Using Chi-square Testing in Modeling Confusion Characteristics for Robust Phonetic Set Generation

Yeou-Jiunn Chen(1) and Chung-Hsien Wu (2)

(1) Advanced Technology Center, Computer & Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, R.O.C.
(2) Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.

chenyj@itri.org.tw, chwu@csie.ncku.edu.tw

Abstract

A phonetic representation of a language is used to describe the corresponding pronunciation and synthesize the acoustic model of any vocabulary. In order to obtain better phonetic representation, context-dependent units are used to model co-articulation effects between phones and have been broadly in speech recognition. However, this representation generally increases the number of recognition units. A phonetic representation with smaller phonetic units such as SAMPA-C for Mandarin Chinese can be applied to reduce the number of recognition units. Nevertheless, smaller phonetic units such as SAMPA-C will contain confusion characters and generally degrade the recognition performance. In this paper, a statistical method based on chi-square testing is used to investigate the confusion characteristics among phonetic units and develop a more reliable phonetic set, named modified SAMPA-C. Finally, experiments on continuous Mandarin telephone speech recognition were conducted. Experimental results show an encouraging improvement on recognition performance can be obtained. In addition, the proposed approaches represent a good compromise between the demands of accurate acoustic modeling.

1. Introduction

From the viewpoint of speech recognition, a phonetic representation is functionally defined by the mapping of the fundamental phonetic units of a language to describe the corresponding pronunciation and synthesize the acoustic model of any vocabulary. In the past years, context-dependent units have been broadly used to model the co-articulation effects such as triphone models, which consider both left and right phonemes at the same time. However, this representation generally increases the number of recognition units. Approaches for designing a smaller number of phonetic units are needed in the context-dependent based recognition.

In recent years, many phoneme-based phonetic representations have been used such as International Phonetic Alphabet (IPA) [1], Speech Assessment Methods Phonetic Alphabet (SAMPA) [2], and SAMPA for Chinese (SAMPA-C) [3]. Among these representations, SAMPA-C is more flexible and consistent than other phoneme-based phonetic representations for Mandarin Chinese. However, in SAMPA-C, several phonetic units with short duration are not easy to be distinguished and therefore degrade the recognition performance.

For Mandarin speech, the confusion characteristics can be found and analyzed in syllable-dependent, subsyllable-dependent, or phoneme-dependent situation. In a training database, syllable-dependent confusion characteristics are difficult to extract due to the sparse data problem. In contrast, the inconsistent phoneme segment in the training data is also not suitable to detect the phone-dependent confusion characteristics. The misdetected phones will result in misrecognition of syllables/subsyllables. Consequently, the phone-dependent confusion characteristic is not helpful for the analysis and representation of confusion characteristics of SAMPA-C based Mandarin speech recognizer. Therefore, the subsyllable is chosen as a compromising unit for the analysis of subsyllable-dependent confusion characteristics.

In this paper, based on the statistical hypothesis, the χ^2 (chi-square) testing [4] is an alternative test for evaluating dependence, which does not assume normally distributed probabilities. The underlying principle is to compare the observed frequencies with the expected frequencies. For investigating the effects of the confusion characteristics, the χ^2 statistic is used to examine the consistencies of two probabilistic distributions and the statistical decision criteria are applied to evaluate the statistical evidence for the confusion degree of two subsyllables. According to the analysis result, a less confusable phonetic set, namely modified SAMPA-C, is applied to develop a new Mandarin speech recognizer and compared to the original SAMPA-C.

The architecture for constructing the recognition model is shown in Fig. 1 and can be divided into two processes: development process and evaluation process. In the development process, an acoustic training database is collected and classified statistically for establishing SAMPA-C based recognition models. By analyzing the output distributions of confusion models, the confusion characteristics are extracted and used to generate the modified SAMPA-C. Moreover, using decision tree, the context-dependent models are generated for evaluating the performance. In the evaluation process, two continuous Mandarin speech recognition systems are developed and used to evaluate the syllable recognition rates using SAMPA-C and modified SAMPA-C HMM-based recognition models, respectively.

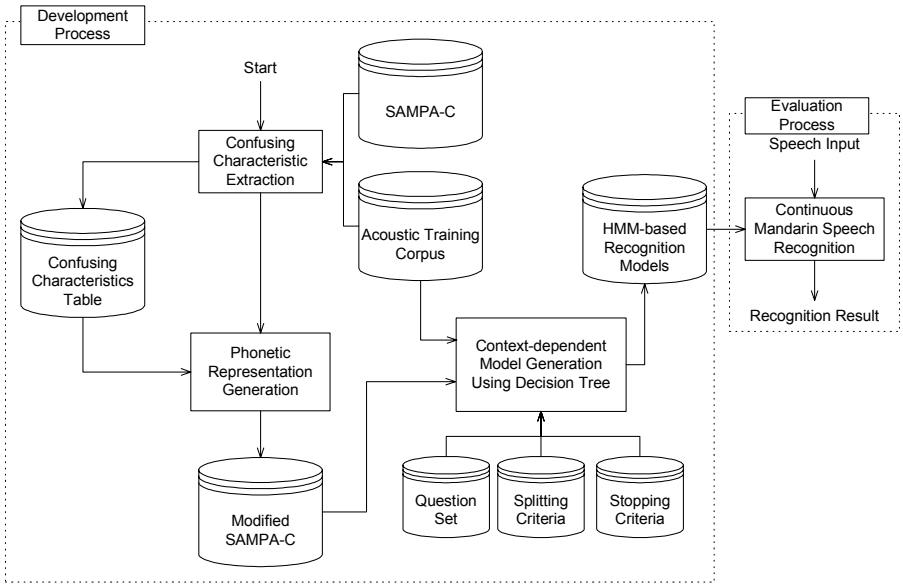

Fig. 1. Overall scheme for developing the HMM-based recognition models using modified SAMPA-C

2. Analysis of Confusion Characteristics

To accurately represent the confusion characteristics of Mandarin speech, the subsyllables are used as the basic units in the analysis process and extracted from the recognition outputs generated by the SAMPA-C based syllable recognizer. In this analytic procedure, 50 context-independent left-to-right HMMs with 4 states and 12 mixtures are built as the basic recognition models. 1551 utterances generated by 80 speakers in Mandarin Speech Database Across Taiwan (MAT) are used for advance analysis. In the following tests, the effect of the confusion characteristics between every two subsyllables is considered.

2-1 Testing for subsyllable-dependent confusion characteristics

To clarify the subsyllable-dependent confusion characteristics, the training and misrecognized data are used and divided into several categories, which are defined as subsyllable attributes (SA). For each SA, the numbers of occurrences and misrecognitions generated by the recognizer are accumulated. Then, these two corresponding frequency distributions of the training and misrecognized data, treated as SA distributions, are utilized to quantitatively analyze the confusion degree by using the χ^2 testing. The χ^2 value, which is greater than a threshold of the predefined significance level, implies that the SA distributions can be regarded as different. Accordingly, several subsyllables are treated as confusable and need further discrimination. The formula to calculate the χ^2 value is defined as follows.

$$\chi^2 = \sum_{i=1}^{N} \frac{(M_i - E_i)(M_i - E_i)}{E_i} \tag{1}$$

where N is the number of SAs, M_i is the number of misrecognitions of the *i-th* SA and E_i is the expected value of M_i and can be defined as

$$E_i = W_i \frac{\sum_{j=1}^{N} M_j}{\sum_{j=1}^{N} W_j} \tag{2}$$

where W_i is the number of appearances of the *i-th* SA.

The effects of confusion characteristics are analyzed and extracted from the recognition outputs generated by the SAMPA-C based syllable recognizer. Table I and Table II show two SA distributions of INITIALs and FINALs represented by SAMPA-C, respectively. It is clear that “d” and “V:” has the largest number of appearances in INITIALs and FINALs. However, the tendency of “dC” and “IM” was misrecognized frequently more than that of “d” and “V:”, respectively. “dC”, “IM”, “d”, and “V:” are the Mandarin syllables represented by SAMPA-C. As a result for a Mandarin speech recognizer, the confusion characteristics seems to strongly depend on the subsyllables. Next, since insufficient training data happen for some SAs, the χ^2 testing conditions might not be satisfied. Thus, the following two conditions in each SA have to be considered [5].

(1) The percentage of the expected value over five is above 80%.

(2) All expected values are more than one.

In Table I and Table II, the χ^2 values are 164 and 97 for INITIALs and FINALs, respectively. It is clear that the χ^2 value is greater than 5% of the significance level. Therefore, the analyzed results

show significant evidences that the confusion characteristics of INITIALs and FINALs can be regarded as subsyllable-dependent.

Table I. SA distributions of INITIALs represented by SAMPA-C, χ^2 value = 164, $p \leq 0.05$

INITIAL	NULL	b	p	m	f	d	t	n	l	g	k
Number of appearances	141	65	46	47	17	227	71	87	97	59	56
Number of misrecognition	49	27	11	10	6	57	38	23	26	21	14
INITIAL	h	dC	tC	C	dZ	tS	S	R	dz	ts	s
Number of appearances	80	54	51	49	62	56	72	52	57	49	56
Number of misrecognition	18	51	31	19	32	51	38	11	36	44	20

Table II. SA distributions of FINALs represented by SAMPA-C, χ^2 value = 97, $p \leq 0.05$

FINAL	NULL	a:	O:	V:	ai	ei	aU	ou	aM	@M	aN	VN	r
Number of appearances	38	72	8	194	60	40	61	53	69	52	59	56	8
Number of misrecognition	11	32	3	33	24	14	36	29	13	20	18	14	2
FINAL	i:	ja:	jE	jai	jaU	jou	jEM	IM	jaN	IN	u:	wa:	wO:
Number of appearances	41	15	37	4	30	29	47	34	25	43	58	21	47
Number of misrecognition	21	11	11	2	6	11	25	25	6	22	13	8	22
FINAL	wai	wei	waM	w@M	waN	wVN	y:	yE	yEM	yM	yN		
Number of appearances	23	57	58	38	27	53	17	23	24	14	16		
Number of misrecognition	8	9	23	20	12	4	11	11	12	4	8		

2-2 Examination of confusable phonetic set

According to the previous analysis, the misrecognition happens in some specific SAs. In general, the misrecognition is caused by the incorrect pronunciation or the confusable phonetic set. The incorrect pronunciation is due to inarticulacy such as the retroflexion in Mandarin speech. For examples, the "tS" and "IN" is usually pronounced as "ts" and "IM" in INITIALs and FINALs, respectively. Thus, in this paper, the confusion characteristic of each recognition units in the SAMPA-C based recognizer has to be examined and the phonetic set should be redefined. Table III shows some examples of SA distributions of confusions for recognition units in SAMPA-C. The upper two measures show the χ^2 values are greater than 5% of the significance level and the phoneme will cause the subsyllable-dependent confusion according to the χ^2 testing. On the other hand, the lower two measures show the χ^2 values are smaller than 5% of the significance level and these subsyllables possess less confusion characteristic.

Table III. Comparison of SA distributions of syllables represented by concatenating (+) phonetic units in SAMPA-C

Subsyllable	d	d+C	d+Z	d+z
Num. of appearances	227	54	62	57
Num. of misrecognition	57	51	32	36
χ^2 value = 55, $p \le 0.05$				

(a)

Subsyllable	y+E	y+E+M	y+M	y+N
Num. of appearances	23	14	14	16
Num. of misrecognition	11	4	4	8
χ^2 value = 1.65, $p \ge 0.05$				

(b)

2-3 Determination of confusable phones

Given a subsyllable *A*, the subsyllable-dependent confusion characteristic between subsyllables A and B can be analyzed in Table IV, which show the four possible outcomes for a given trial. The confusion relationship between subsyllables *A* and *B* can be shown in Fig. 2. According to this representation, the χ^2 testing serves as a way to quantify the confusion between

these two distributions. Hence, based on the four outcomes in Table IV, the χ^2 testing can be applied to determine the degree of confusion between subsyllables *A* and *B* and is given by

$$\chi^2 = \sum_{cells} \frac{\left(f_{ij} - E_{ij}\right)\left(f_{ij} - E_{ij}\right)}{E_{ij}} \tag{3}$$

where f_{ij} is the observed frequency. E_{ij} is the expected frequency and defined as

$$E_{ij} = f_{i0} \frac{f_{0j}}{\sum_{k=1}^{2} f_{k0}} \tag{4}$$

where f_{i0} is the totals of the *i-th* row and f_{0j} is the totals of the *j-th* column. If the value in Table IV is small, Yate's correction method is used to estimate a robust χ^2 value [6]. Therefore, the confusable phone, which causes the subsyllable-dependent confusion, can be found. Table V shows some examples of confusion measure. In this table (a) and (b) have high confusion contrast to (c) and (d). Accordingly, subsyllable "d+C" and subsyllable "U+N" are likely confused with subsyllable "C" and subsyllable "i+U+N", respectively.

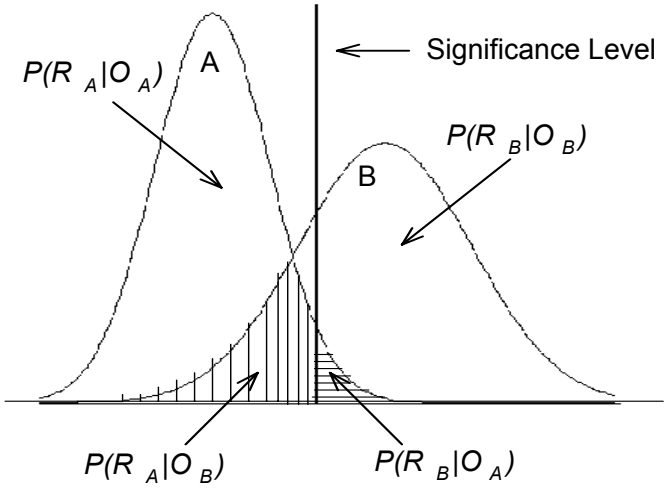

Fig. 2. Confusion relationship of subsyllables *A* and *B*

Table IV. Four possible outcomes for a given trial

		Recognition Result	
		R_A	R_B
Observations	O_A	$P(R_A\|O_A)$	$P(R_B\|O_A)$
	O_B	$P(R_A\|O_B)$	$P(R_B\|O_B)$

Table V. Examples of confusion measure (number of appearances)

		Recognition Result	
		d+C	C
Observations	d+C	3	23
	C	3	30
χ^2 value = 0.00265, $p \geq 0.05$			

(a)

		Recognition Result	
		U+N	i+U+N
Observations	U+N	49	0
	i+U+N	7	8
χ^2 value = 0.00146, $p \geq 0.05$			

(b)

		Recognition Result	
		d+C	d
Observations	d+C	3	12
	d	0	170
χ^2 value = 23.16, $p \leq 0.05$			

(c)

		Recognition Result	
		U+N	V+N
Observations	U+N	49	3
	V+N	2	42
χ^2 value = 76.98, $p \leq 0.05$			

(d)

3. Design of the Modified SAMPA-C

Based on the analysis of confusion characteristics, several confusion subsyllables caused by the confusable phonetic representation can be extracted. The confusable phonetic representation can be automatically detected using the above process. In our experimental results, the automatic speech recognition based on SAMPA-C cannot model the rapid variation between subsyllables. This is because that the confusion always occurs in the short duration between two subsyllables and the phonetic units representing the short phones cannot model this short duration well. Accordingly, a longer phonetic representation similar to subsyllable units is adopted to eliminate the confusion between two confusable subsyllables. These unsuitable phonetic units are manually analyzed. Each unit is concatenated with other phonetic unit to form a new, longer phonetic unit. The testing process is performed on the new representation iteratively. Finally, a modified SAMPA-C phonetic set, which suitably represent Chinese pronunciation is obtained and listed in Table VI. The original SAMPA-C phonetic set is also listed in Table VI for comparison. The phonetic units with boldface are the newly defined units. For example, the new phonetic unit “G” is defined by concatenating the phonetic units “d” and “C.” The total number of phonetic units in the modified SAMPA-C becomes 52 compared to 45 in the original SAMPA-C.

Table VI. Modified SAMPA-C and the examples with the corresponding Chinese characters and PinYin

Modified SAMPA-C	Examples by PINYIN	Modified SAMPA-C	Examples by PINYIN
G(d+C)	GIN (晶 jing1)	z(d+z)	zI: (子 zi3)
Q(t+C)	Qi: (七 qi1)	c(t+s)	cu@M (村 cun1)
X(C)	XiaU (小 xiao3)	aN(a+N)	laN (狼 lang2)
Z(d+Z)	ZUN (中 zhong1)	aM(a+M)	maM (慢 man4)
C(t+S)	Ca: (茶 cha2)	iU(I+U)	XiUN (兄 xiong)

4. Experimental Results

In the experiment setup, a Mandarin Speech Across Taiwan (MAT) telephone speech database, pronounced by 160 speakers (81 males, 79 females), with 8,237 files (sampling rate of 8kHz) was employed. Another speech database with 500 utterances was also collected and used as the testing data. In the following experiments, 12 Mel-Frequency Cepstrum Coefficient (MFCC), 12 delta MFCC, one delta log energy, and one delta delta log energy are extracted as a 26-dimension feature vector.

In the first experiment, the SAMPA-C based recognizer and the modified SAMPA-C based recognizer were built for the comparison of recognition performance. In these systems, the context-independent models were adopted and the subsyllable recognition rates of INITIALs and FINALs for the two systems are listed in Table VII.

Table VII. Recognition rates using SAMPA-C and modified SAMPA-C, respectively

	SAMPA-C	Modified SAMPA-C
INITIAL	55.86%	75.08%
FINAL	66.53%	67.26%

For Mandarin speech, the confusion effects of INITIALs are more obvious than that of FINALs. Due to the channel distortion of telephone network, the unvoiced INITIAL part with short duration is easy to be misrecognized. Therefore, the confusion between INITIALs can be discriminated using the modified SAMPA-C and the recognition performance can be improved

significantly.

Moreover, another phonetic representation set is also developed for evaluating the confusion characteristics analysis. This phonetic representation with 58 fundamental subsyllables [7-9] was adopted in this experiment. With the same training database, the distributions of misrecognition for subsyllable “dC”, “C”, “dZ”, and “d” are shown in Fig. 3. The subsyllable “dC” is usually misrecognized to “C”. However, the subsyllable “C” is not usually misrecognized to “dC”. It is difficult to detect the confusion characteristic of subsyllable “dC”. In our approach, the χ^2 value of “dC” compared with other subsyllables is shown in Fig. 4. The confusion characteristic of subsyllable “dC” can be detected. For the significance level, the subsyllable “C” usually confused with subsyllable “dC”.

In the next experiment, the context-dependent models were applied for evaluation and the experimental results are shown in Fig. 5. It is clear that the modified SAMPA-C can achieve an encouraging recognition performance, which is better than that obtained using the SAMPA-C. Especially, for the context-dependent models, the confusion between syllables can be efficiently discriminated and the recognition performance can also be improved.

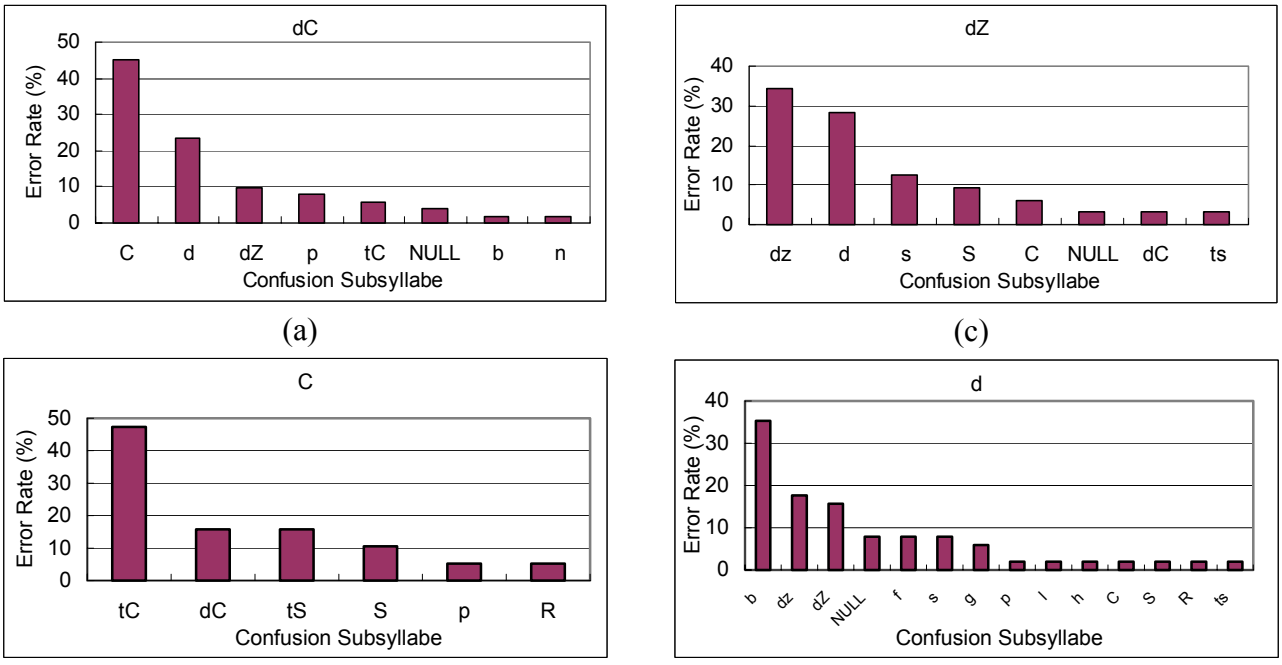

Fig. 3. Distributions of error rate for subsyllables (a) dC, (b) C, (c) dZ, and (d) d.

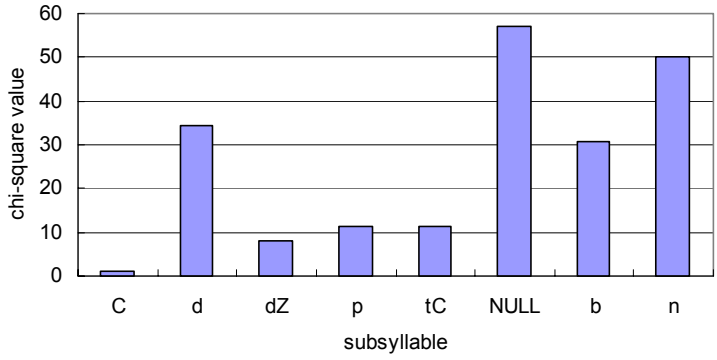

Fig. 4. χ^2 value of "dC" compared with other subsyllables

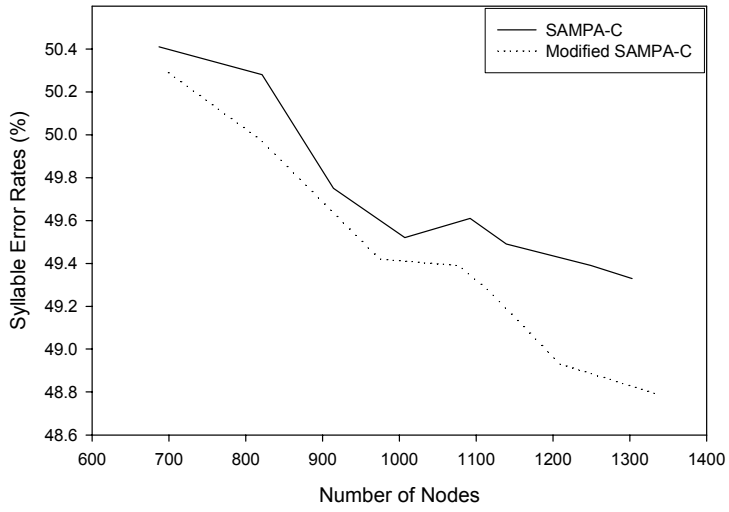

Fig. 5. Syllable error rates with respect to SAMPA-C and modified SAMPA-C based recognition system.

In order to evaluate the performance of different phonetic representations, we conducted experiments on three continuous syllable recognition model types. Three forms of subsyllablic units – right-context dependent INITIAL/FINAL (RCD-IF), SAMPA-C based tri-phones, and modified SAMPA-C based tri-phones were conducted to evaluate the syllable recognition rates (SRR). Table VIII shows the experimental results and the modified SAMPA-C based approach outperformed the other two types.

Table VIII. Syllable recognition rates using RCD IF, SAMPA-C based tri-phones, and modified SAMPA-C based tri-phones

	RCD IF	SAMPA-C Tri-phones	Modified SAMPA-C Tri-phones
No. of Nodes	675	754	812
SRR	46.12%	43.23%	50.23%

5. Conclusions

In this paper, the confusion characteristics for Mandarin speech using SAMPA-C were analyzed. The confusion characteristics generated with respect to confusable phonetic set can be discriminated by incorporating a statistical categorical data analysis method without any model assumption. Redefining the phonetic set, the effect of the confusion characteristics can be reduced and the recognition performance can be improved significantly. Hence, a modified SAMPA-C is proposed to provide a corresponding phonetic representation for building more reliable recognition models. Experimental results show that the proposed approaches give an encouraging improvement. For the portability to other languages, the proposed procedure can be easily applied to detect the confusion phonetic units of that language. Accordingly, a more reliable phonetic set for that language can be obtained.

6. Acknowledgment

The authors would like to thank the National Science Council, R.O.C., for its financial support of this work, under Contract No. NSC89-2614-H-006-004-F20. The paper is also a partial result of Project 3XS1B11 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.

7. References

[1] R. H. Mathews, Mathews' Chinese-English Dictionary, Caves, 13th printing, 1975.

[2] J. Wells, *EAGGLES Handbook on Spoken Language Systems(DRAFT) – SAMPA computer readable phonetic alphabet*, <http:// www.phon.ucl.ac.uk/home/sampa/home.htm>, 1997.

[3] F. Seide, and N. J. C. Wang, "Phonetic modeling in the Philips Chinese continuous-speech recognition system", *Proc. of ISCSLP'98*, 1998, pp. 54-59.

[4] A. Agresti, Categorical Data Analysis, John Wiley & Sons, 1990.

[5] W. G. Cochran, “Some methods for strengthening of common tests”, *J. of the International Biometric Society*, 1954, pp. 417-451.

[6] W. J. Krzanowski, Principles of Multivariate Analysis. Oxford University Press, New York, 1988.

[7] S. H. Chen, Y. F. Liao, S. M. Chiang, and S. Chang, “An RNN-based pre-classification method for fast continuous Mandarin speech recognition”, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 1, 1998, pp. 86-90.

[8] R. Y. Lyc, I. C. Hong, J. L. Shen, M. Y. Lee, and L. S. Lee, “Isolated Mandarin based-syllable recognition based upon the segmental probability model”, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 3, 1998, pp. 293-299.

[9] C. H. Wu, Y. J. Chen, and G. L. Yan, “Integration of phonetic and prosodic information for robust utterance verification”, IEE Proceedings-Vision, Image and Signal Processing, Vol. 147, 2000, pp. 55-61.