W16-0404

Rangwani, H., Kulshreshtha, D., & Singh, A. K. (2018). Nlprl-iitbhu at semeval-2018 task 3: Combining linguistic features and emoji pre-trained cnn for irony detection in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 638-642. https://doi.org/10.18653/v1/S18-1104

Tang, Y.-j., & Chen, H.-H. (2014). Chinese irony corpus construction and ironic structure analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1269-1278.

Turc, I., Chang, M. W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962.

Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons. In *Proceedings of 19th European Conference on Artificial Intelligence*, 765-770. https://doi.org/10.3233/978-1-60750-606-5-765

Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R.,... & Zhang, J. (2022). Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. arXiv preprint arXiv:2209.02970

Xiang, R., Gao, X., Long, Y., Li, A., Chersoni, E., Lu, Q., & Huang, C.-R. (2020). Ciron: a new benchmark dataset for Chinese irony detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5714-5720.

Xie, H., Lin, W., Lin, S., Wang, J., & Yu, L.-C. (2021). A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*, *579*, 832-844. https://doi.org/10.1016/j.ins.2021.08.052

Yang, P., Wang, J., Gan, R., Zhu, X., Zhang, L., Wu, Z., Gao, X., Zhang, J., & Sakai, T. (2022). Zero-shot learners for natural language understanding via a unified multiple choice perspective. arXiv preprint arXiv:2210.08590.

Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 89-138

結合語音辨認及合成模組之台語語音轉換系統

Taiwanese Voice Conversion based on Cascade ASR and TTS Framework

許文漢 [*]、廖元甫[#]、王文俊 [+]、潘振銘[+]

Wen-Han Hsu, Yuan-Fu Liao, Wern-Jun Wang, and Chen-Ming Pan

摘要

台語已被聯合國列為瀕危消失語言，急需傳承。因此，本論文研究如何做出一個可以用任何人的聲音，合成出任何台語語句的台語語音合成系統。為達到此目的，我們首先(1)建置一 Taiwanese Across Taiwan (TAT) 大規模台文語音語料庫，其共有 204 位語者，約 140 小時的語料，其中有兩男兩女，每人約 10 小時的台語語音合成專用語料。然後(2)基於 Tacotron2 之語音合成架構，並加上前端中文字轉台羅拼音模組與後端 WaveGlow 即時語音生成器，建立中文文字轉台語語音合成系統。最後(3)基於串接台語語音辨認與語音合成架構，建置一台語語音轉換系統，並完成同語言：台語對台語語音轉換；以及跨語言：華語對台語語音轉換，兩種台語語音轉換功能。為評估此台語語音轉換系統的成效，我們透過網路公開招募到 29 位實驗者，進行同語言及跨語言轉換台語語音兩項評分任務，並分別進行針對「自然度」與「相似度」的 MOS 分數之主觀評測。實驗結果顯示，在同語言部分，若使用目標語者 10 分鐘，3 分鐘與 30 秒語料進行測試，自然度平均 MOS 分數依序為 3.45 分，3.02 分與 2.23 分，相似度平均 MOS 分數依序為 3.38 分，2.99 分與 2.10 分；而在跨語言部分，若使用目標語者 6 分鐘與 3 分鐘語料進行測試，自然度平均 MOS 分數依序為

[*] 國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
E-mail: jeff3136169@gmail.com

[*] 國立陽明交通大學智能系統系
Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University
E-mail: yfliao@nycu.edu.tw

[+] 中華電信研究院
Chunghwa Telecom Laboratories
E-mail: {wernjun, chenming }@cht.com.tw

2.90 分與 2.70 分，相似度平均 MOS 分數依序為 2.84 分與 2.54 分。由實驗結果，可以顯示我們確實初步達成一個可以用任何人的聲音，合成出任何台語語句的台語語音合成系統。

Abstract

Taiwanese has been listed as an endangered language by the United Nations and is urgent for passing on. Therefore, this study wants to find out how to make a Taiwanese speech synthesis system that can synthesize any Taiwanese sentences via anyone's voice. To achieve this goal, we first (1) built a large-scale Taiwanese Across Taiwan (TAT) corpus, with in total of 204 speakers and about 140 hours of speech. Among those speakers, two men and women, each one has especially about 10 hours of speech recorded for the purpose of speech synthesis, then (2) establish a Chinese Text-to-Taiwanese speech synthesis system based on the Tacotron2 speech synthesis architecture, plus with a frontend sequence-to-sequence-based Chinese characters to Taiwan Minnanyu Luomazi Pinyin (shortened as Tâi-lô) machine translation module and the backend WaveGlow real-time speech generator, and finally, (3) constructed a Taiwanese voice conversion system based on the concatenated speech recognition and speech synthesis framework where two voice conversion functions had been implemented including (1) same-language: Taiwanese to Taiwanese voice conversion, and (2) multi-language: Chinese to Taiwanese voice conversion. In order to evaluate the Taiwanese voice conversion system, we publically recruited 29 subjects from the Internet to conduct two kinds of scoring task: same-language and cross-language voice conversion and carried out the subjective "naturalness" and "similarity" mean opinion score (MOS) evaluations respectively. The test result shows that in the Intra-lingual session, the average naturalness MOS is 3.45, 3.02 and 2.23 points, and average similarity MOS score's 3.38, 2.99 and 2.10 points while using 10 minutes, 3 minutes, and 30 seconds target speech, respectively; in cross-lingual part, the average naturalness MOS score is 2.90 and 2.70 points; average similarity MOS score is 2.84 and 2.54 points while using 6 minutes and 3 minutes target speech, respectively. From those results, it shows that our proposed system indeed could synthesize any Taiwanese sentences via anyone's voice.

關鍵詞：台文語音語料庫、台語語音合成、台語語音轉換

Keywords: Taiwanese Across Taiwan, Taiwanese Speech Synthesis, Taiwanese Voice Conversion

1. 緒論 (Introduction)

1.1 動機與目的 (Motivation)

台灣的台語人口逐漸式微，普遍有越年輕越不使用台語的趨勢，最多情況就是面對長輩時，不懂得如何用台語溝通。因此本論文針對現況，希望能建置（1）中文文字轉台語語音合成系統，與（2）具語音轉換功能的多語者台語語音合成系統，讓使用者在面對講不出台語的窘境時，能知道要如何用台語表達，並能聽到以自己講話的聲音合成的台語語音。

阻礙達成此目標的主要問題是缺乏台語語音語料，因此，我們首先針對此台語語料基礎建設之需求，建置台文語音語料庫（Taiwanese Across Taiwan，TAT）作為研發台語語音合成系統之基礎，其包括（1）TAT-Vol1~2，其為橫跨台灣錄製的 100 小時/200 人之台語語音辨認（automatic speech recognition, ASR）與多語者語音合成（multi-speaker text-to-speech, TTS）用語料庫，與（2）TAT-TTS-M1~2 與 TAT-TTS-F1~2，此為依據台語強勢（高雄）腔與次強勢（台北）腔，各錄製一男一女，每人十小時之台語語音合成用語料庫。此外，為使台語語音合成的自然度能逼近真人，本論文也針對 TAT-TTS 語料庫進行人工台語變調與韻律標註，包括校正語料中每個字的音調，與加上語音韻律階層邊界標記，以改善合成語音的韻律流暢度。我們即利用此 TAT 語料庫，訓練前述之台語語音合成與語音轉換系統。

此外，我們考慮在中文文字轉台語語音合成系統系統方面，需要能即時將中文文字轉譯成台語語音，因此除採用高品質（state-of-the-art）的 end-to-end（E2E）Tacotron 2 語音合成主架構，並再加上基於 convolution neural network（CNN）之 sequence-to-sequence 中文文字轉台語台羅拼音機器翻譯前級，與可即時合成語音的 WaveGlow 語音合成後級。

而為盡量能多樣化合成台語語音的（人物）音色，在語音轉換架構方面，我們改用基於串接語音辨認（automatic speech recognition, ASR）與語音合成（text-to-speech, TTS）模組之 cascaded ASR+TTS 架構。其中 ASR 與 TTS 模組，都將使用基於 end-to-end（E2E）架構的 Transformer 類神經網路。

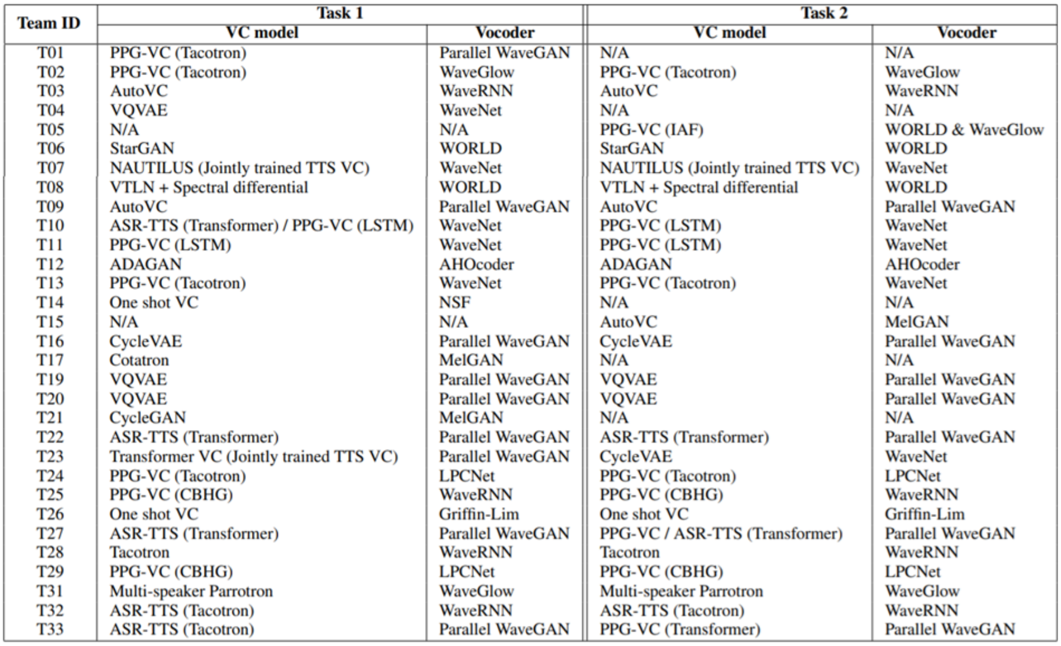
最後，我們使用 mean opinion score（MOS）主觀評分方式，測試台語語音合成與語音轉換系統的『自然度』，與比較語音轉換系統的合成音檔與目標語者音色的「相似度」。此外，如何在有限的目標語者訓練語料限制情況下，盡量維持住語音轉換後音檔的自然度和相似度，也是語音轉換系統評估成效的重點。

1.2 背景 (Background)

目前主流的語音合成系統，幾乎都是基於類神經網路技術，比如 Google 提出的 Tacotron2+WaveNet Vocoder 架構(Shen *et al.*, 2018)。Tacotron2 可直接以類神經網路，進行文脈訊息處理，建立一「文字」轉「Mel-Spectrogram」的 end-to-end 架構。WaveNet Vocoder 接著將「Mel-Spectrogram」轉成「Speech Waveform」。此架構出現以後，語音

合成的音質就幾乎接近人聲。但此處的 WaveNet Vocoder，是一個以 sample 為單位做計算的序列式遞迴網路架構，sample 需要一個接著一個照前後順序產生。除計算量相當大外，也不易平行化，導致語音生成速度非常慢。而 NVIDIA 提出的 WaveGlow (Prenger *et al*., 2018)則正好可以解決此問題，WaveGlow 是一個基於流的生成模型，其利用批次取樣與分佈轉換處理，避開遞迴網路架構計算量大，且不易平行化的問題，合成所需時間比大幅減少，若是合成約 10 秒以下的語音，大幅減少的合成時間已經幾乎接近體感的即時合成，且其公開的平均意見得分（MOS）測試也表明，WaveGlow 的音質也不遜於 WaveNet。

而在語音轉換方面，國際研究社群在 2016，2018 和 2020 年，分別舉辦多次的 voice conversion challenge（VCC）競賽。2020 年的比賽(Zhao *et al*., 2020)分成了 Task1：同語言的半平行語音轉換（Intra-lingual semi-parallel VC)，和 Task2：跨語言語音轉換（Cross-lingual VC）兩種任務。在 VCC 2020 中，各參賽隊伍在語音特徵轉換和後端聲碼器的選擇如圖 1 所示。由 Task1 的語音轉換音檔自然度及相似度比賽結果（如圖 2 所示），顯示在同語言的情況下，得益於 ASR 和 TTS 技術近幾年的快速發展，一些基於聲學後驗圖譜（phonetic posteriorgram，PPG），或是 ASR 加 TTS 架構的語音轉換技術，在語音轉換音檔的自然度與相似度上，被普遍認為優於以往的自動編碼器（Auto-Encoder）或是生成對抗網路（GAN）等系統。而由 Task2 比賽結果（如圖 3 所示）可以看出來，PPG 和 cascaded ASR+TTS 的方法，都很適合用在跨語言語音轉換情況。



Team ID	Task 1		Task 2	
	VC model	Vocoder	VC model	Vocoder
T01	PPG-VC (Tacotron)	Parallel WaveGAN	N/A	N/A
T02	PPG-VC (Tacotron)	WaveGlow	PPG-VC (Tacotron)	WaveGlow
T03	AutoVC	WaveRNN	AutoVC	WaveRNN
T04	VQVAE	WaveNet	N/A	N/A
T05	N/A	N/A	PPG-VC (IAF)	WORLD & WaveGlow
T06	StarGAN	WORLD	StarGAN	WORLD
T07	NAUTILUS (Jointly trained TTS VC)	WaveNet	NAUTILUS (Jointly trained TTS VC)	WaveNet
T08	VTLN + Spectral differential	WORLD	VTLN + Spectral differential	WORLD
T09	AutoVC	Parallel WaveGAN	AutoVC	Parallel WaveGAN
T10	ASR-TTS (Transformer) / PPG-VC (LSTM)	WaveNet	PPG-VC (LSTM)	WaveNet
T11	PPG-VC (LSTM)	WaveNet	PPG-VC (LSTM)	WaveNet
T12	ADAGAN	AHOcoder	ADAGAN	AHOcoder
T13	PPG-VC (Tacotron)	WaveNet	PPG-VC (Tacotron)	WaveNet
T14	One shot VC	NSF	N/A	N/A
T15	N/A	N/A	AutoVC	MelGAN
T16	CycleVAE	Parallel WaveGAN	CycleVAE	Parallel WaveGAN
T17	Cotatron	MelGAN	N/A	N/A
T19	VQVAE	Parallel WaveGAN	VQVAE	Parallel WaveGAN
T20	VQVAE	Parallel WaveGAN	VQVAE	Parallel WaveGAN
T21	CycleGAN	MelGAN	N/A	N/A
T22	ASR-TTS (Transformer)	Parallel WaveGAN	ASR-TTS (Transformer)	Parallel WaveGAN
T23	Transformer VC (Jointly trained TTS VC)	Parallel WaveGAN	CycleVAE	WaveNet
T24	PPG-VC (Tacotron)	LPCNet	PPG-VC (Tacotron)	LPCNet
T25	PPG-VC (CBHG)	WaveRNN	PPG-VC (CBHG)	WaveRNN
T26	One shot VC	Griffin-Lim	One shot VC	Griffin-Lim
T27	ASR-TTS (Transformer)	Parallel WaveGAN	PPG-VC / ASR-TTS (Transformer)	Parallel WaveGAN
T28	Tacotron	WaveRNN	Tacotron	WaveRNN
T29	PPG-VC (CBHG)	LPCNet	PPG-VC (CBHG)	LPCNet
T31	Multi-speaker Parrotron	WaveGlow	Multi-speaker Parrotron	WaveGlow
T32	ASR-TTS (Tacotron)	WaveRNN	ASR-TTS (Tacotron)	WaveRNN
T33	ASR-TTS (Tacotron)	Parallel WaveGAN	PPG-VC (Transformer)	Parallel WaveGAN

圖 1. VCC 2020 參賽者使用模型架構詳細資訊(Zhao et al., 2020)
[Figure 1. Summary of adopted approaches in Voice Conversion Challenge (VCC) 2020]

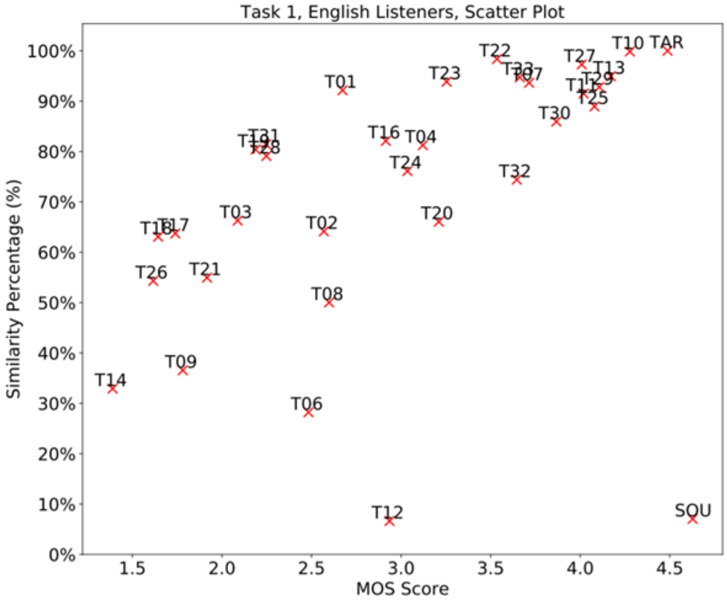


圖 2. VCC 2020 Task1 比賽結果(Zhao et al., 2020)
[Figure 2. Benchmark Results on Task1 of Voice Conversion Challenge (VCC) 2020]

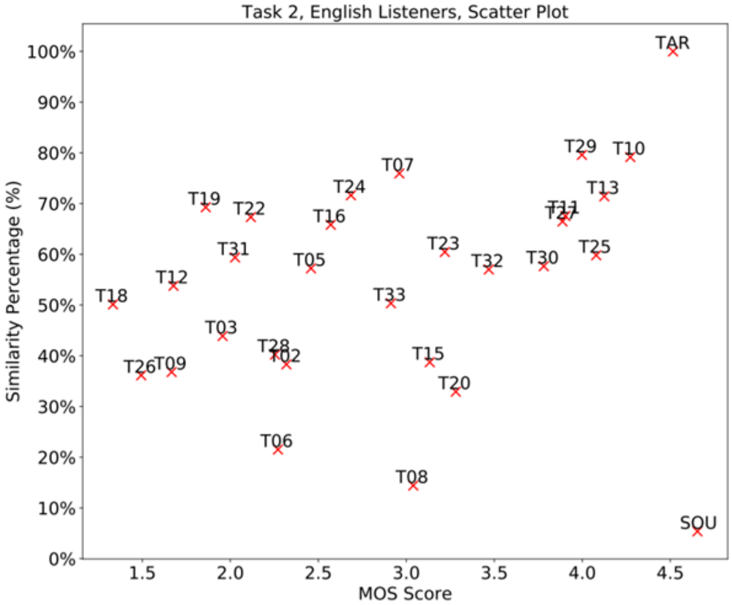


圖 3. VCC 2020 Task2 比賽結果(Zhao et al., 2020)
[Figure 3. Benchmark Results on Task2 of Voice Conversion Challenge (VCC) 2020]

此外，從整體比賽結果來看，同語言任務的自然度與相似度分數，都已經可以達到相當不錯的分數。而跨語言語音轉換，因難度較高，在自然度和相似度都還有進步的空間，不同語言說話的方式還是一定程度上影響了語音轉換的成效。

1.3 研究方法 (Approaches)

目前還較少有人嘗試台語的語音合成與語音轉換技術。以下說明我們建置（1）中文文字轉台語語音合成系統，與（2）具語音轉換功能的多語者台語語音合成系統的主要目標與研究方法。

1.3.1 單人台語語音合成系統 (Single-Speaker Chinese Text-to-Taiwanese Speech Synthesis)

此系統使用 sequence-to-sequence-based 中文文字轉台語台羅拼音機器翻譯前級，串接 Tractron 語音合成主架構，與 WaveGlow 語音合成後級，以實現高品質且即時之台語語音合成系統。其系統流程如圖 4 所示：使用者輸入中文文字後，先透過機器翻譯成相對應的台語語句的台羅拼音文本，再使用 Tacotron 2 將台羅拼音文本轉換成台語合成語音的頻譜，最後用 WaveGlow 將台語語音的頻譜解碼成實際的台語語音。此系統最後並被做成線上展示網頁(http://tts001.iptcloud.net:8801)，公開供大眾測試。

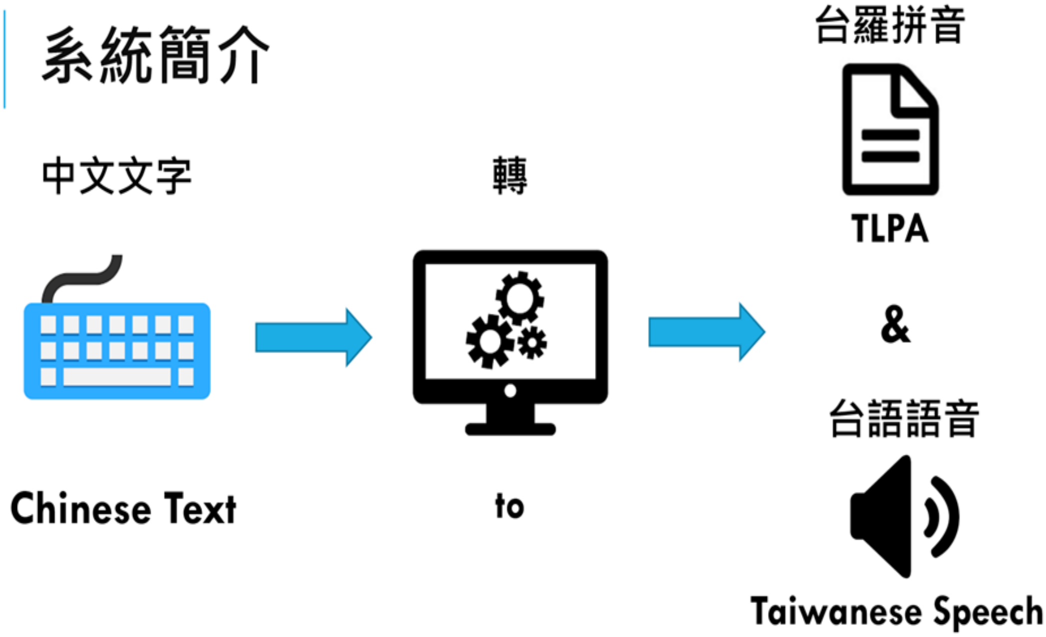


圖 4. 單人台語語音合成系統流程簡介
[Figure 4. Block-diagram of the Single-Speaker Taiwanese Text-to-Speech System]

1.3.2 結合語音辨認及合成模組之台語語音轉換系統 (Cascade ASR and TTS-based Taiwanese Voice Conversion System)

此部分採用 Cascade ASR+TTS 架構(Huang *et al*., 2020)進行台語語音轉換系統的初步建置，其架構如圖 5 所示。其包含三個模型，分別是負責台語語音辨認的語者獨立 Transformer-based ASR model，負責擷取目標語者語音特徵的 X-Vector，以及負責依據目標語者 X-Vector，進行台語語音合成的 Multi-speaker Transformer-TTS model。

其在訓練階段，利用少量目標語者的聲音，求取其 X-Vector，並微調（fine-tuning）事先預備好的多語者 TTS 模型，產生目標語者的 TTS 模型。因此，在測試階段，就能利用 ASR 前級，把來源語者的音檔辨識出文字，並用微調好的 TTS 合成出具目標語者語音特徵的語音檔。

此外，我們也針對跨語言語音轉換任務，將原先 VCC 2020 Task2 中的華語轉英語的專案(Kamo, 2021)，替換成華語轉台語的語音轉換系統。

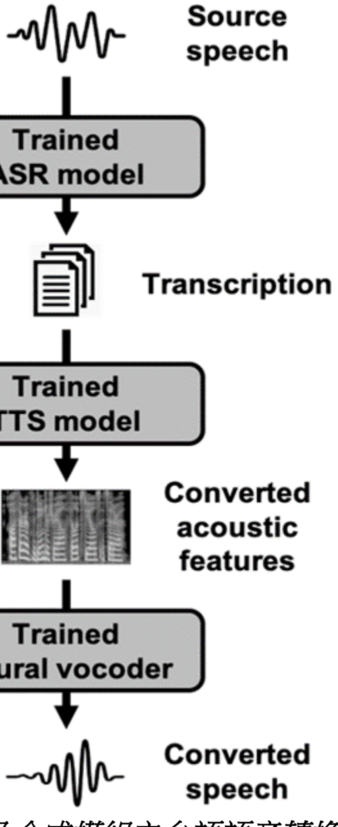


圖 5. 結合語音辨認及合成模組之台語語音轉換系統流程簡介 (Huang et al., 2020)

[Figure 5. Flowchart of the Cascade ASR and TTS-based Taiwanese Voice Conversion System]

2. 相關研究 (Related Works)

2.1 語音合成 (Text-to-Speech)

語音合成，是一種將文本轉換成語音（Text-To-Speech，TTS）的技術，從最早期的音檔串接合成，發展到使用隱藏式馬可夫模型，一直到現今的類神經網路，電腦合成語音的聲音已經到了幾乎跟真人相似的程度。

2.1.1 Tacotron2

Tacotron2 的架構圖(Shen *et al.*, 2018) 如圖 6 所示，其使用 encoder-decoder + Location Sensitive Attention 的架構。將文本資訊輸入 encoder 後，encoder 類神經網路進行文本分析，萃取出輸入文句的文脈特徵參數，decoder 接著依據文脈特徵參數，以 attention 權重機制，對齊（alignment）輸入文字與產出的合成語音的梅爾頻譜圖，最後再由 WaveNet Vocoder 將梅爾頻譜進行解碼，合成高品質的語音。

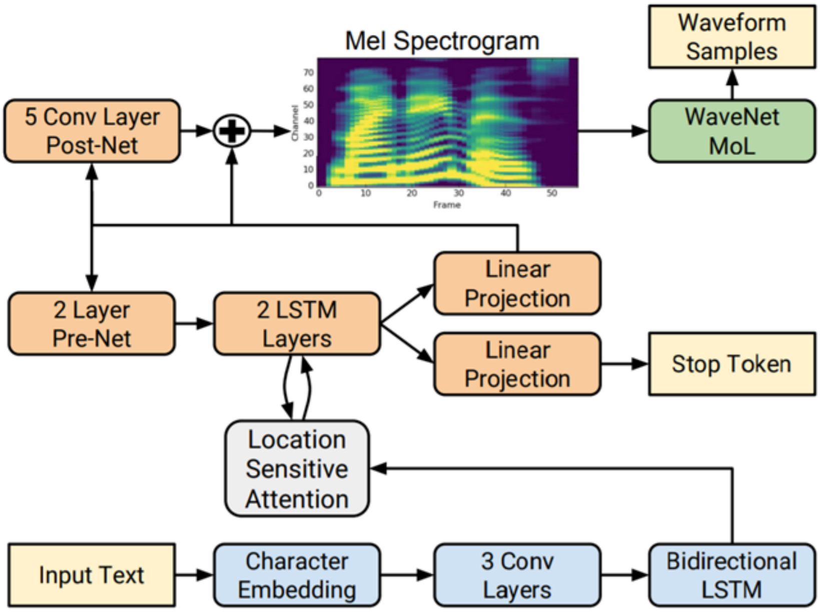


圖 6. Tacotron2 架構流程圖(Shen et al., 2018)
[Figure 6. Architecture of the Tacotron2 Speech Synthesis System]

與同是 Google 提出的對前一代的 Tacotron (Wang *et al.*, 2017)相比，其主要改良是取消 CBHG，改用普通的 LSTM 和 Convolution layer，接著每一步 decoder 改成只生成一個 frame，並在後面增加了一個 5 層的後置 CNN 網路，來使產出的梅爾頻譜更正確，最後重點是將 Tacotron 後端的 Griffin-Lim，改為能合成出更像真人講話聲音的 WaveNet Vocoder。

2.1.2 WaveGlow

上述提到的 Tacotron2 已經可以做到合成出接近人聲的音檔，但後端的 WaveNet Vocoder 卻有著語音生成速度緩慢的缺點。而 NVIDIA 提出的 WaveGlow 架構(Prenger *et al*., 2018)即是此問題的解決方法，WaveGlow 是一種 flow-based generative networks，其架構圖如圖 7 所示，主要是透過一可以完美執行逆轉換的音檔波形正規化轉換類神經網路架構，在給定輸入音檔的相對應頻譜的條件下，將所有可能的語音音檔波形折疊投影到一高斯分布 z 空間。因此，在合成時只需在 z 空間中作取樣，再依據給定的合成音檔的對應頻譜條件，即可以類神經網路執行逆轉換，將取樣出的 z 向量，反折疊回實際的音檔波形。此外，因為這些逆轉換的類神經網路運算過程，都可以進行批次平行化處理，所以最終可以即時合成高品質語音。

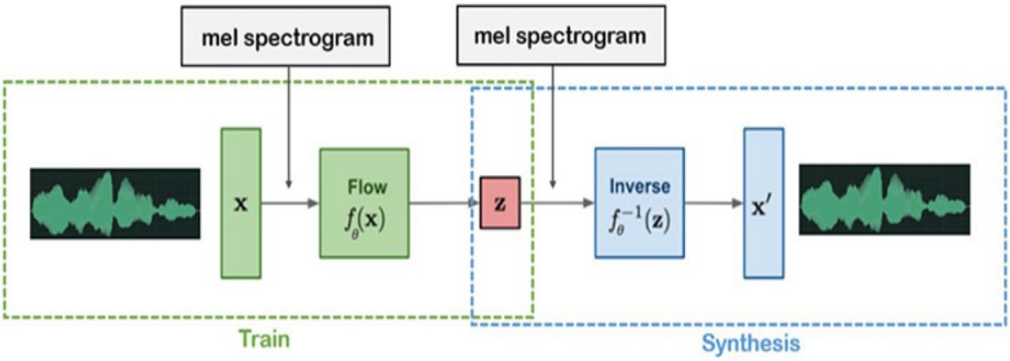


圖 7. WaveGlow 訓練及合成過程(Prenger et al., 2018)
[Figure 7. Training and Synthesis processing of the WaveGlow Approach]

其中，可完美執行逆轉換的類神經網路的實際架構圖如圖 8 所示。訓練時依據基於輸入與輸出音檔相似度的 cost function 導引，依據給定的合成音檔頻譜條件，多次利用可逆卷積（invertible convolution）與耦合層（coupling layer）網路，進行函式轉換，逐步學習如何將真實語音波形訊號 x，投射到一具高斯分佈之隱藏變數 z 的向量空間。並在訓練時限制 mapping 函式為可逆函式。如此，WaveGlow 在生成語音波型時，即可在隱藏變數 z 空間進行取樣，再依據給定的合成音檔頻譜條件，經多次逆函式轉換，逐步將 z 向量，反解成真實語音波形訊號 x。

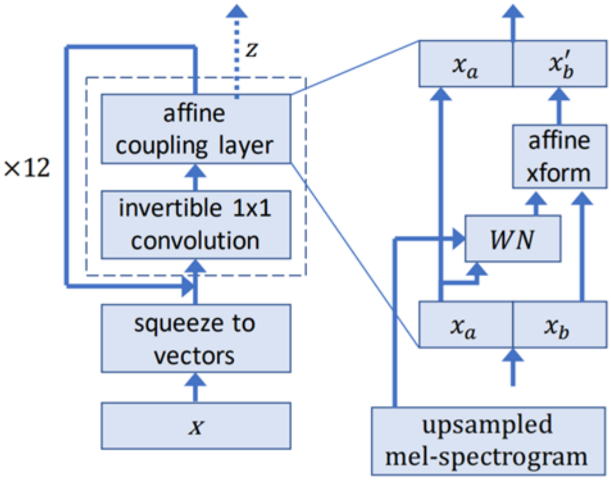


圖 8. WaveGlow 網路結構圖(Prenger et al., 2018)
[Figure 8. The architecture of the WaveGlow System]

2.2 語音轉換 (Voice Conversion)

較早的語音轉換方法，在有平行錄音語料的情形下，有統計式及樣本式兩種方法，例如高斯混合模型（Gaussian mixture model, GMM）與基於局部線性嵌入(locally linear embedding, LLE)的語音轉換方法等。其中，統計式語音轉換方法的主要缺點，為轉換後所得到的語音頻譜有過度平滑的現象，因而降低了語音品質與語者相似度。而樣本式語音轉換方法的優點，則是不需要模型訓練過程，但是樣本數量的多寡會影響轉換後語音的音質，而且轉換過程的運算量會隨著樣本數量增加而提高。

隨著類神經網路的高度發展，基於神經網路的語音轉換技術也漸漸成為主流。像是變分式自動編碼器（variational autoencoder, VAE）架構(Hsu *et al*., 2016)，可以用來串在一通用語音合成系統的後級，用以在只有少量語者語音資料的情形下，轉換通用語音合成系統的輸出音色，如圖 9 所示。

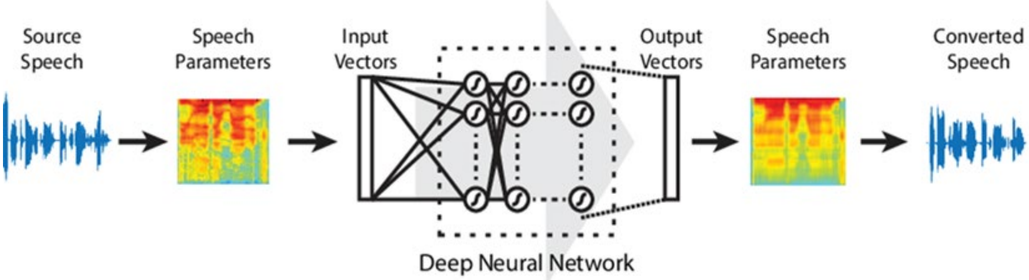


圖 9. 基於變分式自動編碼器（variational autoencoder, VAE）之語音轉換方法 (Hsu et al., 2016)
[Figure 9. Variational autoencoder-based voice conversion]

此外，亦有利用語音辨認器作為輔助的方法，例如，以語音辨認器，計算語音中的聲學後驗圖譜（phonetic posteriorgram，PPG）(Sun *et al*., 2016)，並依此圖譜當作語音中說話內容的資訊，用以輔助語音轉換，如圖 10 所示。此做法可以容忍因前級語音辨認器的錯誤辨認輸出，導致後級語音轉換合成的錯誤。

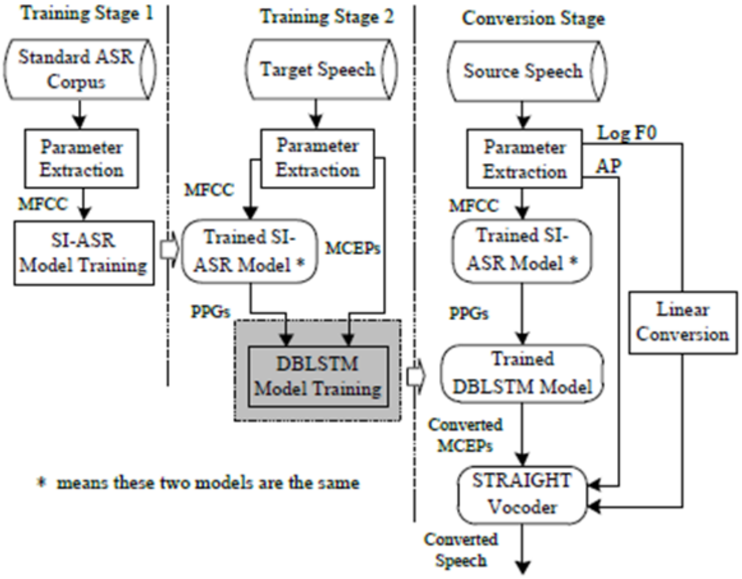


圖 10. 基於聲學後驗圖譜（phonetic posteriorgram，PPG）之語音轉換架構(Sun et al., 2016)
[Figure 10. Phonetic Posteriorgram-based voice conversion]

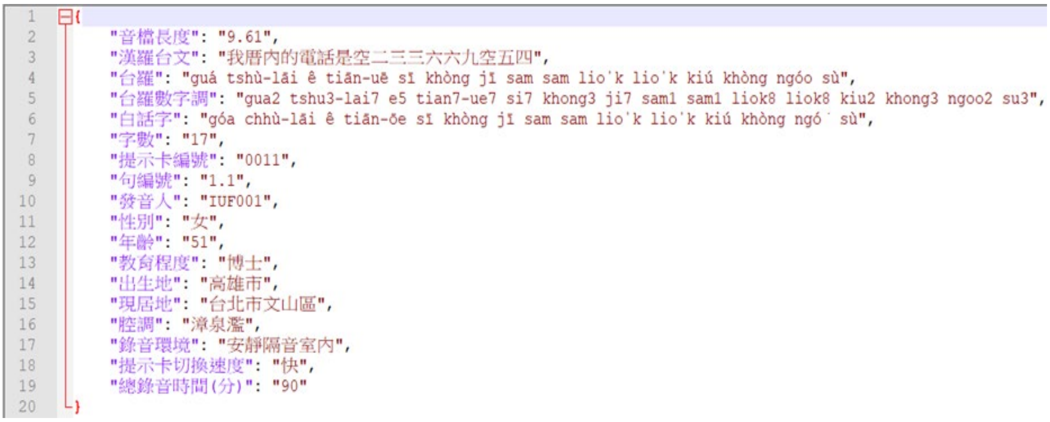
3. Taiwanese Across Taiwan語料庫建置 (Taiwanese Across Taiwan Corpus)

為了處理在此論文中，所述之台語語音合成與語音轉換系統開發，所需的大量台語語料問題，我們首先規劃並建置一大規模台文語音語料庫（Taiwanese across Taiwan，TAT）。其包括 TAT-Vol1~2，共有 200 位語者，約 100 小時的台語語音辨認/多語者語音合成用語料，與 TAT-TTS-M1~2 與 TAT-TTS-F1~2，考慮台語強勢（高雄）與次強勢（台北）腔，共兩男兩女，每人約 10 小時的台語語音合成專用語料。以下說明 TAT 語料庫的設計，錄音標準作業程序，人工校正作業程序與語料庫發行等程序。

3.1 語料庫目的與設計 (Corpus Design)

我們的目的是想要製作一個能反應台文與台語使用現況，並涵蓋大部分台語腔調的台文語音語料庫。因此我們需要收集豐富的原生台語文本，並在台灣各地招募台語發音人，進行語料收集。並整理成具錄音資訊後製資料（metadata）的電子化語料庫。所以，此台語語音語料庫設計如下：

- 目標台語文本內容：使用多種題材的台語原生短文章（不用翻譯文章），以充分反應台文與台語使用現況。並包含日常生活對話、數字，地址等常用語句，以利開發語音辨認與語音合成應用。
- 目標台語發音語者：在台灣各區域招募不同性別與年紀的當地台語語者，並要求其照自己平常的習慣自然發音，以蒐集台灣不同區域的人真實使用的不同台語腔調。
- 目標錄音環境：針對語音辨認應用，需使用多種麥克風，在一般安靜辦公室環境錄音，以收集多種麥克風通道與環境變化。針對語音合成應用，則需進具高階隔音與殘響抑制的專業錄音室，以高階電容式麥克風，錄制無任何背景噪音與殘響的高品質音檔。
- 電子化檔案格式：語料庫必須包含錄音資訊後製資料的電子化檔案格式，最終形式需為一個 Microsoft Waveform 格式的語音音檔，配合一個對應的 json 格式文字檔，其中記錄了人工校正後產生的台羅數字調，音檔長度或是錄音者使用的腔調等相關資訊，檔案格式範例如圖 11 所示。



```
{
    "音檔長度": "9.61",
    "漢羅台文": "我厝內的電話是空二三三六六九空五四",
    "台羅": "guá tshù-lāi ê tiān-uē sī khòng jī sam sam lio̍k lio̍k kiú khòng ngóo sù",
    "台羅數字調": "gua2 tshu3-lai7 e5 tian7-ue7 si7 khong3 ji7 sam1 sam1 liok8 liok8 kiu2 khong3 ngoo2 su3",
    "白話字": "góa chhù-lāi ê tiān-ōe sī khòng jī sam sam lio̍k lio̍k kiú khòng ngó͘ sù",
    "字數": "17",
    "提示卡編號": "0011",
    "句編號": "1.1",
    "發音人": "IUF001",
    "性別": "女",
    "年齡": "51",
    "教育程度": "博士",
    "出生地": "高雄市",
    "現居地": "台北市文山區",
    "腔調": "漳泉濫",
    "錄音環境": "安靜隔音室內",
    "提示卡切換速度": "快",
    "總錄音時間(分)": "90"
}
```

圖 11. TAT 語料庫 json 文檔範例
[Figure 11. Recording Metadata Example]

3.2 錄音作業協定 (Recording Protocol)

依據上述語料庫目的，我們訂定了以下語料庫建置的錄音標準作業流程，如圖 12 所示，包括文本蒐集、提示卡製作、發音人招募、發音人錄音預備、實際錄音、人工校正與語料庫釋出等流程。以下將分別介紹各程序的進行方式。

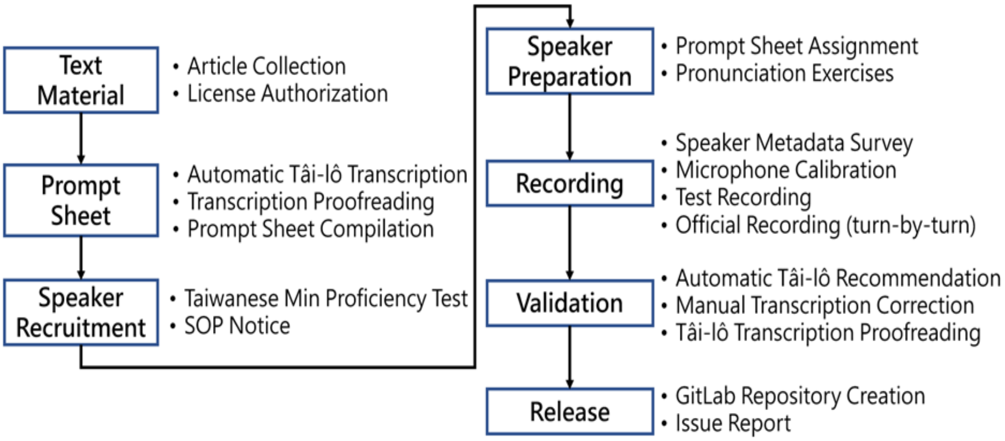
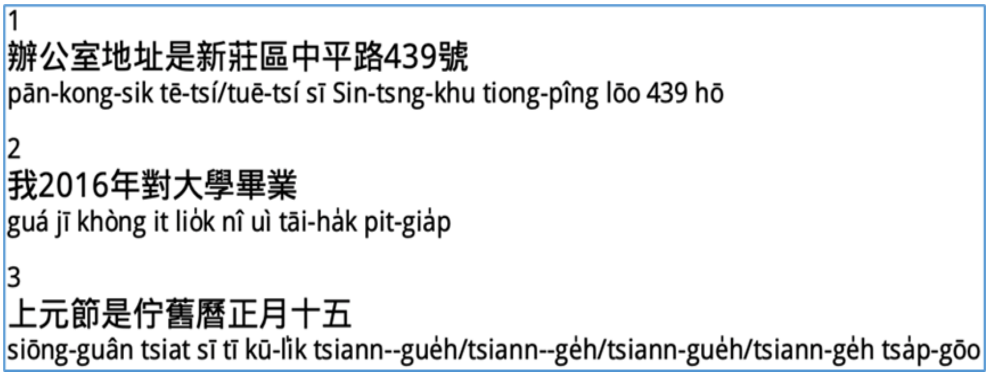

Text Material
• Article Collection
• License Authorization
Prompt Sheet
• Automatic Tâi-lô Transcription
• Transcription Proofreading
• Prompt Sheet Compilation
Speaker Recruitment
• Taiwanese Min Proficiency Test
• SOP Notice
Speaker Preparation
• Prompt Sheet Assignment
• Pronunciation Exercises
Recording
• Speaker Metadata Survey
• Microphone Calibration
• Test Recording
• Official Recording (turn-by-turn)
Validation
• Automatic Tâi-lô Recommendation
• Manual Transcription Correction
• Tâi-lô Transcription Proofreading
Release
• GitLab Repository Creation
• Issue Report

圖 12. TAT 語料庫建置流程
[Figure 12. The recording protocol of the TAT corpus]

3.2.1 台語原生文本蒐集與提示卡製作 (Native Taiwanese Prompt Sheets)

錄音的文本的蒐集來源，主要來自於李江卻台語文教基金會 [1]，由其聯繫曾在基金會出版品發表文章的 50 位作者，每人蒐集約 6000 字的文本。再加上基金會本身出版的台語日常對話課程教材，與程式依樣板隨機產生的數字串、電話號碼與地址等常用語句。

依據上述文本，製作出的約 50 份提示卡，其內容包含三大類，範例如圖 13，圖 14 以及圖 15 所示，分別為(1)數字資料，像是地址、日期、電話等等，還有(2)日常對話以及(3)短文等，分別以句或是短段落（1~3 短句）為單位編號條列之。每一個句子，除提供台文文句外，並標上參考（不限制其發音）用的對應台羅拼音文句，以便不熟習台文


1
辦公室地址是新莊區中平路439號
pān-kong-sik tē-tsí/tuē-tsí sī Sin-tsng-khu tiong-pîng lōo 439 hō

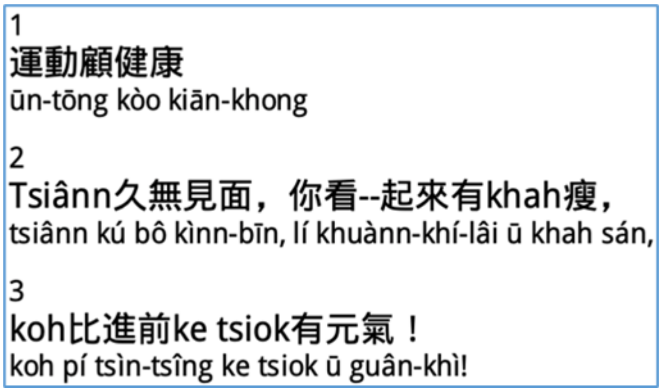
2
我2016年對大學畢業
guá jī khòng it lio̍k nî uì tāi-ha̍k pit-gia̍p

3
上元節是佇舊曆正月十五
siōng-guân tsiat sī tī kū-li̍k tsiann--gue̍h/tsiann--ge̍h/tsiann-gue̍h/tsiann-ge̍h tsa̍p-gōo

圖 13. 錄音提示卡-數字資料部分範例
[Figure 13. A typical exampleof prompt sheet: address, date and digits]

[1] i kang khioh taiwanese cultural and educational foundation
https://www.tgb.org.tw/

的發音人事先溫稿。最後，每份提示卡內容長度，則是設定成一人份，總共約錄製出 30 分鐘語音檔。

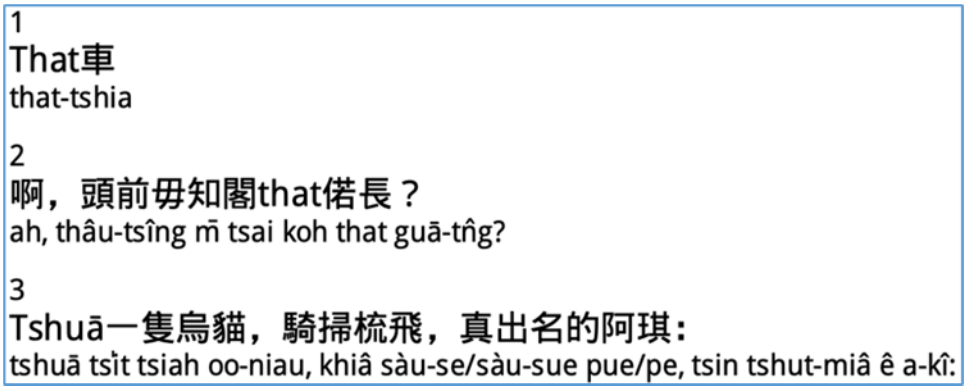


1
運動顧健康
ūn-tōng kòo kiān-khong

2
Tsiânn久無見面，你看--起來有khah瘦，
tsiânn kú bô kìnn-bīn, lí khuànn-khí-lâi ū khah sán,

3
koh比進前ke tsiok有元氣！
koh pí tsìn-tsîng ke tsiok ū guân-khì!

圖 14. 錄音提示卡-日常對話部分範例
[Figure 14. A typical exampleof prompt sheet: daily conversation session]



1
That車
that-tshia

2
啊，頭前毋知閣that偌長？
ah, thâu-tsîng m̄ tsai koh that guā-tn̂g?

3
Tshuā一隻烏貓，騎掃梳飛，真出名的阿琪：
tshuā tsi̍t tsiah oo-niau, khiâ sàu-se/sàu-sue pue/pe, tsin tshut-miâ ê a-kî:

圖 15. 錄音提示卡-短文部分範例
[Figure 15. A typical exampleof prompt sheet: short article session]

3.2.2 ASR錄音程序 (Recording Procedure for ASR Corpus)

3.2.2.1 錄音員招募與錄音員預備 (Speaker Recruitment)

錄音提示卡準備好後，我們與台灣各地的教授合作，就近在其學校附近，招募適合的台語發音人，並指定其所使用的提示卡的編號，要求其先溫稿後，約定時間進行錄音。以涵蓋全台灣各地使用各種不同台語腔調的台語使用者。在 TAT-Vol1~2 語料庫錄製過程中，我們共與五位教授合作，包括師範大學許慧如教授，台中教育大學楊允言教授及程俊源教授，中正大學蔡素娟教授與成功大學陳麗君教授，其學校所在區域如圖 16 所示，每位教授負責招募與錄製約 50 位語者，因此一份提示卡，最多只會能給四個人使用。

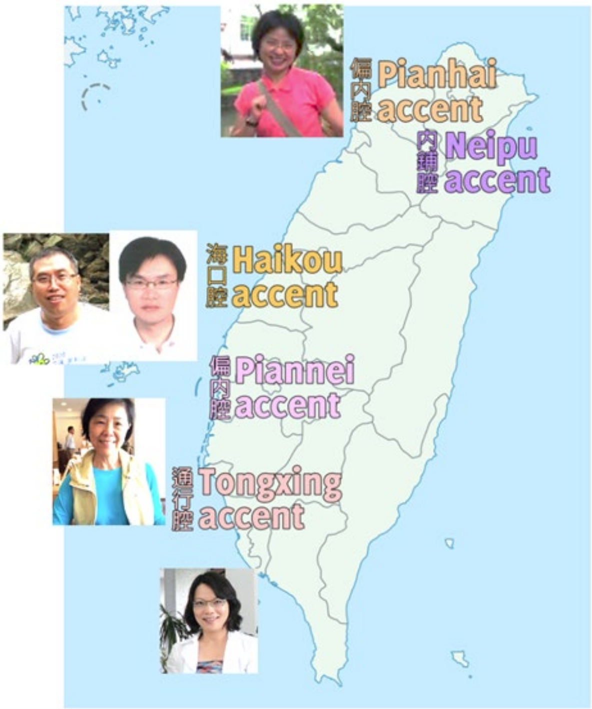


圖 16. TAT 語料庫錄製合作教授分布區域圖

[Figure 16. Distribution of locations of recording campuses]

3.2.2.2 錄音設備配置 (Equipment Configuration)

ASR 語料蒐集，通常選在安靜的一般辦公室，會議室或教室作為錄音環境。錄音時，為了模擬不同的使用情境下錄到的聲音，整個錄音設備配置如圖 17 所示，使用筆電，透過 USB 介面，連接 Zoom H6[2]多軌數位錄音介面，同時抓取 6 隻麥克風的訊號，一次錄製 6 軌音檔，以蒐集不同麥克風與錄音通道的影響。

[2] ZOOM CORPORATION, H6 Handy Recorder Operation Manual, Available: https://www.zoom.co.jp/sites/default/files/products/downloads/pdfs/E_H6v2.pdf

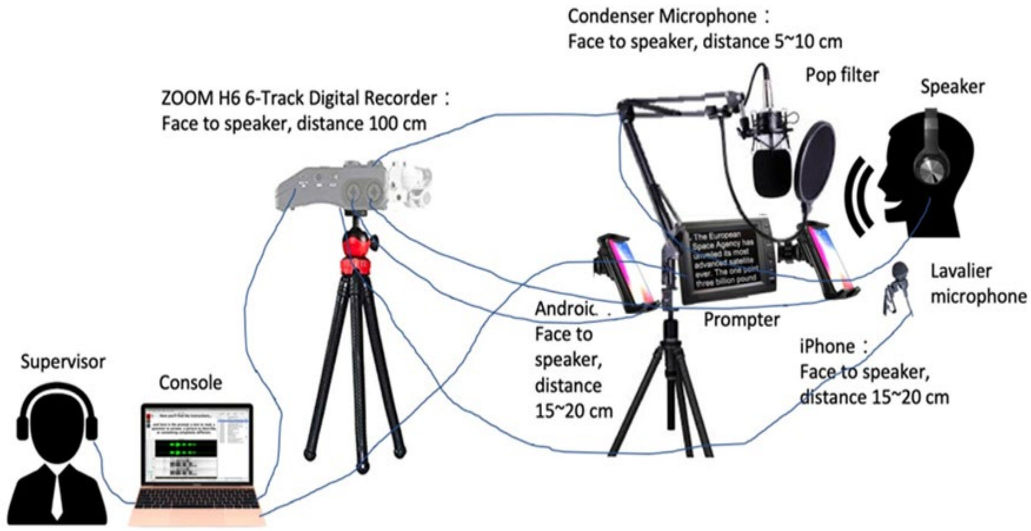


圖 17. TAT 語料庫錄製情形模擬圖
[Figure 17. Configuration of the recording equipments]

其中，Zoom H6 的最高取樣頻率，可以設置到 192 kHz。而此六隻麥克風的模擬使用情境，可分為三類，第一種是距離最遠的兩個麥克風，約距離錄音者 1 公尺，分別為 ZOOM XYH-6 左聲道和 ZOOM XYH-6 右聲道，用以模擬遠距麥克風收音效應，其會含有較多空間殘響與背景噪音：第二種是距離最近的兩個麥克風，約距離錄音者 5 到 10 公分，分別為放在發音人嘴巴正前方的電容式麥克風，和別在胸口的領夾式麥克風，用以收集近場語音，其聲音應該最為乾淨清楚：第三種是約距離錄音者 15 到 20 公分，位居下方偏右的 ios 手機，以及下方偏左的 android 手機的內建麥克風，分別用來呈現一般手機使用情形下的收音情況。

3.2.2.3 錄音程序 (Recording Procedures)

錄音現場示意圖如圖 18 所示，由一名監錄員使用筆電，執行德國慕尼黑大學開發的 SpeechRecoder 語料蒐集錄音軟體 [3]，一次一句將提示卡內容，投放至發音人正前方的提示卡螢幕（外接螢幕），給發音人觀看。請發音人按照提示卡螢幕上的文句，進行錄音。同時，監錄員利用耳機，與筆電主螢幕上 SpeechRecoder 錄音軟體所畫出的六軌音檔波形，進行音檔監聽與波形監看。確認每次錄音，音檔聲音的大小聲，錄音說話內容及說話順暢度皆正確後，再接續投放提示卡中的下一個文句。如此重複進行直至全部錄音文本錄音完畢。

[3] SpeechRecorder. Available: https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

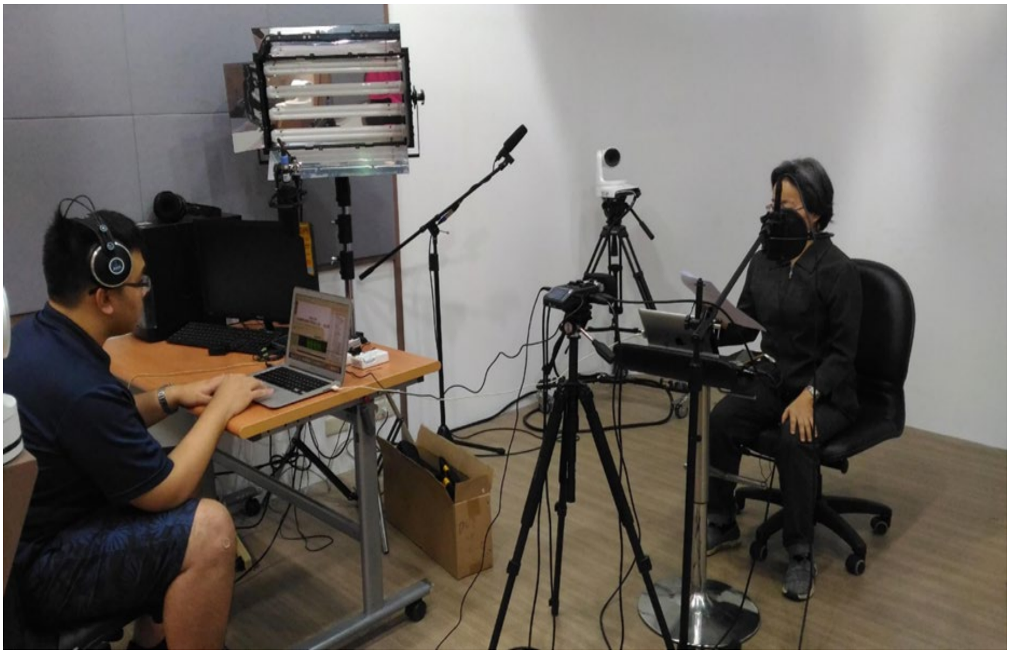


圖 18. ASR 錄音現場示意圖
[Figure 18. Photo of the recording site]

3.2.3 TTS錄音程序 (Recording Procedures for TTS)

3.2.3.1 錄音員招募與錄音員預備 (Speaker Recruitment)

TTS 的部分則是請李江却台語文教基金會招募了共兩男兩女 4 名語者，代號分別為 M1，M2，F1 與 F2，4 位語者的相關資訊如表 1 到表 4 所示。其中 M1 以及 F1 語者的腔調為台灣台語強勢腔（偏漳州腔/高雄腔），M2 跟 F2 則為次強勢腔（偏泉州腔/台北腔）。此四位發音人需把所有提示卡全部念完，完成共約十小時的台語語音，以蒐集足夠台語語音合成用語料。

表 1. M1 語者資訊
[Table 1. Personal information about Speaker M1]

性別	年齡	教育程度	出生地	現居地	腔調
男	34 歲	大學	台北市士林區	台北市士林區	偏漳州腔

表 2. M2 語者資訊
[Table 2. Personal information about Speaker M2]

性別	年齡	教育程度	出生地	現居地	腔調
男	55 歲	大學	新北市汐止區	新北市汐止區	泉州安溪腔

表 3. F1 語者資訊
[Table 3. Personal information about Speaker F1]

性別	年齡	教育程度	出生地	現居地	腔調
女	52 歲	碩士	高雄市新興區	新北市新店區	漳州腔

表 4. F2 語者資訊
[Table 4. Personal information about Speaker F2]

性別	年齡	教育程度	出生地	現居地	腔調
女	41 歲	碩士	台中市梧棲區	台北市中正區	泉州腔

3.2.3.2 錄音設備配置 (Equipment Configuration)

與 ASR 不同， TTS 用合成語料必須排除背景雜音，讓聲音越乾淨越好。因此 TTS 專用語料是在能隔絕背景底噪（如圖 19 左方所示）及抑制空間殘響（如圖 19 右方所示）的專業錄音室中錄音。並且錄製 TTS 時必須使用音質最好的高階電容式麥克風，近距離進行錄製。

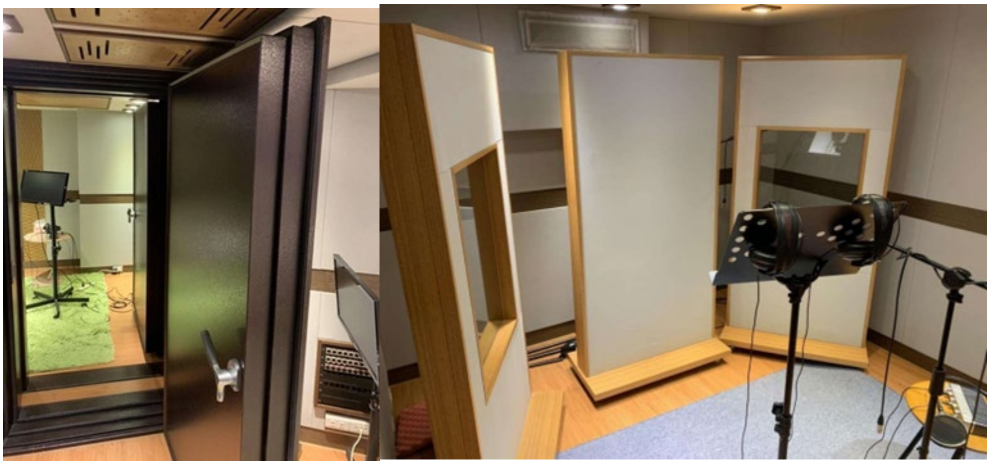


圖 19. 專業錄音室示意圖
[Figure 19. Photos of the professional recording studio]

3.2.3.3 錄音程序 (Recording Procedures)

TTS 語料蒐集的錄音現場示意圖如圖 20 所示。需由一名音響工程師，操作專業錄音工作站軟體，確認音檔聲學特性平穩一致（大小聲、速度與韻律等等）。此外，並需再加上一名具台語老師等級的發音監錄員，同時一句一句監督每句台語發音的正確性。此外，第一次錄音時，需先錄製數句校正句，作為範本，在每次開始新的錄音工作（session）前，播放給發音人聆聽，讓發音人校正對齊其發音特性。而且，一次錄音工作的時間不能太長，必須讓發音人有足夠休息時間，以維持其發音特型一致，不要偏掉。

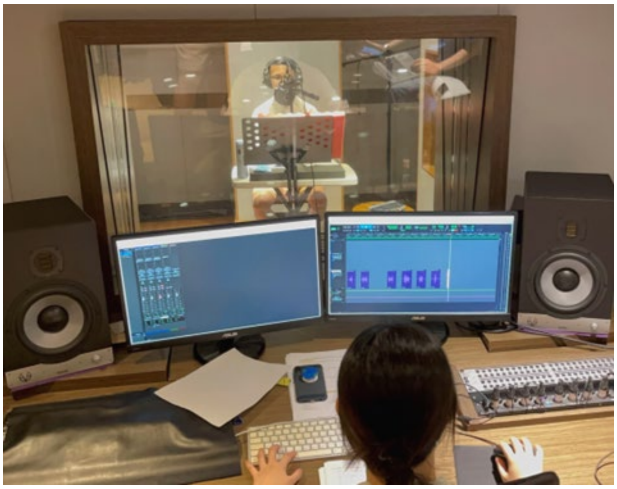
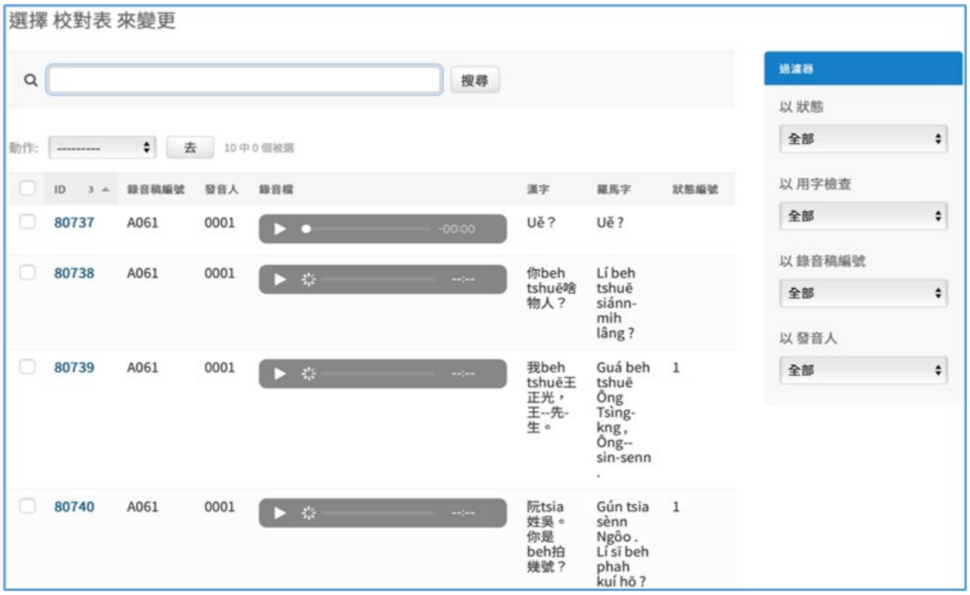


圖 20. TTS 錄音現場示意圖
[Figure 20. Photo of the recording workstation]

3.2.4 人工校正 (Transcription)

錄製好的語音檔，最後分別由五位合作教授的團隊（負責 TAT-Vol1~2），與李江卻台語文教基金會 的工作人員（負責 TAT-TTS-M1~2 與 TAT-TTS-F1~2），利用意傳科技的線上語料庫校正輔助工具（如圖 21 所示），以人工逐句聆聽校正文字檔，與依據發音人的實際發音，標上台羅正確拚音，產生最終的語料庫。



選擇 校對表 來變更

□	ID 3 ▲	錄音稿編號	發音人	錄音檔	漢字	羅馬字	狀態編號
□	80737	A061	0001	▶ -00:00	Uě？	Uě？	
□	80738	A061	0001	▶ --:--	你beh tshuē啥物人？	Lí beh tshuē siánn-mih lâng？	
□	80739	A061	0001	▶ --:--	我beh tshuē王正光，王--先-生。	Guá beh tshuē Ông Tsìng-kng, Ông--sin-senn.	1
□	80740	A061	0001	▶ --:--	阮tsia 姓吳。你是beh拍幾號？	Gún tsia sènn Ngôo. Lí sī beh phah kuí hō？	1

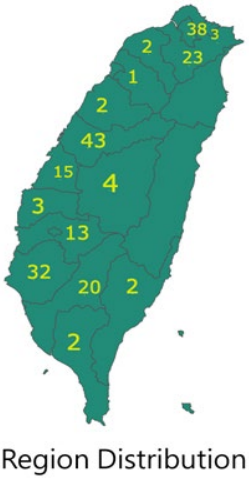
圖 21. 意傳科技語料庫校正輔助工具示意圖
[Figure 21. User interface of the online corpus annotation tool]

3.3 語料庫內容統計 (Statistics of TAT Corpus)

以下介紹完成的 TAT-Vol11~2 與 TAT-TTS-M1~2 與 TAT-TTS-F1~2 語料庫的內容統計資料。

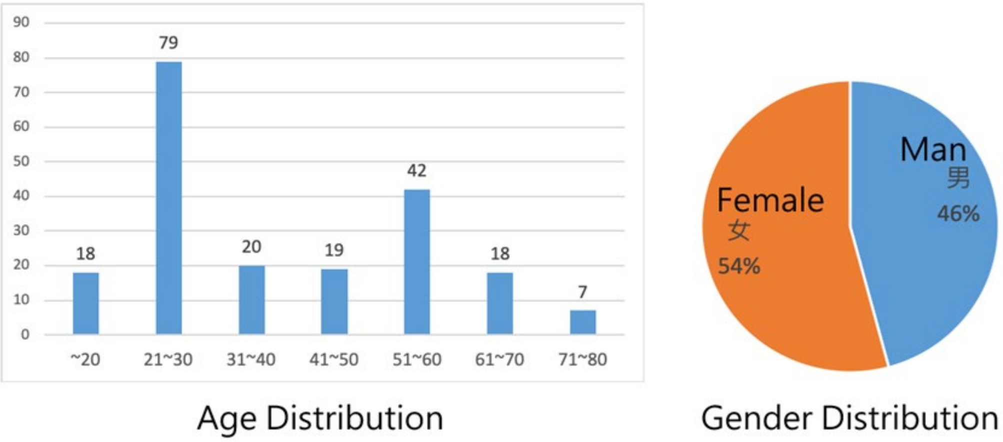
3.3.1 TAT-Vol1~2

在台語語音辨認用語料蒐集部分，經過一年的辛苦錄音後，最終招募了台灣各地使用各種不同台語腔調的台語語者共約 200 人，包括男生 91 人，女生 109 人，其在台灣的地域/人數分布如圖 22 所示，年齡分佈從 18 到 80 歲都有，分佈如圖 23 所示。



Region Distribution

圖 22. TAT-Vol1~2 語料庫，錄音語者地域分布
[Figure 22. Distribution of speakers in TAT-Vol1~2 corpus]

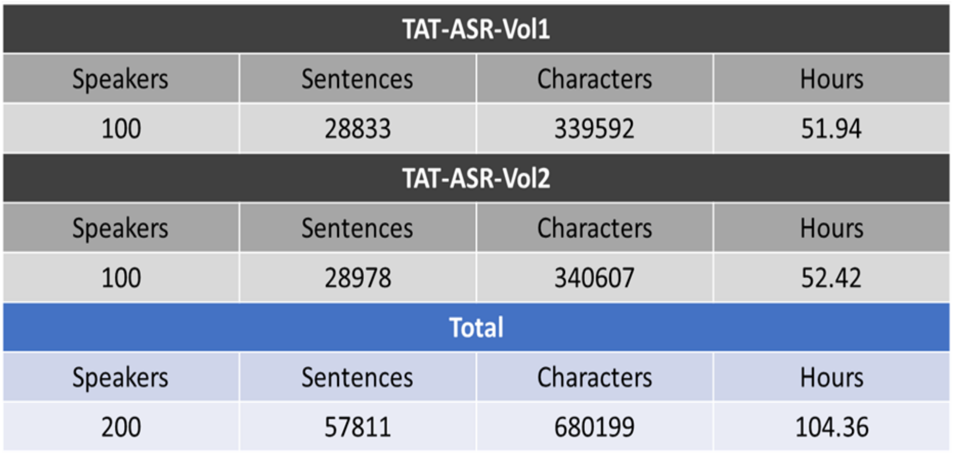


Age Distribution

Gender Distribution

圖 23. TAT-ASR 錄音語者年齡分布及性別比
[Figure 23. Distribution of ages and sexs of recorded speakers in TAT-Vol1~2 corpsu]

目前完成的語料庫總時數共約 104.36 小時，經均分成 TAT-Vol1~2 兩集語料庫，委託『社團法人中華民國計算語言學學會』[4]公開發行。各子集的詳細的句數，字數與時數統計數據如圖 24 所示。

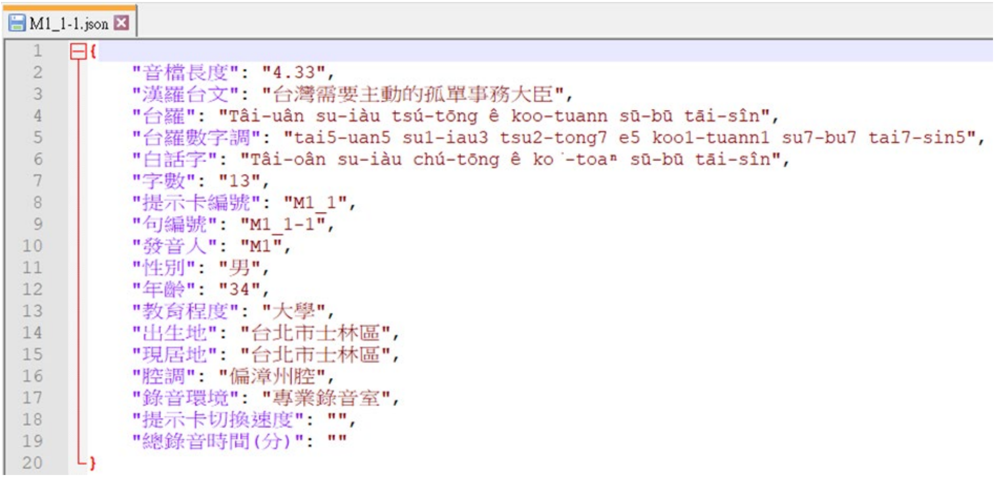


TAT-ASR-Vol1			
Speakers	Sentences	Characters	Hours
100	28833	339592	51.94
TAT-ASR-Vol2			
Speakers	Sentences	Characters	Hours
100	28978	340607	52.42
Total			
Speakers	Sentences	Characters	Hours
200	57811	680199	104.36

圖 24. TAT-ASR 語料庫詳細統計數據
[Figure 24. Statistics of the TAT-Vol1~2 corpus]

3.3.2 TAT-TTS-M1~2與TAT-TTS-F1~2 (TAT-TTS-M1~2 and TAT-TTS-F1~2)

在台語語音合成專用的語料庫方面，最終錄製了 2 名男生和 2 名女生的語料。語料庫完成品形式同樣為一個音檔，配對一個相同檔名的 json 檔。此 json 檔的格式，如圖 25 所示，含有漢羅台文、台羅數字調、音檔長度或是發音人的屬性等相關資訊。

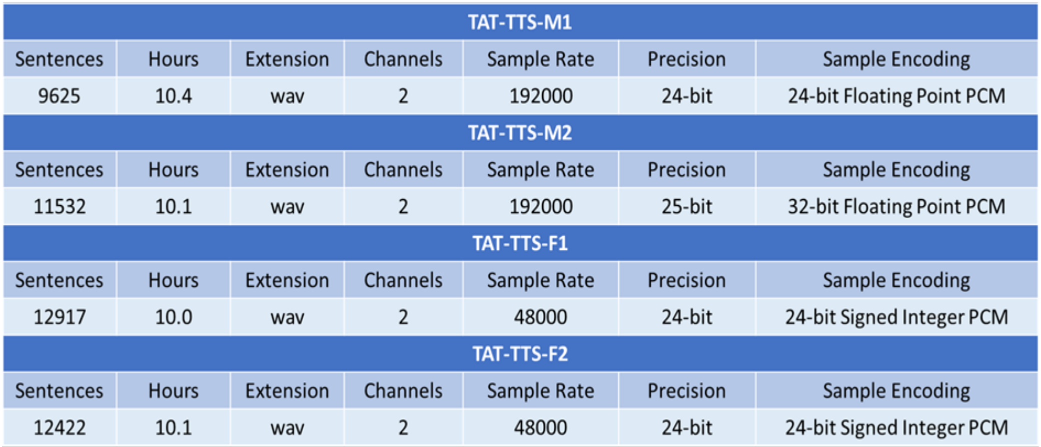


```
{
    "音檔長度": "4.33",
    "漢羅台文": "台灣需要主動的孤單事務大臣",
    "台羅": "Tâi-uân su-iàu tsú-tōng ê koo-tuann sū-bū tāi-sîn",
    "台羅數字調": "tai5-uan5 su1-iau3 tsu2-tong7 e5 koo1-tuann1 su7-bu7 tai7-sin5",
    "白話字": "Tâi-oân su-iàu chú-tōng ê ko͘-toaⁿ sū-bū tāi-sîn",
    "字數": "13",
    "提示卡編號": "M1_1",
    "句編號": "M1_1-1",
    "發音人": "M1",
    "性別": "男",
    "年齡": "34",
    "教育程度": "大學",
    "出生地": "台北市士林區",
    "現居地": "台北市士林區",
    "腔調": "偏漳州腔",
    "錄音環境": "專業錄音室",
    "提示卡切換速度": "",
    "總錄音時間(分)": ""
}
```

圖 25. TAT-TTS json 文檔範例
[Figure 25. A typical example of the recording metadata]

[4] https://www.aclclp.org.tw/corp_c.php

此語料也已經委託『社團法人中華民國計算語言學學會』公開發行。目前共完成 4 位語者，每位語者約 10 小時的語料，總時數共約 40.6 小時。音檔詳細資訊如圖 26 所示，以人為單位分成四集，包括 TAT-TTS-M1~2 與 TAT-TTS-F1~2，其中 M 與 F 分別為男生與女生語者的代號，1 與 2 則分別為強勢腔與次強勢腔的編碼。



TAT-TTS-M1						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
9625	10.4	wav	2	192000	24-bit	24-bit Floating Point PCM

TAT-TTS-M2						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
11532	10.1	wav	2	192000	25-bit	32-bit Floating Point PCM

TAT-TTS-F1						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
12917	10.0	wav	2	48000	24-bit	24-bit Signed Integer PCM

TAT-TTS-F2						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
12422	10.1	wav	2	48000	24-bit	24-bit Signed Integer PCM

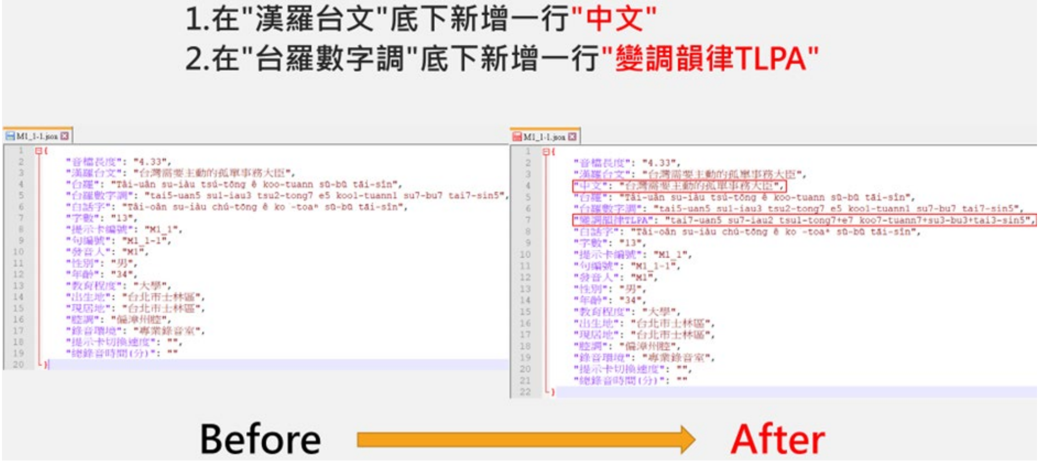
圖 26. TAT-TTS-M1~2 與 TAT-TTS-F1~2 語料庫的詳細音檔資料與句數、時數等統計資訊
[Figure 26. Statistics of the TAT-TTS-M1~2 and TAT-TTS-F1~2 corpus]

3.4 台語變調與韻律邊界標註 (Annotation of Tone and Prosodic Boundary)

我們以 TAT-TTS 的 M1 語者作為訓練語料，完成單人台語語音合成系統後，發現合成音有時有台語變調與停頓不通順的問題，這主要是當初因人力問題，在建置 TAT-TTS-M1~2 與 TAT-TTS-F1~2 時，只以台語本調做標記，並沒有針對台語變調，進行人工校正，或是加上韻律邊界標注。因此，我們針對 TAT-TTS-M1 語料庫，進一步加上中文翻譯、標註台語變調與台語韻律詞或韻律片語邊界。

3.4.1 語料庫設計 (Corpus Design)

我們訂定了以下的變調與韻律標註標準作業程序。標註方法為在原先語料的 json 檔裡面新增兩行字串，分別是(1)中文翻譯，(2)台語變調校正以及（3）加入兩種新韻律邊界符號的台羅數字調。新增標註資訊後的 json 檔的前後比較如圖 27 所示。



1.在"漢羅台文"底下新增一行"中文"
2.在"台羅數字調"底下新增一行"變調韻律TLPA"

M1_1-1.json

```
{
    "音檔長度": "4.33",
    "漢羅台文": "台灣需要主動的孤單事務大臣",
    "台羅": "Tâi-uân su-iàu tsú-tōng ê koo-tuann sū-bū tāi-sîn",
    "台羅數字調": "tai5-uan5 su1-iau3 tsu2-tong7 e5 koo1-tuann1 su7-bu7 tai7-sin5",
    "白話字": "Tâi-oân su-iàu chú-tōng ê ko͘-toaⁿ sū-bū tāi-sîn",
    "字數": "13",
    "提示卡編號": "M1_1",
    "句編號": "M1_1-1",
    "發音人": "M1",
    "性別": "男",
    "年齡": "34",
    "教育程度": "大學",
    "出生地": "台北市士林區",
    "現居地": "台北市士林區",
    "腔調": "偏漳州腔",
    "錄音環境": "專業錄音室",
    "提示卡切換速度": "",
    "總錄音時間(分)": ""
}
```

M1_1-1.json

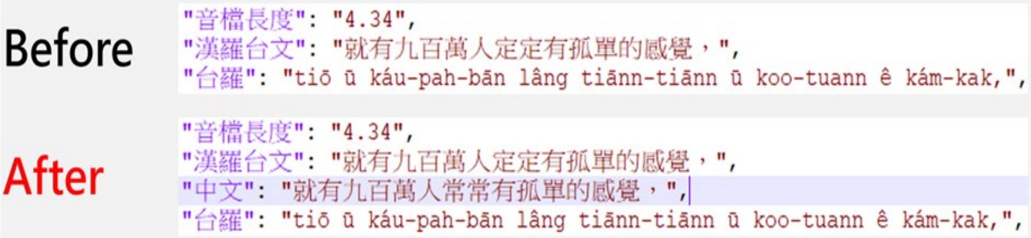
```
{
    "音檔長度": "4.33",
    "漢羅台文": "台灣需要主動的孤單事務大臣",
    "中文": "台灣需要主動的孤單事務大臣",
    "台羅": "Tâi-uân su-iàu tsú-tōng ê koo-tuann sū-bū tāi-sîn",
    "台羅數字調": "tai5-uan5 su1-iau3 tsu2-tong7 e5 koo1-tuann1 su7-bu7 tai7-sin5",
    "變調韻律TLPA": "tai7-uan5 su7-iau2 tsu1-tong7+e7 koo7-tuann7+su3-bu3+tai3-sin5",
    "白話字": "Tâi-oân su-iàu chú-tōng ê ko͘-toaⁿ sū-bū tāi-sîn",
    "字數": "13",
    "提示卡編號": "M1_1",
    "句編號": "M1_1-1",
    "發音人": "M1",
    "性別": "男",
    "年齡": "34",
    "教育程度": "大學",
    "出生地": "台北市士林區",
    "現居地": "台北市士林區",
    "腔調": "偏漳州腔",
    "錄音環境": "專業錄音室",
    "提示卡切換速度": "",
    "總錄音時間(分)": ""
}
```

圖 27. TAT-TTS 台語語料校正前後比較
[Figure 27. Comparison of metadata before and after Chinese Text, tone and prosodic boundary annotation]

其中加入中文是為了以後能進行中文轉台文與中文轉台羅拼音兩種機器翻譯。而校正變調與加入自訂的兩種新韻律符號，是為了讓語音合成系統，可以學習變調規則與韻律停頓方式。

3.4.2 人工標註程序協定 (Annotation Protocol)

校正者應先聆聽每一句台語音檔，以人工整句翻譯成對應的中文文字，並在 json 中加入一行中文文字標註。範例如圖 28 所示。



Before

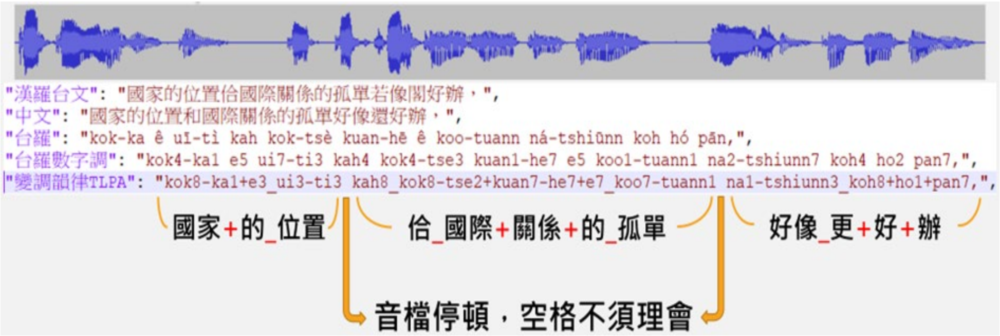
```
"音檔長度": "4.34",
"漢羅台文": "就有九百萬人定定有孤單的感覺，",
"台羅": "tiō ū káu-pah-bān lâng tiānn-tiānn ū koo-tuann ê kám-kak,",
```

After

```
"音檔長度": "4.34",
"漢羅台文": "就有九百萬人定定有孤單的感覺，",
"中文": "就有九百萬人常常有孤單的感覺，",
"台羅": "tiō ū káu-pah-bān lâng tiānn-tiānn ū koo-tuann ê kám-kak,",
```

圖 28. TAT-TTS 台語語料之加入中文翻譯範例
[Figure 28. A typical example of Chinese translation]

此外，校正者需一句一句聆聽台語音檔中的變調與韻律邊界結構現象，首先將需要變調的數字調進行更正。接下來必須留意音檔停頓的位置，語音停頓的地方如為空格或標點符號為正常，不須理會。講者連續念過去的地方如為連字號或是輕聲符號(雙重連字號)，也不須理會。但如連續念過去的地方為空格，則需將空格取代為適當的韻律符號，經討論後我們定義出兩種新的韻律符號，分別為(1)韻律詞，代表符號為加號"+"，(2)韻律片語，代表符號為底線"_"。json 檔校正完成後的範例如圖 29 所示，此範例音檔為 TAT-TTS-M1 裡面的 M1_1-6.wav 音檔。

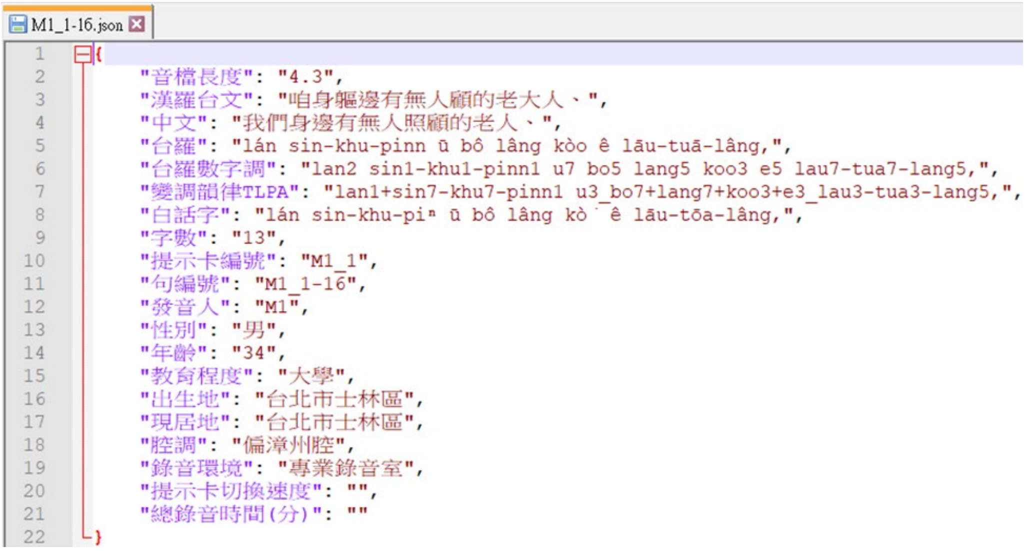


```
"漢羅台文": "國家的位置佮國際關係的孤單若像閣好辦，",
"中文": "國家的位置和國際關係的孤單好像還好辦，",
"台羅": "kok-ka ê uī-tì kah kok-tsè kuan-hē ê koo-tuann ná-tshiūnn koh hó pān,",
"台羅數字調": "kok4-ka1 e5 ui7-ti3 kah4 kok4-tse3 kuan1-he7 e5 koo1-tuann1 na2-tshiunn7 koh4 ho2 pan7,",
"變調韻律TLPA": "kok8-ka1+e3_ui3-ti3 kah8_kok8-tse2+kuan7-he7+e7_koo7-tuann1 na1-tshiunn3_koh8+ho1+pan7,",
```

圖 29. TAT-TTS 台語語料之變調與韻律符號校正範例
[Figure 29. A typical example of tone sandhi and prosodic boundary annotation]

3.4.3 語料庫完成品統計 (Statistics of Corpus)

經過校正人員的努力，目前已經將 M1 語者的語料全數校正完畢，校正完成後的部分 json 文檔範例如圖 30 所示。後續也將針對剩下的 M2，F1 和 F2 語者的語料進行校正。



```
{
        "音檔長度": "4.3",
        "漢羅台文": "咱身軀邊有無人顧的老大人、",
        "中文": "我們身邊有無人照顧的老人、",
        "台羅": "lán sin-khu-pinn ū bô lâng kòo ê lāu-tuā-lâng,",
        "台羅數字調": "lan2 sin1-khu1-pinn1 u7 bo5 lang5 koo3 e5 lau7-tua7-lang5,",
        "變調韻律TLPA": "lan1+sin7-khu7-pinn1 u3_bo7+lang7+koo3+e3_lau3-tua3-lang5,",
        "白話字": "lán sin-khu-piⁿ ū bô lâng kò͘ ê lāu-tōa-lâng,",
        "字數": "13",
        "提示卡編號": "M1_1",
        "句編號": "M1_1-16",
        "發音人": "M1",
        "性別": "男",
        "年齡": "34",
        "教育程度": "大學",
        "出生地": "台北市士林區",
        "現居地": "台北市士林區",
        "腔調": "偏漳州腔",
        "錄音環境": "專業錄音室",
        "提示卡切換速度": "",
        "總錄音時間(分)": ""
}
```

圖 30. TAT-TTS-M1 校正後 json 文檔範例
[Figure 30. A typical example of recording metadata]

4. 中文文字轉台語語音合成系統 (Chinese Text to Taiwanese Speech Synthesis System)

在踏入較複雜的台語語音轉換系統前，我們先以做出單人台語語音合成系統為目標，從中汲取台語語音合成相關的經驗，後面再繼續做多人台語語音合成系統，跟台語語音轉換系統。此外，因大多數人無法讀寫台文或是台羅拼音，因此我們額外在台語語音合成系統的前端，再加上一個中文文字轉台羅拼音的機器翻譯模組，製作一『中文文字轉台語語音合成系統』。

4.1 系統架構 (System Architecture)

此次單人台語語音合成系統的建置，以前端的中文轉台羅拼音（Chinese to Taiwanese Tâi-Lô Pinyin (TLPI), C2T）機器翻譯模組，加上後端的 Tacotron2+WaveGlow 語音合成架構為 baseline。訓練 Tacotron2 的台語語料則選用 TAT-TTS-M1 的男生強勢腔語者，並選擇 json 文檔中「台羅數字調」當作訓練文本，機器翻譯使用語料則為開源之 iCorpus 臺華平行新聞語料庫漢字臺羅版(Sih4, 2015)。具體系統架構如圖 31 所示，使用者輸入中文後，C2T 將中文轉換成台羅拼音（依據『臺灣閩南語羅馬字拼音方案』(教育部，2008)，台羅拼音作為文本輸入 Tacotron2 轉為頻譜，最後透過 WaveGlow 將頻譜即時的轉為波形並合成語音。

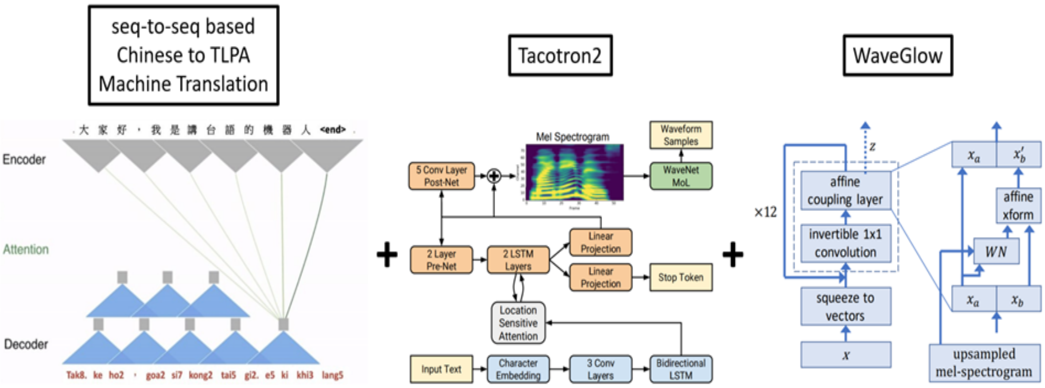


圖 31. 中文文字轉台語語音合成系統

[Figure 31. The architecture of the Chinese text to Taiwanese speech synthesis system]

4.1.1 中文文字轉台羅拼音機器翻譯 (Chinese to Taiwanese translation)

在訓練此機器翻譯所使用的中文對應台語拼音平行語料方面，使用了開源之 iCorpus 臺華平行新聞語料庫漢字臺羅版(Sih4, 2015)。並以上述語料為基礎，自行人工去除專有名詞，地名及人名等英文，以及少部分中文錯誤的地方，使訓練文本更為正確，經整理後得到 60323 句原始平行語料。而原始的 iCorpus 平行語料並不包含標點符號，為使成品之機器翻譯能夠看懂基本的標點符號，進而對語料進行以下處理。

做法為將文本複製六份後，每一份負責加入一種標點符號，分別為逗號、句號、驚嘆號、問號、冒號與分號六種標點符號。加入的方式為，每一句平行語料的結尾加入標點符號後，隨後接上一句隨機分配的平行語料，這種將標點符號加在前後為不一樣句子的方式，能讓標點符號的存在較為自然，使標點符號能正確地被訓練進去。最後將平行語料中文的部分以空格隔開每一個中文字，台羅拼音的部分連字號也改成以空格表示，讓平行語料以 phone 對 phone 的方式對應。以上全部完成處理後的部分範例如表 5 與表 6 所示。

表 5. iCorpus 平行語料中文部分範例
[Table 5. Examples of the Chinese transcriptions in the iCorpus corpus]

駐美特派員曹郁芬華府報導,出海捕獲一條將近三百公斤
蔡英文也說到時薪應該調整?每天得花四個小時在訓練上面
現在正好芒果盛產的季節:北部早晚低溫可能只有二十度上下

表 6. iCorpus 平行語料台羅拼音部分範例
[Table 6. Examples of the Taiwanese transcriptions in the iCorpus corpus]

tsu3 bi2 tik8 phai3 uan5 tso5 hiok4 hun1 hua5 hu2 po3 to7 , tshut4 hai2 liah8 tioh8 tsit8 tiau5 ua2 beh4 sann1 pah4 kong1 kin1
tshua3 ing1 bun5 ma7 tam5 kau3 si5 sin1 ing1 kai1 tiau5 tsing2 ? tak8 kang1 khai1 iong7 si3 tiam2 tsing1 ti7 hun3 lian7 bin7 ting2
tsit4 ma2 tu2 ho2 suainn7 a2 tua7 tshut4 e5 kui3 tsiat4 : pak4 poo7 tsa2 am3 ke7 un1 kho2 ling5 tsi2 u7 ji7 tsap8 too7 ting2 e7

準備好平行語料後，我們參考網路上開源的 fairseq 機器翻譯演算法(Hsu, 2021)，其基於 sequence-to-sequence 架構，網路如圖 32 所示，包括一 encoder 前端與一 decoder 後端。前端 encoder 負責接收輸入中文文字序列，分析其語意並擷取出文脈資訊向量。後端 decoder 在文脈資訊向量之間加入 attention 之機制與 convolutional neural network 之訓練模型下每個 encoder 權重，利用中文對應台語拼音平行語料庫進行訓練，以此得到最佳的轉譯台羅拼音序列。

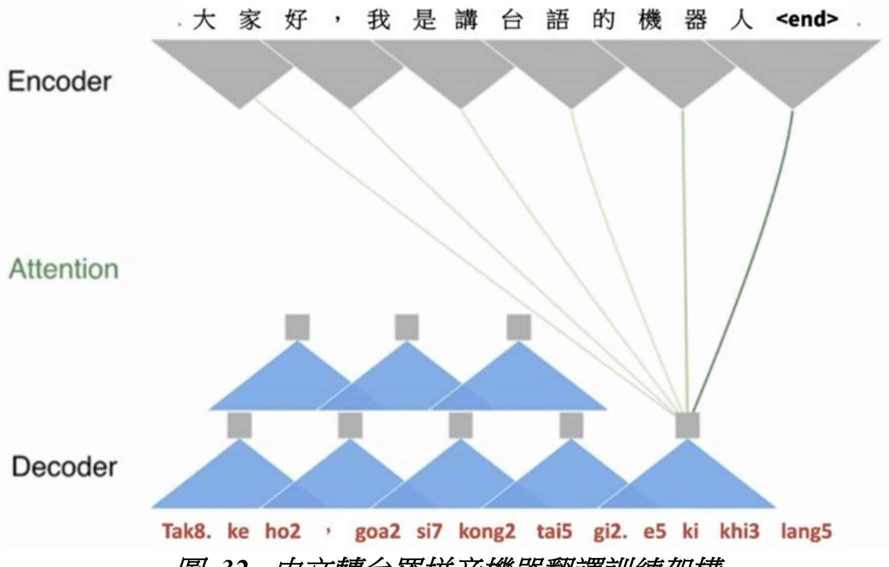

. 大 家 好 ， 我 是 講 台 語 的 機 器 人 <end> .
Encoder
Attention
Decoder
Tak8. ke ho2 ， goa2 si7 kong2 tai5 gi2. e5 ki khi3 lang5

圖 32. 中文轉台羅拼音機器翻譯訓練架構
[Figure 32. The architecture of the Chinese to Taiwanese machine translation system]

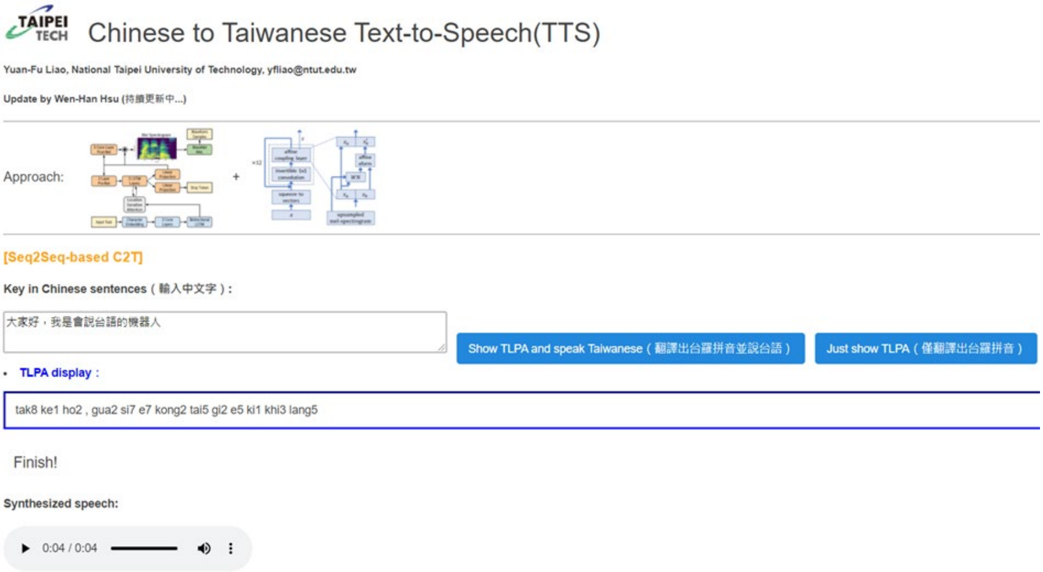
4.1.2 Tacotron2+WaveGlow

我們使用 TAT-TTS-M1 約 10.4 小時的台語語料，以 22050 赫茲的音檔取樣頻率，以及原始的「台羅數字調」當作文本，進行 Tacotron2 的訓練(Valle, 2020)。WaveGlow 聲碼器僅負責將頻譜合成語音，訓練時僅使用大量的音檔，不須配合文本，意即使用的語料語言與前端 Tacotron2 並無衝突。因此 WaveGlow 的部份，使用實驗室已經事先用數量與豐富度較多的英文語料 LJ Speech 訓練出的 WaveGlow 模型(Valle, 2020)，不須使用台語語料重新訓練。

另外我們也使用了校正好變調以及韻律符號的 TAT-TTS-M1 語料，裡面新增的"變調韻律 TLPI"作為訓練文本，將新增的兩種符號，加號"+"以及底線"_"新增進去訓練模型時考慮的特殊符號，訓練了一版考慮變調以及新增兩種韻律符號的 Tacotron2，作為後續實驗的比較。

4.1.3 雛型系統展示網頁 (Prototype System Demonstration)

我們將結合了中文轉台羅拼音機器翻譯的單人台語語音合成系統，做成了一展示網頁[5]，如圖 33 所示。使用者輸入中文文字後，按下合成按鈕就能撥放對應的台語語音，並能一併顯示出翻譯過後的台羅拼音供使用者查詢。且另外設計了可輸入台羅拼音的欄位，讓擁有相關台羅知識的使用者可以鍵入不同的發音並合成想要的台語語音。


TAIPEI TECH
Show TLPA and speak Taiwanese (翻譯出台羅拼音並說台語)
Just show TLPA (僅翻譯出台羅拼音)
0:04 / 0:04

Chinese to Taiwanese Text-to-Speech(TTS)

Yuan-Fu Liao, National Taipei University of Technology, yfliao@ntut.edu.tw

Update by Wen-Han Hsu (持續更新中...)

Approach:

[Seq2Seq-based C2T]

Key in Chinese sentences (輸入中文字) :

大家好，我是會說台語的機器人

- TLPA display :

tak8 ke1 ho2 , gua2 si7 e7 kong2 tai5 gi2 e5 ki1 khi3 lang5

Finish!

Synthesized speech:

圖 33. 中文文字轉台語語音合成系統展示網頁
[Figure 33. Demo website of the Chinese to Taiwanese machine translation]

[5] NTUT's Chinese to Taiwanese Text-to-Speech(TTS), http://tts001.iptcloud.net:8801/

5. 結合語音辨認及合成模組之多語者台語語音轉換系統 (Multi-Speaker Voice Conversion System based on Cascade ASR and TTS framework)

有了台語語音合成相關經驗的累積後，我們開始探究適合的台語語音轉換進行方法，目標做出一個初版可行的台語語音轉換系統。

5.1 同語言之台語對台語語音轉換系統建置 (Intra-Lingual Voice Conversion)

我們以 VCC 2020 中出現的 Cascade ASR and TTS 方法(Huang *et al.*, 2020)為 baseline，目標建置台語對台語語音轉換系統。具體系統架構如圖 34 所示，將來源語者的音檔，以台語語音辨認器轉成台羅拼音後，輸入已經預先以目標語者的語料微調過的台語多語者語音合成器，合成出符合來源語者文本以及目標語者音色的語音轉換音檔。

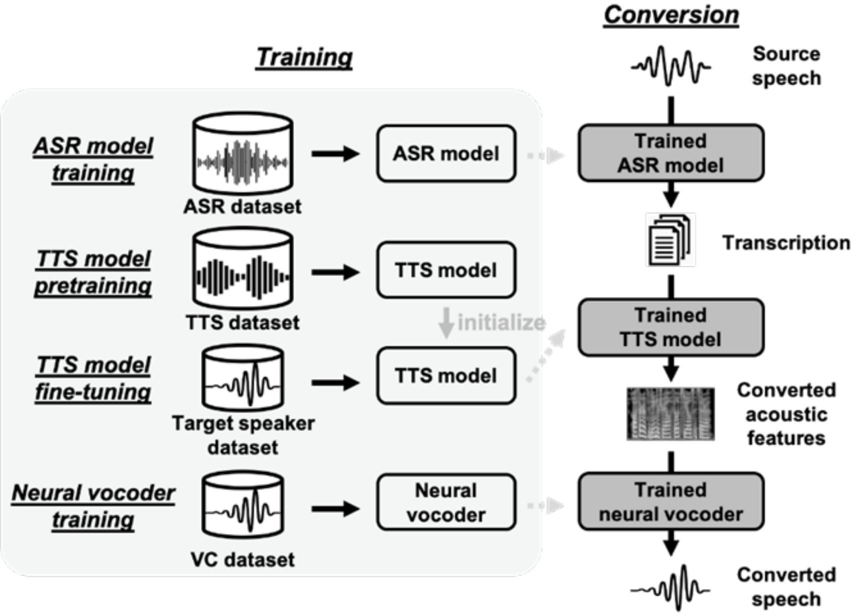
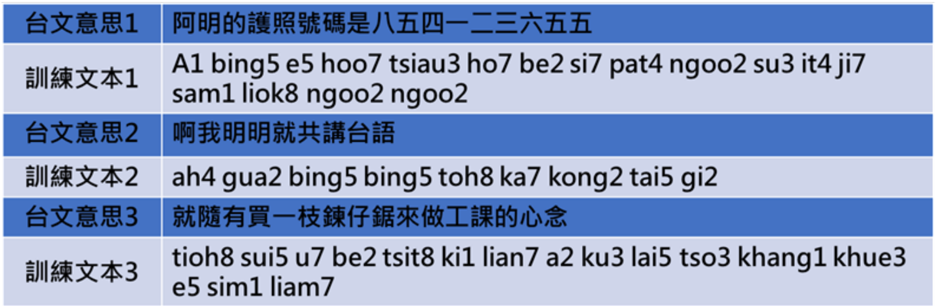


圖 34. 結合語音辨認及合成模組之台語轉台語語音轉換系統架構 (Huang et al., 2020)
[Figure 34. The architecture of the Taiwanese voice conversion system]

此方法需以三種預訓練模型為基礎，分別是(1)X-Vectors，(2)Transformer-based ASR model 以及(3)Multi-speaker Transformer-TTS model。在語音轉換的領域，因為涉及到語者辨識的技術，因此訓練語料的語者數量是越豐富越好。我們訓練三個預訓練模型使用的台語語料分別使用了(1)TAT-TTS- M1~2 與 TAT-TTS-F1~2 四位語者，共 2 男 2 女，每人約有 10 小時語料，總長度約為 40.6 小時，以及(2) TAT-Vol1~2 語料庫裡面的 200 位全部語者，共 91 男 109 女，每人約有半小時語料，總長度約為 104.4 小時。

在訓練文本的選擇上，我們一樣使用"台羅數字調"，並將文本的連字號取消，以較單純的字對字去訓練，部分範例如圖 35 所示，這樣一方面可以降低訓練的難度，也可以跟後面跨語言任務使用的華語語料做相同的對應。



台文意思1	阿明的護照號碼是八五四一二三六五五
訓練文本1	A1 bing5 e5 hoo7 tsiau3 ho7 be2 si7 pat4 ngoo2 su3 it4 ji7 sam1 liok8 ngoo2 ngoo2
台文意思2	啊我明明就共講台語
訓練文本2	ah4 gua2 bing5 bing5 toh8 ka7 kong2 tai5 gi2
台文意思3	就隨有買一枝鍊仔鋸來做工課的心念
訓練文本3	tioh8 sui5 u7 be2 tsit8 ki1 lian7 a2 ku3 lai5 tso3 khang1 khue3 e5 sim1 liam7

圖 35. 台語語料訓練文本部分範例

[Figure 35. A typical Example of the Taiwanese speech transcription data]

而後端聲碼器則沿用原本的 Parallel WaveGAN（PWG），原因同單人台語語音合成的 WaveGlow，聲碼器的部分不需要重新訓練。

5.1.1 語者向量編碼器 (Speaker Embedding)

使用了較為基礎的 X-Vectors 方法(Snyder *et al*., 2018)，如圖 36 所示。把輸入的語音截成多段，將每一小段語音信號輸出的特徵算一個 mean 以及 variance 並且 concat 起來，輸入 DNN 後來判斷這一小段語音是哪位語者的語者資訊，最後各小段語音的平均結果即為 speaker embedding。將 204 人的台語音檔以 train set 194 人，test set 10 人的設定進行語者向量編碼器的訓練(esdeboer, 2020)。

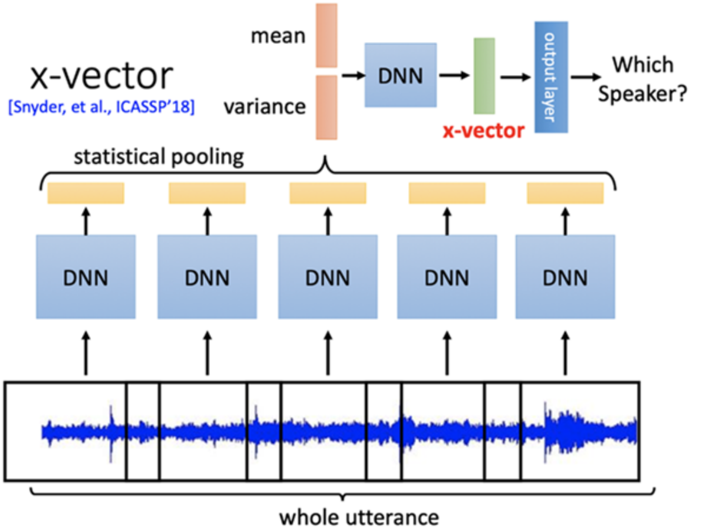


圖 36. X-Vectors 架構(Snyder et al., 2018)

[Figure 36. The architecture of X-Vector speaker embedding encoder]

5.1.2 台語語音辨認器 (Taiwanese Speech Recogntizer)

以端對端 ASR 架構(Dong *et al.*, 2018)，如圖 37 所示，以台語音檔和對應的文本，和上述已經訓練好的 X-Vectors，進行 Transformer-based ASR model 的訓練(shirayu, 2021)。在資料集分配上使用跟上述語者向量編碼器一樣 194 人的 train set，並將相同 10 人的 test set 分出 5 人給 dev set。

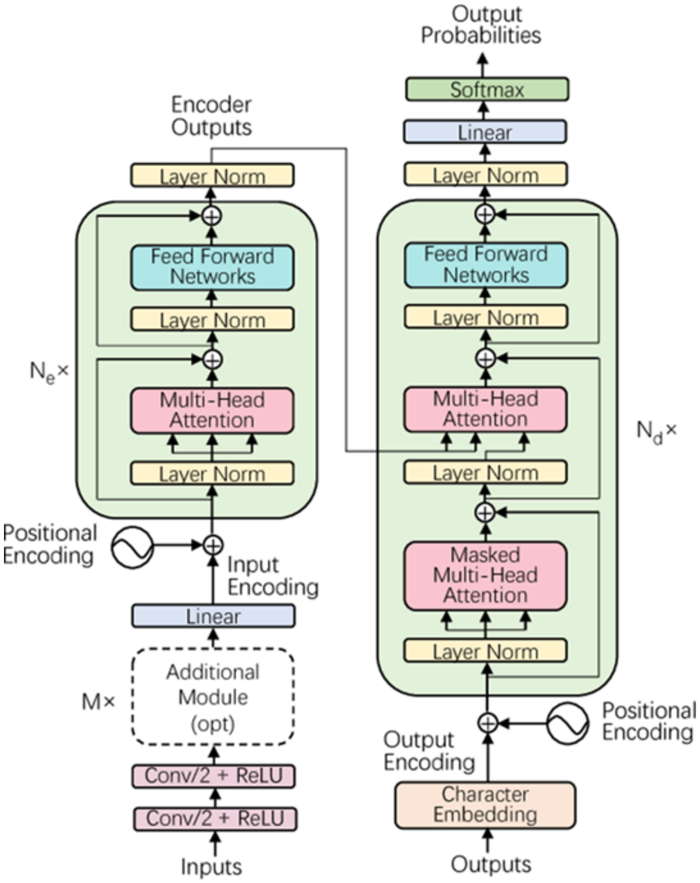


圖 37. E2E-ASR 架構(Dong et al., 2018)
[Figure 37. The architecture of Taiwanese speech recognizer]

5.1.3 多語者語音合成器 (Multi-speaker TTS)

多語者語音合成器的部分採用類似如圖 38 的架構(Chen *et al.*, 2020)，以台語音檔和對應的文本，和上述已經訓練好的 X-Vectors，進行 Multi-speaker Transformer-TTS model 的訓練(shirayu, 2021)。資料集分配上則跟上述 ASR 完全一致，train：dev：test＝194：5：5。

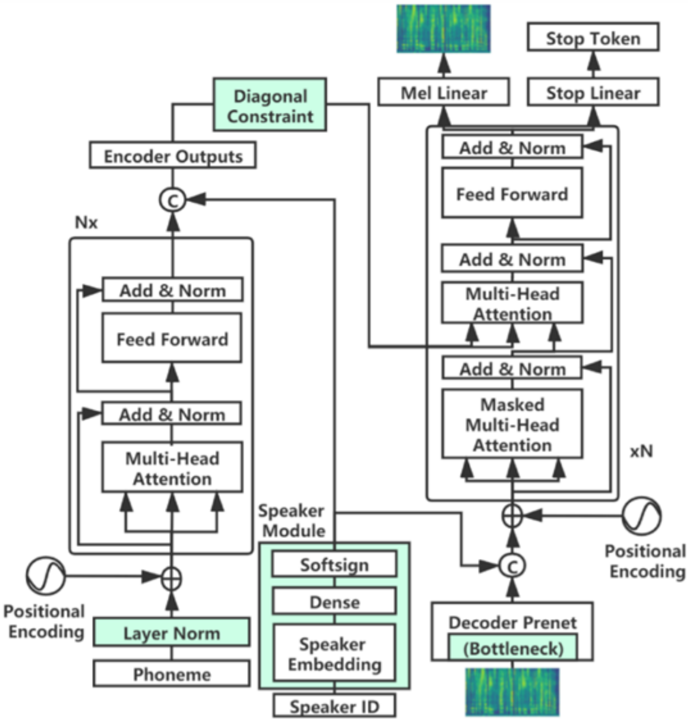


圖 38. E2E-TTS 架構(Chen et al., 2020)
[Figure 38. The architecture of multi-speaker Taiwanese speech synthesis]

5.2 跨語言之華語對台語語音轉換系統建置 (Cross-Lingual Voice Conversion)

在 VCC 2020 中，比賽又分為 Task1：同語言任務以及 Task2：跨語言任務，如圖 39 所示。而在跨語言任務中，比賽中設置目標語者的語言分別為華語，德語和法語，並需要利用語音轉換使用目標語者的語音特徵說出英文句子。

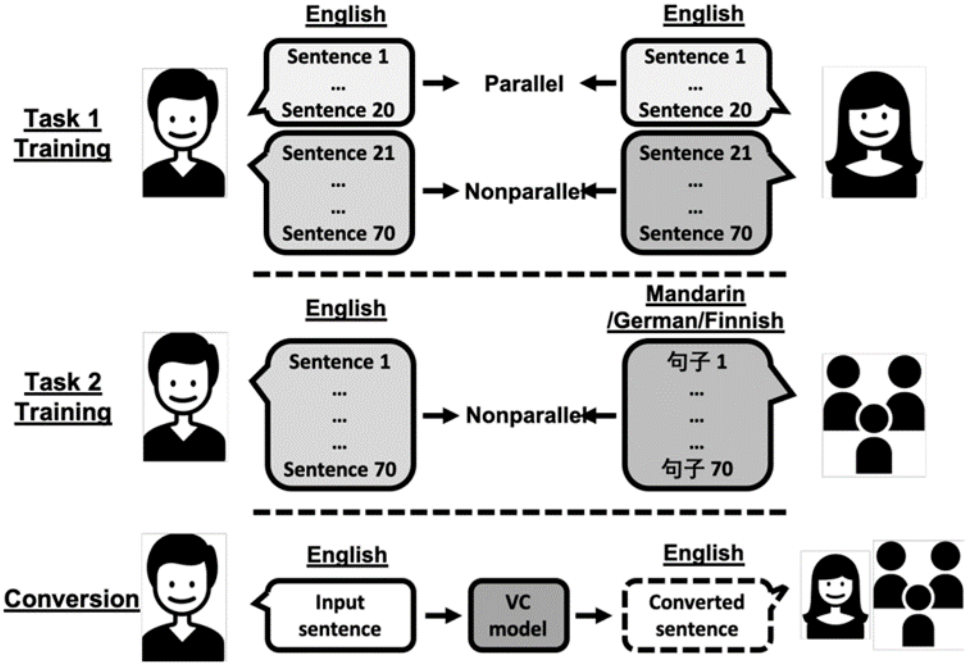


圖 39. VCC 2020 兩種任務介紹(Zhao et al., 2020)
[Figure 39. The Intra- and cross-lingual voice conversion tasks in VCC 2020]

Cascade ASR and TTS 方法因為架構的特性，也可以做到跨語言的語音轉換，我們將目標語者的語言設定為華語，目標建置一個華語對台語語音轉換系統。跟同語言的台語對台語語音轉換系統相比，跨語言最大的差異在於 TTS 端的部分，如圖 40 所示。目標語者的語言變成華語後，要做到跨語言的語音轉換就必須重新以華語和台語 2 種語言去訓練雙語言的多語者語音合成器。

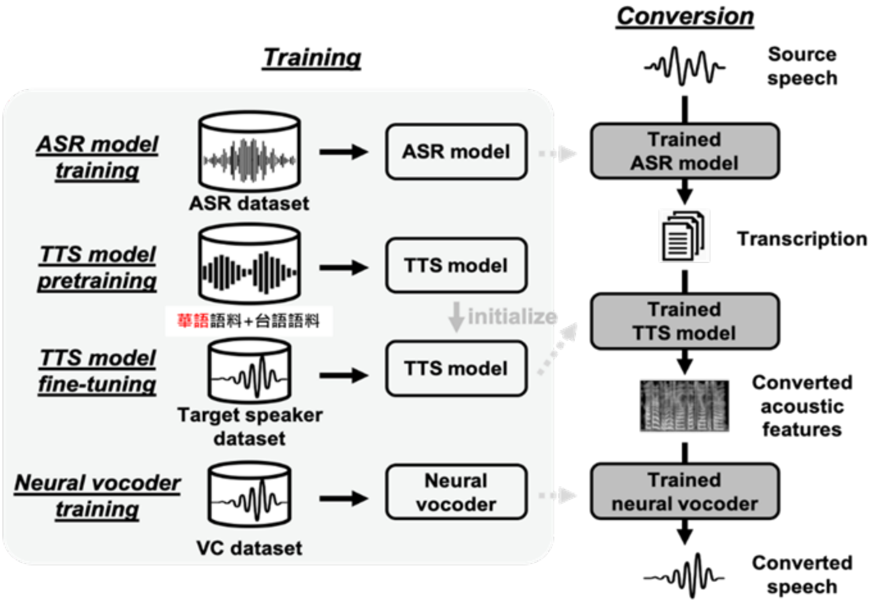


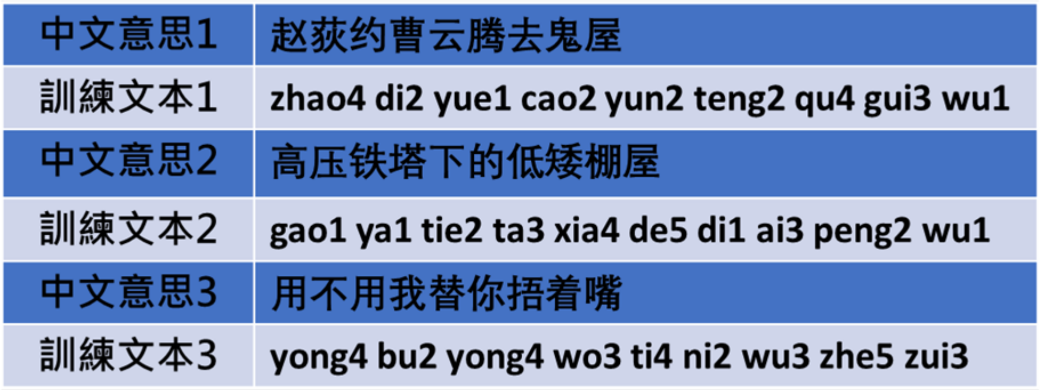
圖 40. 結合語音辨認及合成模組之華語轉台語語音轉換系統架構 (Snyder et al., 2018)

[Figure 40. The architecture of the cross-lingual voice conversion framework]

我們在同語言的部分，也使用了原先由 librispeech 英文語料所訓練的 X-Vectors，並以英文的 X-Vectors 重新對台語的多語者語音合成器進行訓練後，發現語音轉換的音檔聽感其實差異不大，因此在跨語言的部分，我們在 X-Vectors 的部分沿用了原先英文語料訓練的模型。而 ASR 的部分沿用同語言任務已經訓練好的模型。聲碼器也依舊使用原本訓練好的 PWG。

此處應用在跨語言語音轉換的華語語料，為 VCC 2020 中，Cascade ASR and TTS 方法在華語對英語的跨語言語音轉換中，訓練雙語言多語者語音合成器(Kamo, 2021)時使用的語料。名稱為 csmsc，共有 10000 筆音檔，語料總長度約 11.86 小時，為一大陸腔女生所錄製的華語語料。

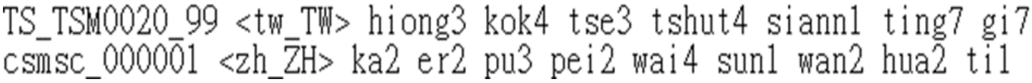
為了使華語文本和台語文本一致，我們使用華語語料中跟台語文本一樣使用子音加母音的文本形式作為訓練文本，部分範例如圖 41 所示。



中文意思1	赵荻约曹云腾去鬼屋
訓練文本1	zhao4 di2 yue1 cao2 yun2 teng2 qu4 gui3 wu1
中文意思2	高压铁塔下的低矮棚屋
訓練文本2	gao1 ya1 tie2 ta3 xia4 de5 di1 ai3 peng2 wu1
中文意思3	用不用我替你捂着嘴
訓練文本3	yong4 bu2 yong4 wo3 ti4 ni2 wu3 zhe5 zui3

圖 41. 華語語料訓練文本部分範例
[Figure 41. A typical example of the Mandarin speech transcription data]

因為使用了兩種語言作為訓練文本，為了區分文本的語言屬性，需要在訓練文本前加上語言碼，如英文預設為<en_US>，華文預設為<zh_ZH>，我們在台語方面則訂為<tw_TW>，加上語言碼後的訓練文本部分範例如圖 42 所示。



```
TS_TSM0020_99 <tw_TW> hiong3 kok4 tse3 tshut4 siann1 ting7 gi7
csmsc_000001 <zh_ZH> ka2 er2 pu3 pei2 wai4 sun1 wan2 hua2 ti1
```

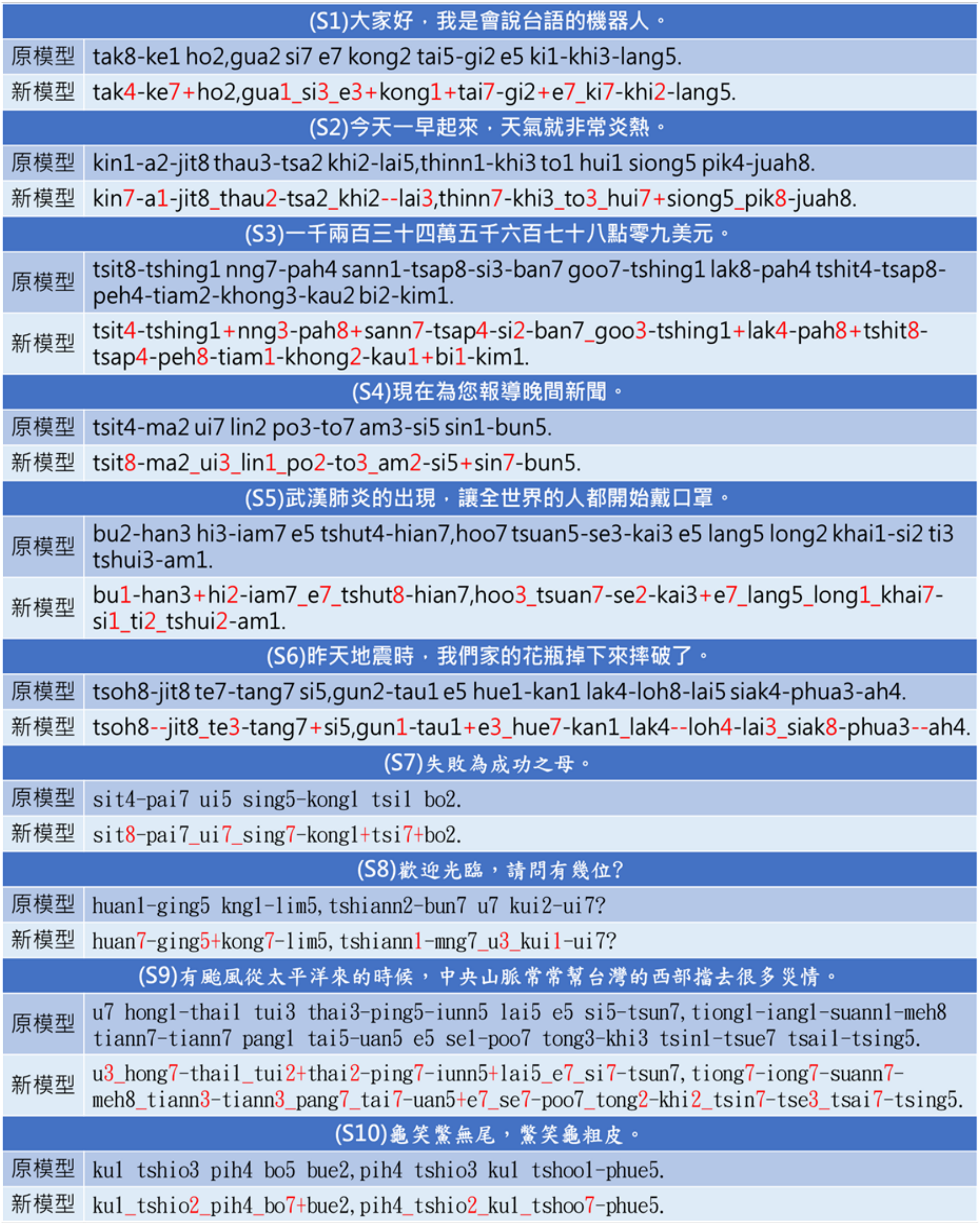
圖 42. 跨語言任務語言碼部分範例
[Figure 42. A typical example of the language code embedding for cross-lingual voice conversion task]

在使用 csmsc 華語語料以及同語言任務使用的台語語料，以及原先英文語料訓練的 X-Vectors，混合訓練好雙語言多語者語音合成器(Kamo, 2021)後，我們也成功建置了一個跨語言之華語對台語語音轉換系統。

6. 實驗 (Experimental Results)

6.1 單人台語語音合成系統實驗 (Single-Speaker Taiwanese Speech Synthesis)

我們使用變調更正以及增加韻律符號的新文本，重新訓練台語語音合成的新模型，並以原本使用沒有變調以及沒有考慮兩種新的韻律符號的舊文本訓練的原模型做比較，簡單設計了以下實驗。分別準備 10 句中文句子原始還沒校正過的台羅拼音，作為原模型的輸入文本合成語音，然後請校正人員以相同規則為這 10 句台羅拼音進行變調更正以及韻律符號的添加，作為新模型輸入的文本並合成語音。10 句實驗句子（S1~S10）校正前後的台羅拼音比較如圖 43 所示，標記紅色的地方為變調更正和韻律符號不同的地方。



(S1)大家好，我是會說台語的機器人。	
原模型	tak8-ke1 ho2,gua2 si7 e7 kong2 tai5-gi2 e5 ki1-khi3-lang5.
新模型	tak4-ke7+ho2,gua1_si3_e3+kong1+tai7-gi2+e7_ki7-khi2-lang5.
(S2)今天一早起來，天氣就非常炎熱。	
原模型	kin1-a2-jit8 thau3-tsa2 khi2-lai5,thinn1-khi3 to1 hui1 siong5 pik4-juah8.
新模型	kin7-a1-jit8_thau2-tsa2_khi2--lai3,thinn7-khi3_to3_hui7+siong5_pik8-juah8.
(S3)一千兩百三十四萬五千六百七十八點零九美元。	
原模型	tsit8-tshing1 nng7-pah4 sann1-tsap8-si3-ban7 goo7-tshing1 lak8-pah4 tshit4-tsap8-peh4-tiam2-khong3-kau2 bi2-kim1.
新模型	tsit4-tshing1+nng3-pah8+sann7-tsap4-si2-ban7_goo3-tshing1+lak4-pah8+tshit8-tsap4-peh8-tiam1-khong2-kau1+bi1-kim1.
(S4)現在為您報導晚間新聞。	
原模型	tsit4-ma2 ui7 lin2 po3-to7 am3-si5 sin1-bun5.
新模型	tsit8-ma2_ui3_lin1_po2-to3_am2-si5+sin7-bun5.
(S5)武漢肺炎的出現，讓全世界的人都開始戴口罩。	
原模型	bu2-han3 hi3-iam7 e5 tshut4-hian7,hoo7 tsuan5-se3-kai3 e5 lang5 long2 khai1-si2 ti3 tshui3-am1.
新模型	bu1-han3+hi2-iam7_e7_tshut8-hian7,hoo3_tsuan7-se2-kai3+e7_lang5_long1_khai7-si1_ti2_tshui2-am1.
(S6)昨天地震時，我們家的花瓶掉下來摔破了。	
原模型	tsoh8-jit8 te7-tang7 si5,gun2-tau1 e5 hue1-kan1 lak4-loh8-lai5 siak4-phua3-ah4.
新模型	tsoh8--jit8_te3-tang7+si5,gun1-tau1+e3_hue7-kan1_lak4--loh4-lai3_siak8-phua3--ah4.
(S7)失敗為成功之母。	
原模型	sit4-pai7 ui5 sing5-kong1 tsi1 bo2.
新模型	sit8-pai7_ui7_sing7-kong1+tsi7+bo2.
(S8)歡迎光臨，請問有幾位?	
原模型	huan1-ging5 kng1-lim5, tshiann2-bun7 u7 kui2-ui7?
新模型	huan7-ging5+kong7-lim5, tshiann1-mng7_u3_kui1-ui7?
(S9)有颱風從太平洋來的時候，中央山脈常常幫台灣的西部擋去很多災情。	
原模型	u7 hong1-thai1 tui3 thai3-ping5-iunn5 lai5 e5 si5-tsun7, tiong1-iang1-suann1-meh8 tiann7-tiann7 pang1 tai5-uan5 e5 se1-poo7 tong3-khi3 tsin1-tsue7 tsai1-tsing5.
新模型	u3_hong7-thai1_tui2+thai2-ping7-iunn5+lai5_e7_si7-tsun7, tiong7-iong7-suann7-meh8_tiann3-tiann3_pang7_tai7-uan5+e7_se7-poo7_tong2-khi2_tsin7-tse3_tsai7-tsing5.
(S10)龜笑鱉無尾，鱉笑龜粗皮。	
原模型	ku1 tshio3 pih4 bo5 bue2, pih4 tshio3 ku1 tshoo1-phue5.
新模型	ku1_tshio2_pih4_bo7+bue2, pih4_tshio2_ku1_tshoo7-phue5.

圖 43. 校正前後的台羅拼音比較

[Figure 43. Comparison of Taiwanese transcriptions before and after tone sandi annotation]

我們請聽者對原模型以及新模型各 10 個句子合成出的音檔進行自然度的評分，最後收集到 27 位聽者的評分，原模型以及新模型的實驗結果如圖 44 及圖 45 所示。

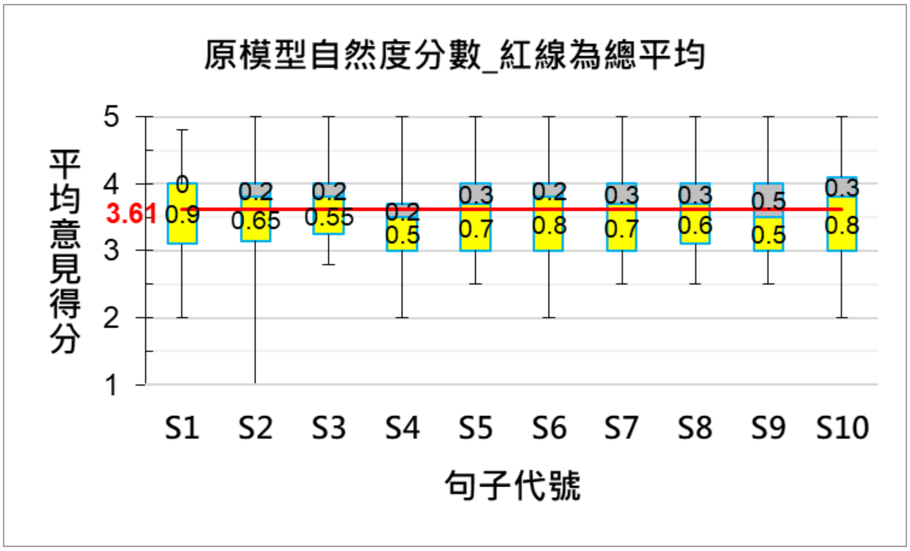


圖 44. 原模型自然度分數實驗結果盒鬚圖
[Figure 44. The box-and-whisker plot of naturalness scores of the baseline model]

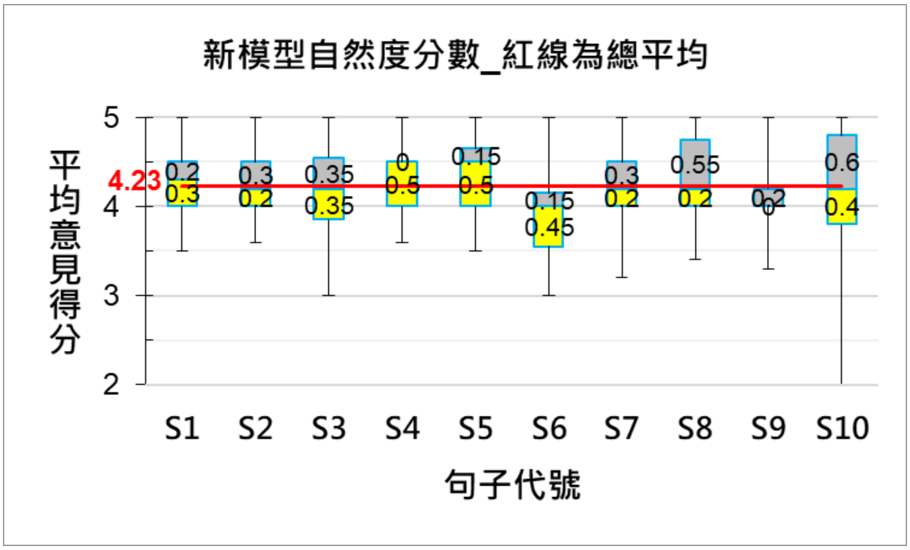


圖 45. 新模型自然度分數實驗結果盒鬚圖
[Figure 45. The box-and-whisker plot of naturalness scores of the improved model]

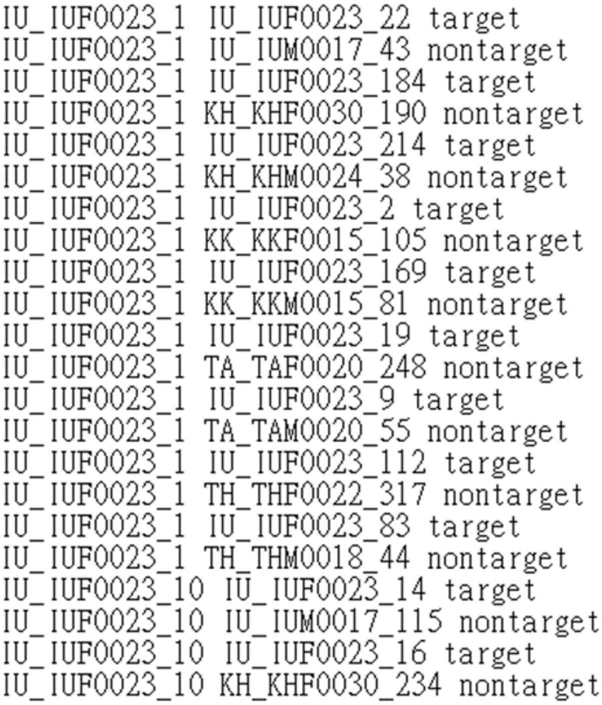
由實驗結果得知，輸入同樣韻律規則的文本合成的語音，比起只有以連字號和空格作為韻律符號的文本合成的語音，的確在聽感上有更加接近真人在講同一句話時該有的順暢度，較少會在奇怪的地方停頓。因此下一個目標，即為使用此新語料訓練能產出同樣韻律規則的中文轉台語機器翻譯，以完成整個新台語語音合成系統的建置。

6.2 結合語音辨認及合成模組之台語多語者語音轉換系統實驗 (Multi-Speaker Voice Conversion)

以下將介紹使用 TAT-ASR 以及 TAT-TTS 台文語料庫總共 204 名語者，共約 145 小時的台語語料訓練的模型成果，以及結合語音辨認及合成模組之台語語音轉換系統的相關實驗。

6.2.1 語者向量編碼器EER結果 (Performance on Speaker Recognition)

將 204 人的台語音檔以 train set 194 人，test set 10 人的設定進行語者向量編碼器的訓練 (esdeboer, 2020)。並以 test set 的台語語料製作 EER 的測試檔案，將 10 位測試語者取出 1 位，以相同語者，非相同語者的設計平均的跟另外 9 位測試語者做語者辨識的測試，並且以此方法將 10 位測試語者全部測試完畢，測試檔案部分範例如圖 46 所示。最後用 194 人台語語料訓練出的語者向量編碼器，得出了 EER 為 5.506%的測試結果，minDCF 在 p-target = 0.01 的情況下為 0.7101，在 p-target = 0.001 的情況下為 0.8318。



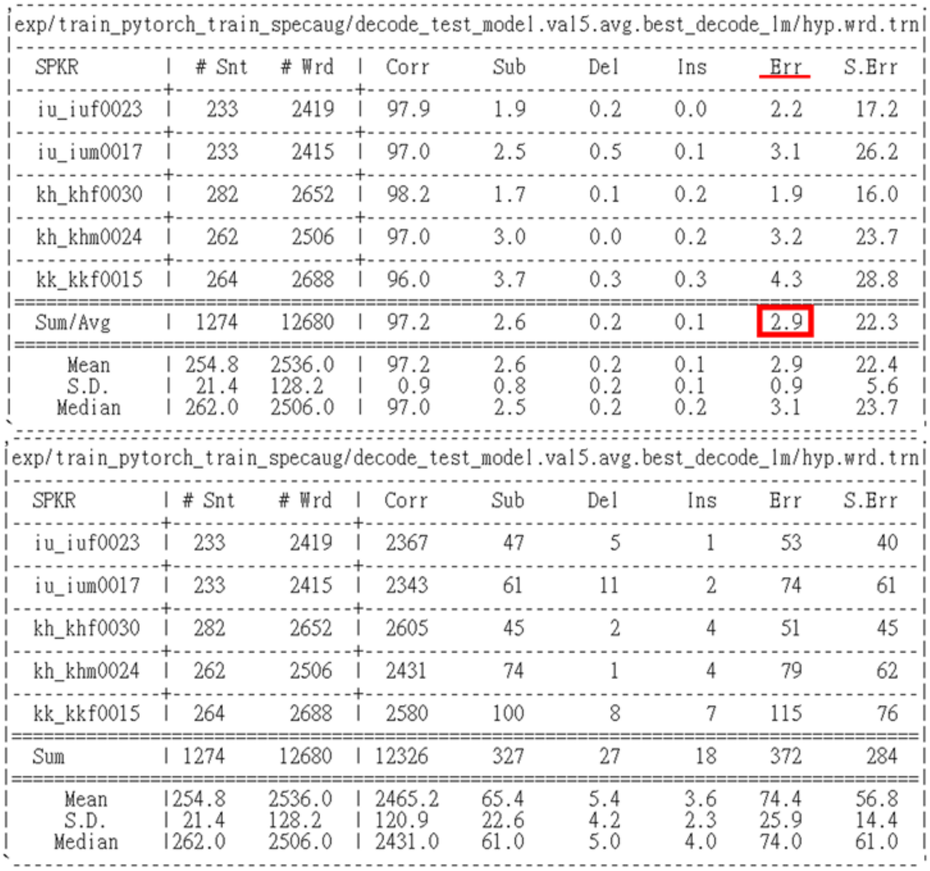
```
IU_IUF0023_1  IU_IUF0023_22 target
IU_IUF0023_1  IU_IUM0017_43 nontarget
IU_IUF0023_1  IU_IUF0023_184 target
IU_IUF0023_1 KH_KHF0030_190 nontarget
IU_IUF0023_1  IU_IUF0023_214 target
IU_IUF0023_1 KH_KHM0024_38 nontarget
IU_IUF0023_1  IU_IUF0023_2 target
IU_IUF0023_1 KK_KKF0015_105 nontarget
IU_IUF0023_1  IU_IUF0023_169 target
IU_IUF0023_1 KK_KKM0015_81 nontarget
IU_IUF0023_1  IU_IUF0023_19 target
IU_IUF0023_1 TA_TAF0020_248 nontarget
IU_IUF0023_1  IU_IUF0023_9 target
IU_IUF0023_1 TA_TAM0020_55 nontarget
IU_IUF0023_1  IU_IUF0023_112 target
IU_IUF0023_1 TH_THF0022_317 nontarget
IU_IUF0023_1  IU_IUF0023_83 target
IU_IUF0023_1 TH_THM0018_44 nontarget
IU_IUF0023_10 IU_IUF0023_14 target
IU_IUF0023_10 IU_IUM0017_115 nontarget
IU_IUF0023_10 IU_IUF0023_16 target
IU_IUF0023_10 KH_KHF0030_234 nontarget
```

圖 46. 語者向量編碼器測試檔案部分範例

[Figure 46. A typical example of the speaker tranacriptions for speaker recognition]

6.2.2 台語語音辨認器錯誤率 (Performance on Taiwanese Speech Recognition)

在資料集分配上使用跟上述語者向量編碼器一樣 194 人的 train set，並將相同 10 人的 test set 分出 5 人給 dev set，最後訓練出來的語音辨認器，錯誤率約為 2.9%，詳情如圖 47 所示。訓練過程相關 loss 資訊如圖 48 所示。



```
,-------------------------------------------------------------------------------------.
|exp/train_pytorch_train_specaug/decode_test_model.val5.avg.best_decode_lm/hyp.wrd.trn|
|-------------------------------------------------------------------------------------|
| SPKR          | # Snt    # Wrd  | Corr      Sub      Del      Ins      Err     S.Err |
|---------------+-----------------+---------------------------------------------------|
| iu_iuf0023    |  233      2419  | 97.9      1.9      0.2      0.0      2.2      17.2 |
|---------------+-----------------+---------------------------------------------------|
| iu_ium0017    |  233      2415  | 97.0      2.5      0.5      0.1      3.1      26.2 |
|---------------+-----------------+---------------------------------------------------|
| kh_khf0030    |  282      2652  | 98.2      1.7      0.1      0.2      1.9      16.0 |
|---------------+-----------------+---------------------------------------------------|
| kh_khm0024    |  262      2506  | 97.0      3.0      0.0      0.2      3.2      23.7 |
|---------------+-----------------+---------------------------------------------------|
| kk_kkf0015    |  264      2688  | 96.0      3.7      0.3      0.3      4.3      28.8 |
|=====================================================================================|
| Sum/Avg       | 1274     12680  | 97.2      2.6      0.2      0.1      2.9      22.3 |
|=====================================================================================|
|     Mean      | 254.8   2536.0  | 97.2      2.6      0.2      0.1      2.9      22.4 |
|     S.D.      |  21.4    128.2  |  0.9      0.8      0.2      0.1      0.9       5.6 |
|    Median     | 262.0   2506.0  | 97.0      2.5      0.2      0.2      3.1      23.7 |
`-------------------------------------------------------------------------------------'
,-------------------------------------------------------------------------------------.
|exp/train_pytorch_train_specaug/decode_test_model.val5.avg.best_decode_lm/hyp.wrd.trn|
|-------------------------------------------------------------------------------------|
| SPKR          | # Snt    # Wrd  | Corr      Sub      Del       Ins      Err    S.Err |
|---------------+-----------------+---------------------------------------------------|
| iu_iuf0023    | 233       2419  | 2367       47        5         1       53       40 |
|---------------+-----------------+---------------------------------------------------|
| iu_ium0017    | 233       2415  | 2343       61       11         2       74       61 |
|---------------+-----------------+---------------------------------------------------|
| kh_khf0030    | 282       2652  | 2605       45        2         4       51       45 |
|---------------+-----------------+---------------------------------------------------|
| kh_khm0024    | 262       2506  | 2431       74        1         4       79       62 |
|---------------+-----------------+---------------------------------------------------|
| kk_kkf0015    | 264       2688  | 2580      100        8         7      115       76 |
|=====================================================================================|
| Sum           | 1274     12680  | 12326     327       27        18      372      284 |
|=====================================================================================|
|     Mean      |254.8    2536.0  | 2465.2   65.4      5.4       3.6     74.4     56.8 |
|     S.D.      | 21.4     128.2  | 120.9    22.6      4.2       2.3     25.9     14.4 |
|    Median     |262.0    2506.0  | 2431.0   61.0      5.0       4.0     74.0     61.0 |
`-------------------------------------------------------------------------------------'
```

圖 47.台語語音辨認器訓練結果

[Figure 47. Experimental results of the Taiwanese speech recognitizer]

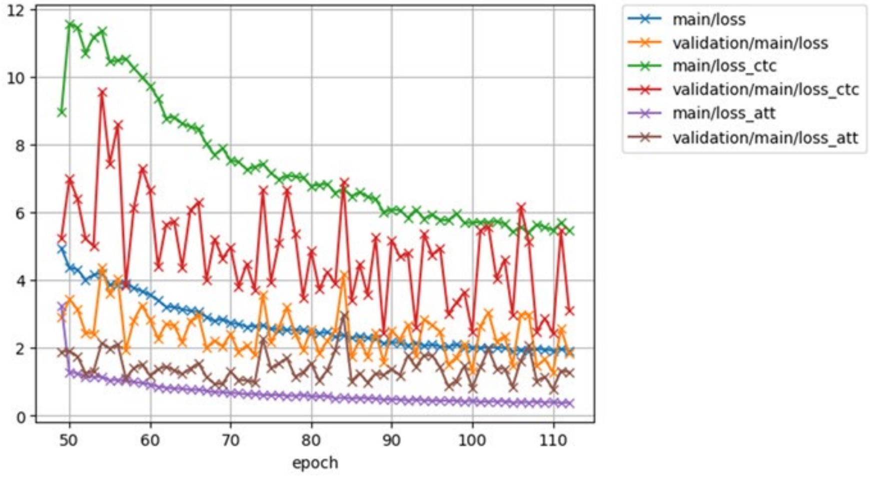


圖 48.台語語音辨認器訓練過程 loss

[Figure 48. The learning curves of the Taiwanese speech synthesis system]

6.2.3 台語語音轉換系統實驗 (Taiwanese Voice Conversion)

最後，我們針對使用台語語料做出的結合語音辨認及合成模組之台語語音轉換系統，設計了轉換音檔「自然度」和「相似度」的主觀評測實驗。此實驗最終收集了 29 位評分者的評分結果，評分人分別有來自中華電信研究院的語音相關專業人士，長問科技的語音相關專業人士，台語老師以及同實驗室的研究生。

6.2.3.1 實驗方法 (Experimental Settings)

實驗問卷分成(1)台語對台語語音轉換和(2)華語對台語語音轉換兩部分，第一部分有 4 位目標語者的合成音檔，第二部分有 3 位目標語者的合成音檔。每 1 位目標語者有數個需要進行評分的合成音檔，和 1 個原始音檔以供對照。

(1) 台語對台語語音轉換的部分，12 個合成音檔分別為

(1-1)4 個使用約 10 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

(1-2)4 個使用約 3 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

(1-3)4 個使用約 30 秒 fine-tuning 語料量做出的語音轉換合成音檔

(2) 華語對台語語音轉換的部分，8 個合成音檔分別為

(2-1)4 個使用約 6 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

(2-2)4 個使用約 3 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

評分者聽完合成音檔後，依據主觀感受對每個合成音檔評兩種分數。

(一) 自然度分數

根據聽到的「自然度」進行 1.0 到 5.0 的評分，最多評分到小數第一位

最低分 1.0 分 為完全不像真人講話的聲音

最高分 5.0 分 為完全像是真人講話的聲音

(二)相似度分數

根據聽到的「相似度」進行 1.0 到 5.0 的評分，最多評分到小數第一位

最低分 1.0 分 <原始音檔>和<合成音檔>完全不像同一個人講話的聲音

最高分 5.0 分 <原始音檔>和<合成音檔>完全像同一個人講話的聲音

6.2.3.2 同語言任務實驗結果 (Intra-Lingual Voice Conversion)

台語對台語語音轉換音檔的 MOS 分數盒鬚圖（Box-Plot）

(1-1)10 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 49 和圖 50 所示

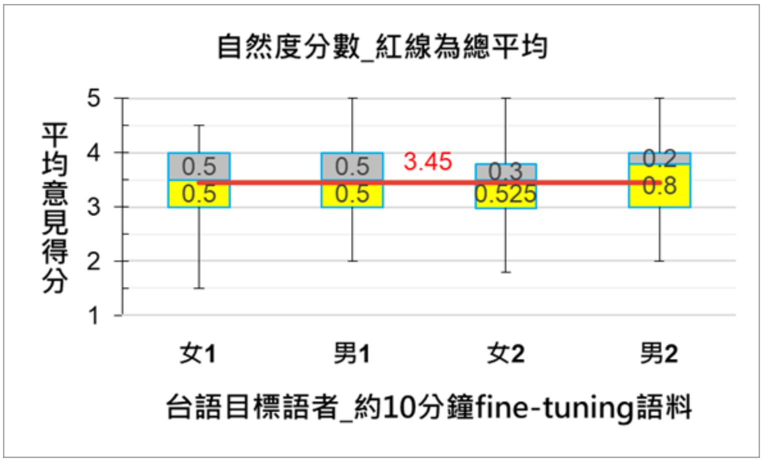


圖 49. 同語言任務使用 10 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 49. The box-and-whisker plot of naturalness scores using 10-minute fine-tuning data for intra-lingual voice conversion]

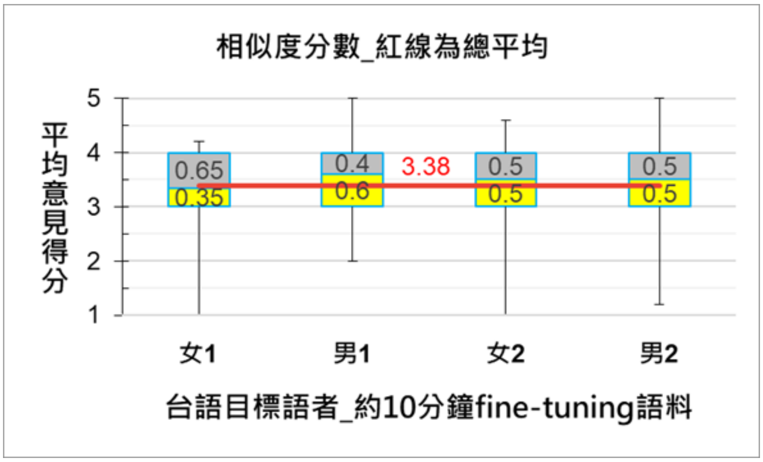


圖 50. 同語言任務使用 10 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 50. The box-and-whisker plot of similarity scores using 10-minute fine-tuning data for intra-lingual voice conversion]

(1-2)3 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 51 和圖 52 所示

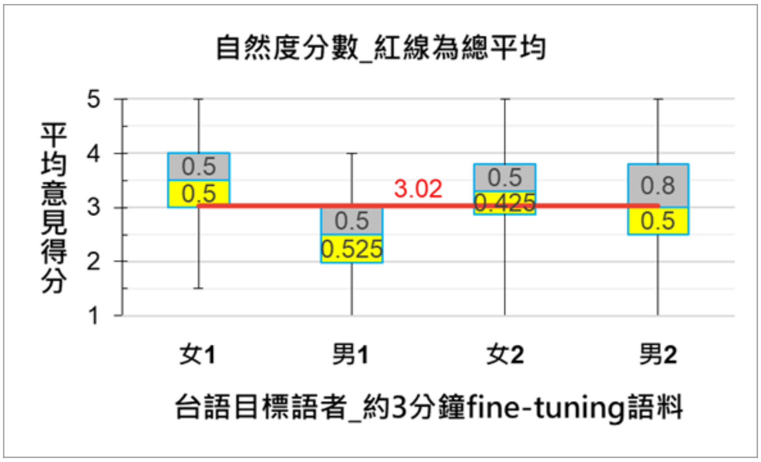


圖 51. 同語言任務使用 3 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 51. The box-and-whisker plot of naturalness scores using 3-minute fine-tuning data for intra-lingual voice conversion]

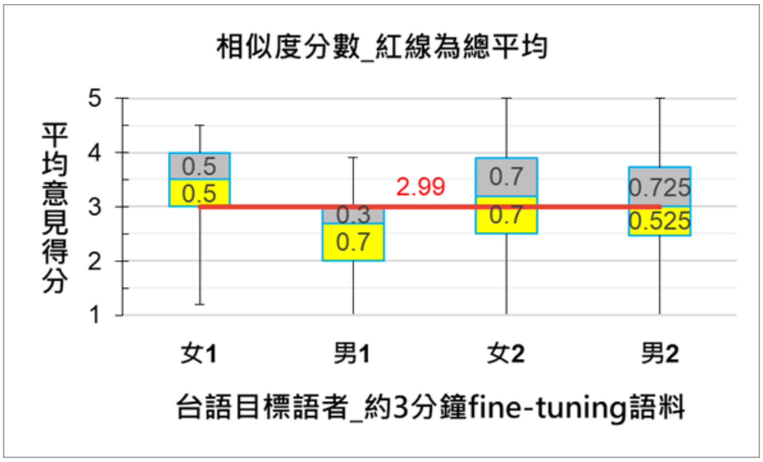


圖 52. 同語言任務使用 3 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 52. The box-and-whisker plot of similarity scores using 3-minute fine-tuning data for intra-lingual voice conversion]

(1-3)30 秒 fine-tuning 語料量，自然度分數和相似度分數如圖 53 和圖 54 所示

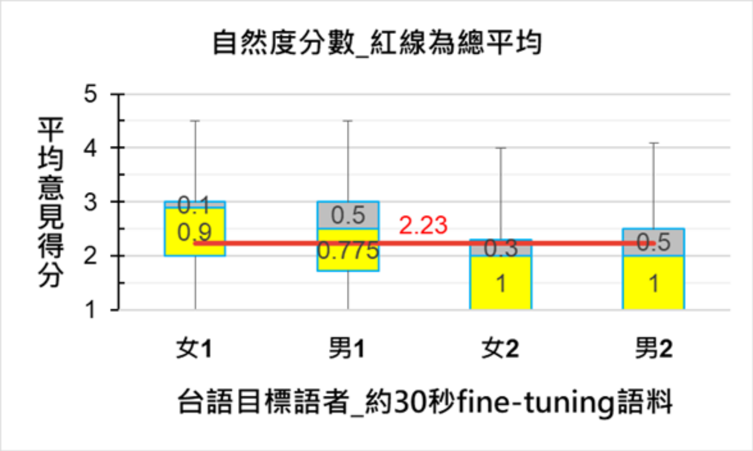


圖 53. 同語言任務使用 30 秒 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 53. The box-and-whisker plot of naturalness scores using 30-second fine-tuning data for intra-lingual voice conversion]

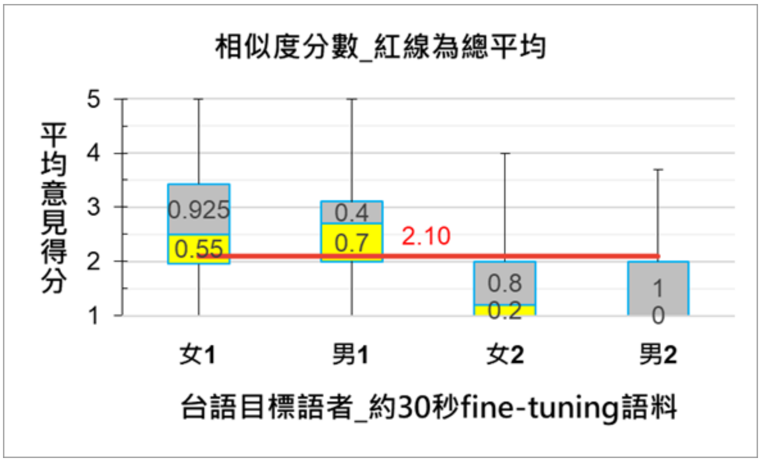


圖 54. 同語言任務使用 30 秒 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 54. The box-and-whisker plot of similarity scores using a 30-second fine-tuning data for intra-lingual voice conversion]

6.2.3.3 跨語言任務實驗結果 (Cross-Lingual Voice Conversion)

華語對台語語音轉換音檔的 MOS 分數盒鬚圖（Box-Plot）

(2-1)6 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 55 和圖 56 所示

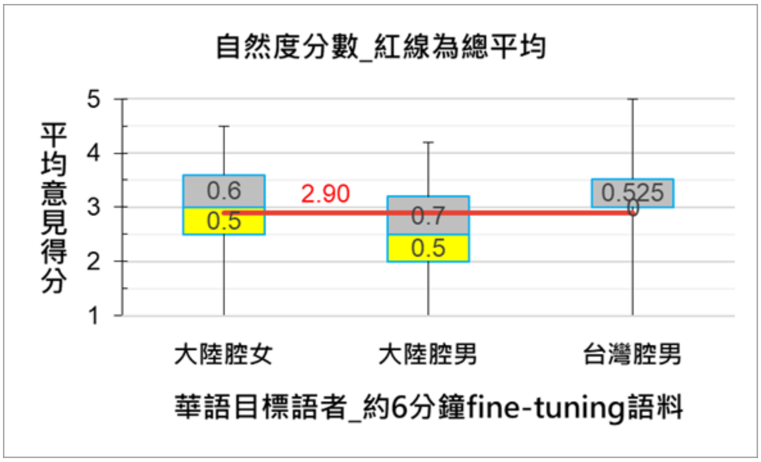


圖 55. 跨語言任務使用 6 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 55. The box-and-whisker plot of naturalness scores using 6-minute fine-tuning data for cross-lingual voice conversion]

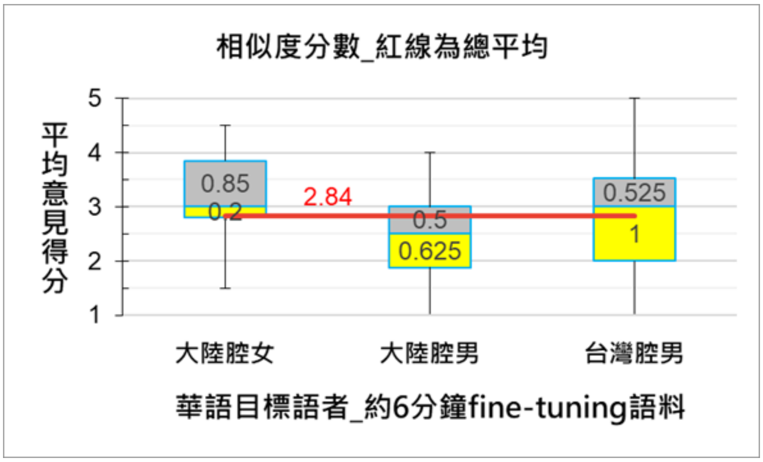


圖 56. 跨語言任務使用 6 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 56. The box-and-whisker plot of similarity scores using 6-minute fine-tuning data for cross-lingual voice conversion]

(2-2)3 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 57 和圖 58 所示

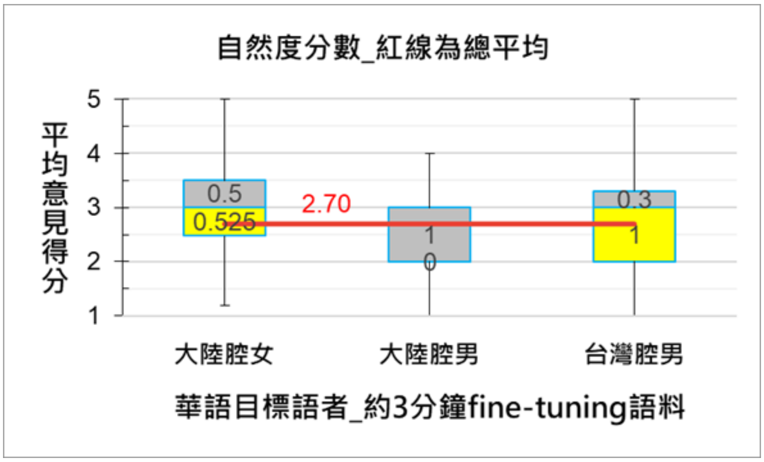


圖 57. 跨語言任務使用 3 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 57. The box-and-whisker plot of naturalness scores using 3-minute fine-tuning data for cross-lingual voice conversion]

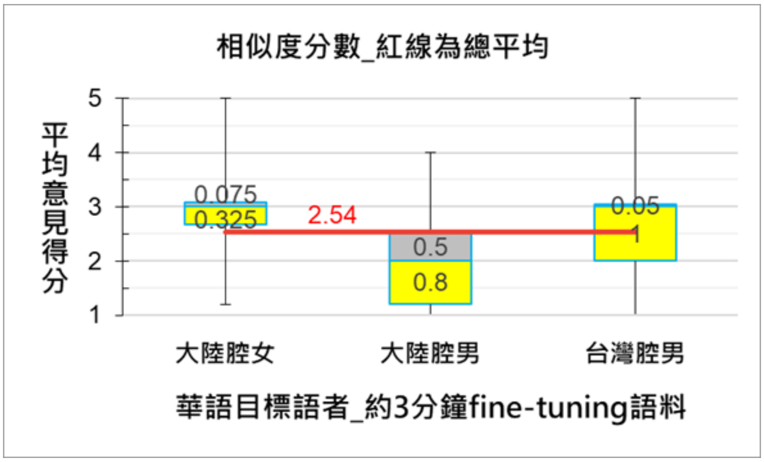


圖 58. 跨語言任務使用 3 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 58. The box-and-whisker plot of similarity scores using 3-minute fine-tuning data for cross-lingual voice conversion]

從實驗結果可以得知，無論是同語言還是跨語言任務，使用的 fine-tuning 語料量越少，音檔越難達到高品質的自然度與相似度，且跨語言的情況又比同語言更艱難。

7. 結論 (Conclusions)

在此論文中，我們利用所蒐集的 Taiwanese Across Taiwan (TAT) 大規模台文語音語料庫，包括，TAT-Vol1~2、TAT-TTS-M1~2 與 TAT-TTS-F1~2，完成了中文文字轉台語語音合成系統，與台語語音轉換系統（包括同語言（台語對台語）與跨語言（華語對台語）兩項語音轉換任務）。

其中的中文文字轉台語語音合成系統，在經利用校正台語變調以及新增兩種韻律符號，訓練出的新模型後，由實驗的結果也得知，合成音檔的自然度，可提升到 4.23 分，如圖 59 所示。

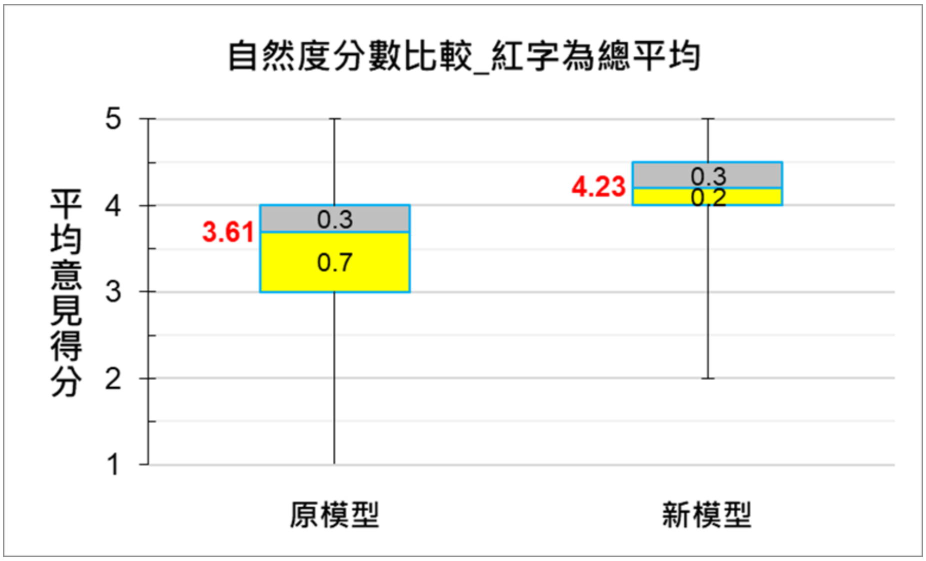


圖 59. 單人台語語音合成系統校正前後自然度實驗結果
[Figure 59. Experimental results on the naturalness of the single-speaker Taiwanese synthesis with and without tone snadi and prosodic boundary annotations]

在結合語音辨認及合成模組之台語多語者語音轉換實驗部分，同語言（台語對台語）語音轉換任務在語料量較充足，如 10 分鐘的情況下，已經可以達到音檔自然度與相似度都 3.45 與 3.38 分，如圖 60 和圖 61 所示。

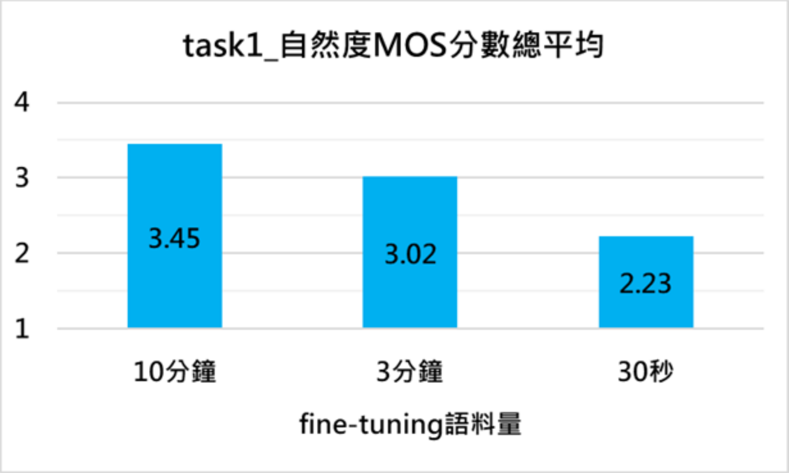


圖 60. 台語轉台語語音轉換系統自然度實驗結果
[Figure 60. Experimental results on the naturalness of the intra-lingual voice conversion]

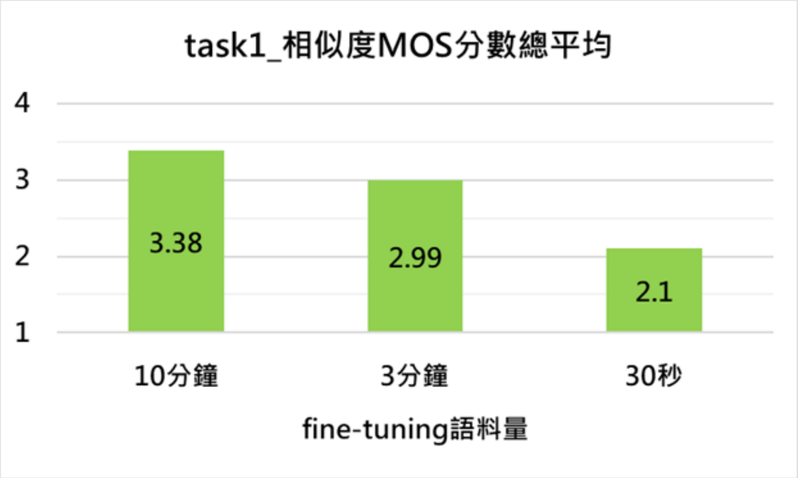


圖 61. 台語轉台語語音轉換系統相似度實驗結果
[Figure 61. Experimental results on the similarity of the intra-lingal voice conversion]

而在跨語言（華語對台語）語音轉換任務難度較高，但在只使用 6 分鐘語料量的情況下，自然度以及相似度的，也還可以達到 2.9 分跟 2.84 分，如圖 62 和圖 63 所示。

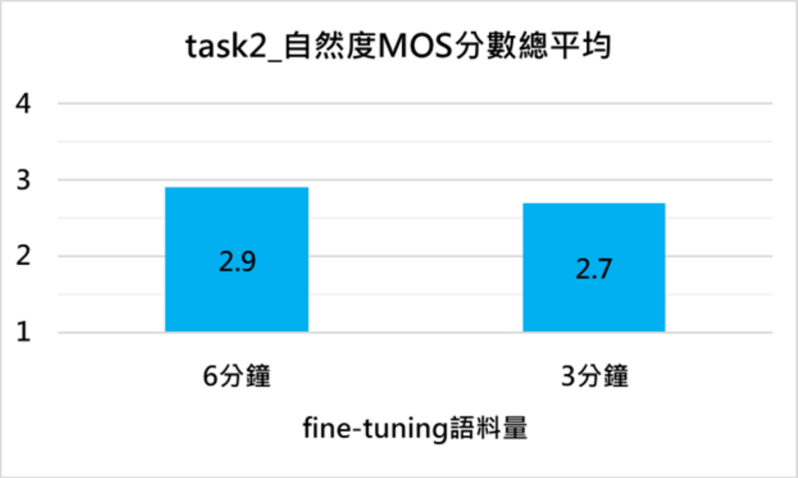


圖 62. 華語轉台語語音轉換系統自然度實驗結果
[Figure 62. Experimental results on the naturalness of the cross-lingual voice conversion]

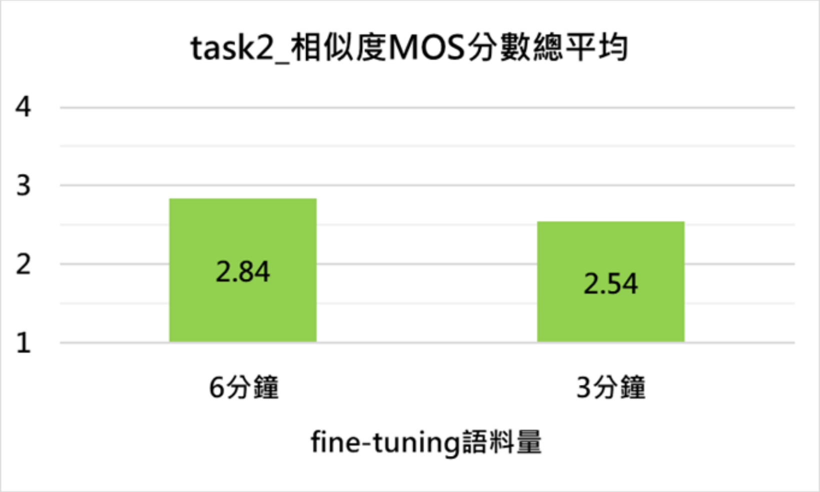


圖 63. 華語轉台語語音轉換系統相似度實驗結果
[Figure 63. Experimental results on the similarity of the cross-lingual voice conversion]

因此，由以上結果可知，我們所蒐集的 TAT 大規模台文語音語料庫，的確可以有效做為開發台語語音合成技術的語料庫。所做出來的中文文字轉台語語音合成系統，與台語語音轉換系統有都有還不錯的效能。

致謝 (Acknowledgements)

This work is supported partially by Taiwan’s Ministry of Education under project ”教育部閩

南語語音語料庫建置計劃", partially by Ministry of Science and Technology under contract No. 109-2221-E-027-108, 110-2221-E-027- 082, 110-2622-8-002-018, partially by National Taiwan University (NTU 109-3111-8-002-002), Cathay Life Insurance, Cathay Century Insurance, Cathay United Bank, Cathay Securities Corporation, and Cathay Securities Investment Trust, partially by Chunghwa Telecom. Lab. under contract No. TL-109-D304 and partially by the Department of Industrial Technology of Ministry of Economic Affairs under contract No. 110-EC-17-A-02-S5-008.

參考文獻 (References)

Chen, M., Tan, X., Ren, Y., Xu, J., Sun, H., Zhao, S., Qin, T., & Liu, T.-Y. (2020). MultiSpeech: Multi-Speaker Text to Speech with Transformer. arXiv preprint arXiv: 2006.04664v2

Dong, L., Xu, S., & Xu, B. (2018). Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). https://doi.org/10.1109/ICASSP.2018.8462506

esdeboer. (2020). GitHub-X-Vector,. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

howardhsu. (2021). GitHug-Facebook AI Research Sequence-to-Sequence Toolkit written in Python. Available https://github.com/pytorch/fairseq

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., & Wang, H.-M. (2016). Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder. arXiv preprint arXiv:1610.04019

Huang, W.-C. Huang, Hayashi, T., Watanabe, S., & Toda, T. (2020). The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS. arXiv preprint arXiv:2010.02434

kamo-naoyuki. (2021). GitHub-Mandarin to English Multi-speaker Transformer-TTS model. Available https://github.com/espnet/espnet/tree/master/egs/vcc20/tts1_en_zh

Prenger, R., Valle, R., & Catanzaro, B. (2018). WaveGlow: A Flow-based Generative Network for Speech Synthesis. arXiv preprint arXiv:1811.00002

rafaelvalle. (2020).GitHub-Tacotron 2 - PyTorch implementation with faster-than-realtime inference. Available: https://github.com/NVIDIA/tacotron2

rafaelvalle. (2020). GitHub-A Flow-based Generative Network for Speech Synthesis. Available https://github.com/NVIDIA/waveglow

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv preprint arXiv:1712.05884v2

shirayu. (2021). GitHub-Transformer-based ASR model. Available: https://github.com/espnet/espnet/tree/master/egs/librispeech/asr1 shirayu. (2021). GitHub-Multi-speaker Transformer-TTS model. Available: https://github.com/espnet/espnet/tree/master/egs/libritts/tts1

sih4sing5hong5. (2015). GitHub-iCorpus 臺華平行新聞語料庫語料加漢字. Available https://github.com/Taiwanese-Corpus/icorpus_ka1_han3-ji7

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. In proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). https://doi.org/10.1109/ICASSP.2018.8461375

Sun, L., Li, K., Wang, H., Kang, S., & Meng, M. (2016). Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In proceedings of 2016 IEEE International Conference on Multimedia and Expo (ICME). https://doi.org/10.1109/ICME.2016.7552917

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. arXiv preprint arXiv:1703.10135v2

Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., &Toda, T. (2020). Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. arXiv preprint arXiv:2008.12527

教育部。(2008)。臺灣閩南語羅馬字拼音方案使用手冊。Available: https://ws.moe.edu.tw/001/Upload/FileUpload/3677-15601/Documents/tshiutsheh.pdf [Ministry of Education. (2008). Tâi-oân Bân-lâm-gú Lô-má-jī Pheng-im Hong-àn.]

The individuals listed below are reviewers of this journal during the year of 2022. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Tao-Hsing Chang
Yung-Chun Chang
Miao-Hsia Chang
Jason S. Chang
Chun Chang
Kuan-Yu Chen
Alvin C.-H. Chen
Chung-Chi Chen
Chen-Yu Chiang
Chin-Chin Chiang
Yu-Tai Chien
Siaw-Fong Chung
Hong-Jie Dai
Min-Yuh Day
Janice Fon
Zhao-Ming Gao
Mei-Ching Ho
Shu-Kai Hsieh
Chen-Yu Chester Hsieh
Fu-Hui Hsieh
Hen-Hsen Huang
Yi-Chin Huang
Jeih-Weih Hung
Win Ping Kuo
Yen-Liang Lin
Yi-Fen Liu
Wei-Yun Ma
Shu-chen Ou
Ming-Hsiang Su
I-Ping Wan
Shih-ping Joe Wang
Yu-Fang Wang
Jheng-Long Wu
Jian-Cheng Wu
Jui-chuan Yeh
Jui-Feng Yeh

2022 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 27

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2022

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

SUBJECT INDEX

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan, R.O.C.

Tel.：886-2-2788-1638 Fax：886-2-2651-9386

E-mail: aclclp@aclclp.org.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#：________________

Name：__________________________________ Date of Birth：_______________

Country of Residence：____________________ Province/State：_________________

Passport No.：___________________________ Sex: ____________________________

Education(highest degree obtained)：____

Work Experience：____

____

Present Occupation：____

Address：____

____

Email Add：____

Tel. No：____________________________ Fax No：__________________________

Membership Category：□ Regular Member □ Life Member

Date：_____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member	：	US$ 50.-（NT$ 1,000）
Life Member	：	US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

社團法人中華民國計算語言學學會

宗旨：

（一）從事計算語言學之研究
（二）推行計算語言學之應用與發展
（三）促進國內外中文計算語言學之研究與發展
（四）聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁填妥入會申請表，填妥後E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115022南港區舊莊街一段3巷34號1樓
電話：(02) 2788-1638　　傳真：(02) 2651-9386
E-mail：aclclp@aclclp.org.tw　網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

社團法人中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號	(由本會填寫)		
姓名		性別		出生日期	年 月 日
				身分證號碼	
現職		學歷			
通訊地址	□□□				
戶籍地址	□□□				
電話		E-Mail			
申請人: (簽章) 中華民國 年 月 日					

審查結果:

1. 年費：

終身會員：	10,000.-
個人會員：	1,000.-
學生會員：	500.-（限國內學生）
團體會員：	20,000.-

2. 連絡處：

   地址：台北市115022南港區舊莊街一段3巷34號1樓
   電話：(02) 2788-1638 傳真：(02) 2651-9386
   E-mail：aclclp@aclclp.org.tw 網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :____(Please print) Date: ______________

Please debit my credit card as follows: US$: _________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Banl:________________

Card No.: _________- _________ - _________ - _________ Exp. Date:___________(M/Y)

3-digit code: ____________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : ___________________________________

Address: ____

Tel.: _____________________________ E-mail : _________________________

PAYMENT FOR

US$___________❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: _______

US$___________❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: _______

US$___________❑ Publications:_______________________

US$___________❑ Text Corpora:_________________________

US$___________❑ Speech Corpora:__________________

US$___________❑ Others :______________________

US$___________❑ Membership Fees: ❑ Life Membership ❑ New Membership ❑Renew

US$___________= Total

*** Fax 886-2-2651-9386 or Mail this form to :**

Association for Computational Linguistics and Chinese Language Processing
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan, R.O.C
E-mail: aclclp@aclclp.org.tw **Website: http://www.aclclp.org.tw**

社團法人中華民國計算語言學學

信用卡付款單

姓名: ____________________________(請以正楷書寫)　　日期:：________________

卡別：□ VISA CARD □ MASTER CARD □ JCB CARD 發卡銀行：___________

信用卡號：________-_________-_________-_________　有效日期：__________(m/y)

卡片後三碼：______（卡片背面簽名欄上數字後三碼

持卡人簽名：_________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：___________________________E-mail：______________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________ □ 中文計算語言學期刊(IJCLCLP) _________________

NT$__________ □ Journal of Information Science and Engineering (JISE)

NT$__________ □ 文字語料庫 ___________________________________

NT$__________ □ 語音資料庫 ___________________________________

NT$__________ □ 光華雜誌語料庫1976~2010

NT$__________ □ 中文資訊檢索標竿測試集/文件集

NT$__________ □ 會員年費：□續會　　□新會員　　□終身會員

NT$__________ □ 其他: ____

NT$_______________ ＝ 合計

填妥後請傳真至 02-26519386 或郵寄至:
115022台北市南港區舊莊街一段3巷34號1樓 中華民國計算語言學學會 收
E-mail: aclclp@aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表 (甲)	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or "中華民國計算語言學學會"
- E-mail：aclclp@aclclp.org.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

社團法人中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本) ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05 中文詞類分析	185	205	_____	_____
7.	no.93-06 現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02 古漢語字頻表	180	200	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01 古漢語詞頻表 (甲)	400	450	_____	_____
15.	no.97-02 論語詞頻表	90	100	_____	_____
16	no.98-01 詞頻詞典	395	440	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊 (一年兩期) 年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內（台灣地區）使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-1638

聯絡人：　黃琪 小姐、何婉如 小姐　　E-mail:aclclp@aclclp.org.tw

訂購者：　______________________　收據抬頭：______________________

地　　址：____

電　　話：______________________　E-mail: ______________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.
1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).
Please visit the following websites for details.
(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)
(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Online Submission: https://ijclclp.aclclp.org.tw/servlet/SignInHandler

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

For more information, please email to ijclclp@aclclp.org.tw

International Journal of
Computational Linguistics & Chinese Language Processing
vol.27 no.2 December 2022

Contents

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing