SUB-SYLLABIC ACOUSTIC MODELING ACROSS CHINESE DIALECTS

Wai-Kit LO[1], Helen M. MENG[2] and P.C. CHING[1]

[1]Digital Signal Processing Laboratory, Dept. of Electronic Engineering,
[2]Human-Computer Communications Laboratory, Dept. of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong, Hong Kong
wklo@ieee.org, hmmeng@se.cuhk.edu.hk and pcching@ee.cuhk.edu.hk

ABSTRACT

This paper presents a series of experiments on sub-syllabic unit selection across the two Chinese dialects – Mandarin and Cantonese. Evaluations are based on syllable recognition using only acoustic information, and no lexical knowledge is incorporated. We use a variety of subsyllabic acoustic models, motivated by phonological and lingustic structures charactersitics of Chinese. Our results should provide a useful reference for work in large-vocabulary Chinese speech recognition, as well as related tasks, e.g. spoken document retrieval.

1 INTRODUCTION

This work utilizes a variety of acoustic models for the task of syllable recognition across two Chinese dialects – Mandarin and Cantonese. Our ultimate goal is to tackle a Chinese spoken document retrieval task, and we start off by finding a set of acoustic units which can be used to index Chinese spoken documents with high efficacy. In this study, we used a variety of syllable-based acoustic models for indexing Chinese audio, because such units should be able to fully index the audio and circumvent the out-of-vocabulary problem [1, 2, 3].

2 PROPERTIES OF CHINESE

We work with two key dialects of the Chinese language – Mandarin (or Putonghua), the official spoken language used in China; as well as Cantonese, the commonly used language in Hong Kong and Macau. Both Mandarin and Cantonese are based on the same writing system with Chinese characters. In their spoken form, they are both monosyllabic and tonal. Each syllable can be decomposed into a syllable initial, a syllable final and a tone.

However, the two dialects also differ significantly, to the extent that a speaker knowing only one of the dialects may not be able to communicate with another speaker knowing only the other dialect. The differences between Mandarin and Cantonese reside in phonetics, syntax and vocabulary selection. Mandarin has 24 initials and 37 finals, constituting 410 distinct base syllables. The dialect also has 5 lexical tones, giving a total of 1,400 tonal syllables [6]. Cantonese has 20 initials and 53 finals, constituting 660 base syllables. It has 6 lexical tones, giving a total of 1800 distinct tonal syllables. The inventory of syllable intials and finals differ between the two dialects.

This work utilizes a variety of sub-syllabic acoustic units for recognition. The selection is motivated by the phonological and linguistic structures of Chinese.

3 CORPORA

The experiments conducted in this work are based on two common and publicly available Chinese corpora. Hence, our results should provide some useful benchmarks on large-vocabulary Chinese speech recogntion.

3.1 Mandarin corpus – Hub 4 Non-English

Our Mandarin corpus is the Voice of America (VOA) news portion of the Hub4 Non-English corpus. It is a transcribed radio broadcast news corpus. It consists of around 11 hours of orthographically transcribed audio data together with time aligment on a sentence-by-sentence basis. The evaluation set used is the VOA portion of the formal evaluation set in Hub4 [11].

3.2 Cantonese corpus – CUSENT

Our Cantonese corpus selected is CUSENT [9], a continuous speech corpus of Cantonese read sentences from Hong Kong newspapers. To our knowledge, this is the only publicly available transcribed Cantonese corpus. The evaluation set is the designated test set in CUSENT. This work provides the the first published evaluation based on the entire CUSENT corpus.

Because of the difference in nature and properties between the two corpora, the reader should exercise caution when interpreting the evaluation results. Comparisons should only be made with consideration on their differences. Table 1 lists a number of differences between the corpora.

	Hub4 Mandarin (VOA)	CUSENT
Duration	11 hours	19 hours
Style	Broadcast news	Read speech
Training speakers	A few	68
Testing speakers	A few speakers, Overlaps with training speakers	12 speakers. No ovelap with training speakers

Table 1. Description of the Hub4 Mandarin and CUSENT corpora.

4 ACOUSTIC UNIT SELECTION

The syllable is a natural and intuitive unit for acoustic modeling of Chinese dialects. The main advantage of the syllable unit is that it provides full phonological coverage of the Chinese language. However, for training speech recognizers, it may occur that we do not have sufficient data to train the large set of syllable models. In this case, we may choose to break down the syllables into smaller units. In our experiments, a number of sub-syllabic units have been used for the acoustic modeling of the Chinese dialects. The selection of the units is phonologically motivated.

4.1 Acoustic Units for Mandarin

The units selected for acoustic modeling of Mandarin are: base syllable (BS), tonal syllable (TS), initial-final (IF), initial-tonal final (ITF), preme-core final (PC), preme-toneme (PT).

Base syllables have no information about tones but tonal syllables do. Initial-final is a common decomposition of the Mandarin syllable. Initial-tonal final is similar to initial-final units except for the inclusion of the lexical tone in the syllable final. Preme-core final is a special decomposition for the Mandarin syllable. It is the same as initial-final unless the syllable has four phones (i.e. the final is a triphthong) – here the initial is combined with the glide of the final to form the *preme*. The remaining portion(s) of the syllable constitutes the *core-final*. *Tonemes* are tone-carrying core-finals.

4.2 Acoustic Units for Cantonese

Cantonese syllables do not contain the medial glide. Consequently, there is no preme/core-final/toneme decomposition. The acoustic units selected for Cantonese include: base syllable (BS), tonal syllable (TS), initial-final (IF), and initial-tonal final (ITF). The syllable decompositions are same as that in Mandarin, but involves a different set of sub-syllabic units. Figure 1 shows the decomposition of the syllable unit in (a) Mandarin and (b) Cantonese.

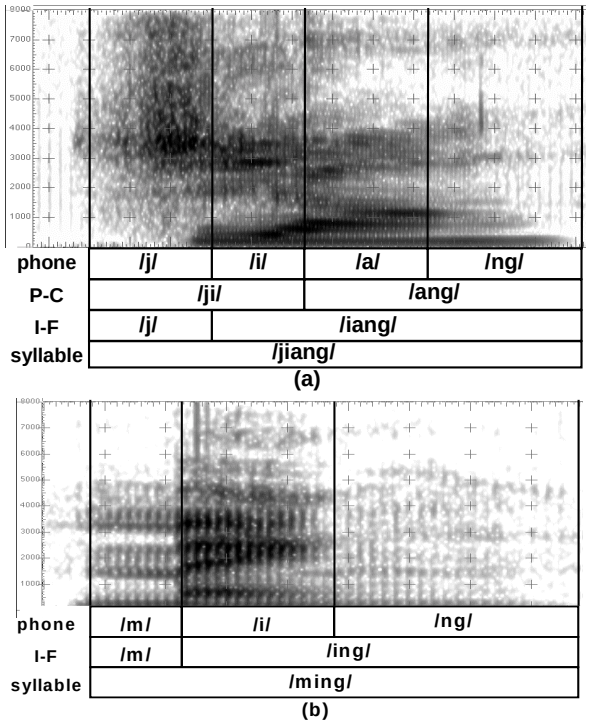

Figure 1*. Illustrations of (a) a Mandarin syllable /jiang/ and (b) a Cantonese syllable /ming/ together with their respective syllable decompositions[8].*

5 ACOUSTIC MODEL TRAINING

5.1 Segmentation and Feature Extraction

The Hub4 data is segmented according to the sentence boundaries provided. We obtain the phonetic transcription by means of the orthographic transcription together with a dictionary lookup using the CALLHOME lexicon [10]. CUSENT is processed differently since it contains short, read sentences. The transcription provided was used directly.

The acoustic features used are the mel-frequency cepstral coeffcients (MFCC) which includes 12 cepstral coefficients together with the energy. The features are also energy normalized and cepstral mean normalized based on each segment. By including the first and second derivatives of the parameters, the final feature vectors have 39 elements.

5.2 Model Training and Evaluation

The estimation of the HMM model parameters is based on the Baum-Welsh re-estimation algorithm. HMM models of different number of mixtures are trainned iteratively and is exponentially increased from 1, 2, 4, 8 to 16. The models used in the evaluation are all 16 mixtures.

In order to capture contextual information in sub-syllabic modeling, context-dependent models are built. For each of the sub-syllabic units, context independent (CI), right context-dependent (BI) and

left-and-right context-dependent (TRI) models are built.

When contextual information is considered, we obtained a very large number of models. To cater for the insufficiency of data, HMM states are combined using phonetic tree-based clustering. The questions used are all phonetically-oriented based on the place and manner of articulation of the units. Since the phonetic properties of Mandarin and Cantonese are different, the set of questions for the two dialects are formulated separately.

Evaluations are carried out using the trained models on the designated test sets of the corpora. A one-pass Viterbi beam search is applied *without* any language model. Hence performance comparison is based purely on acoustic modeling using the different syllabic / sub-syllabic units. Our measure of performance is the base syllable / tonal syllable accuracies.

6 RESULTS

Results are tabulated in Tables 2 and 3.

	Base Syllable					
	BS	**TS**	**IF**	**ITF**	**PC**	**PT**
CI	57.31 (392)	57.01 (1050)	50.37 (68)	55.49 (204)	50.47 (84)	55.44 (175)
BI			65.22 (833)	66.22 (1368)	65.45 (835)	66.52 (1362)
TRI			67.34 (5307)	67.52 (4857)	**68.07 (5932)**	67.62 (4896)
	Tonal Syllable					
CI		43.88 (1050)		37.80 (204)		36.30 (175)
BI				47.28 (1368)		47.32 (1362)
TRI				**49.00 (4857)**		48.98 (4896)

Table 2*. Syllable accuracies (%) of different Mandarin sub-syllabic units. (model counts in parentheses)*

	Base Syllable			
	BS	**TS**	**IF**	**ITF**
CI	56.13 (638)	52.25 (1615)	62.05 (81)	64.99 (301)
BI			78.37 (1031)	77.68 (1993)
TRI			**79.47 (8458)**	78.66 (45556)
	Tonal Syllable			
CI		32.25 (1615)		38.92 (301)
BI				49.15 (1993)
TRI				**50.69 (45556)**

Table 3*. Syllable accuracies (%) of different Cantonese sub-syllabic units. (model counts in parentheses)*

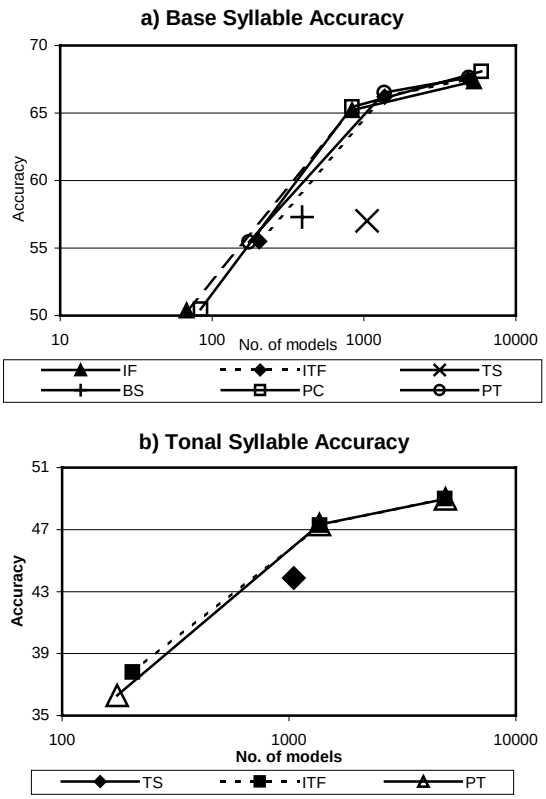

Figure 2. *Mandarin sub-syllabic unit operating curves (recognition accuracy against the number of models)*

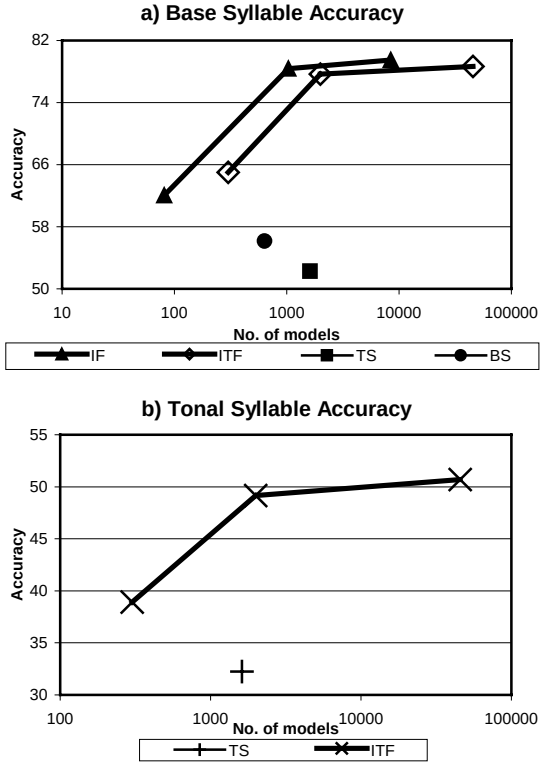

Figure 3 *Cantonese sub-syllabic unit operating curves (recognition accuracy against the number of models)*

If we plot the accuracies of the syllabic / sub-syllabic units against the model-counts, we will have

an illustration of the relationship between the model complexity and recognition accuracy. Complexity increases as a greater amount of articulatory context is captured in the acoustic models. Figures 2 and 3 are the operating curves for Mandarin and Cantonese respectively.

For base syllable recognition in Mandarin, tri-PC gave the best accuracy at 68.07%. The preme-core-final/preme-toneme decompositions seem to do better than the initial-final/initial-tonal final decompositions. For tonal syllable recognition, the tri-ITF gave the best performance at 49.00%. There is no consistent difference between the use of preme-toneme or initial-tonal final.

For base syllable recognition in Cantonese, tri-IF gave the best performance of 79.47%. For tonal syllable recognition, tri-ITF gave the best performance at 50.69%. The advantage of using context is apparent in these cases.

In the operating curves, we made use of the number of models as an indication of computation complexity. It can be seen moving from CI to BI context gave a significant gain in all our curves, but the incremental gain achieved as we move from BI to TRI context decreased. This suggests that a desirable tradeoff between accuracy and computation is obtained from the use of BI context models.

7 CONCLUSIONS

In this paper, we have presented a series of experiments involving a variety of syllabic / sub-syllabic acoustic models for continuous speech recognition across two Chinese dialects – Mandarin and Cantonese.

For base syllable recognition in Mandarin, the sub-syllabic models have a clear advantage over the syllabic models, mainly due to the availability of training data. Among the sub-syllabic models, the preme-core-final / preme-toneme decompositions have a slight advantage over the initial-final / initial-tonal final decompositions. The inclusion of a greater amount of context improves recognition performance as expected, at the expense of a higher computation complexity. Similar trends are observed in tonal syllable recognition for Mandarin, as well as base/tonal syllable recognition in Cantonese.

Our operating curves suggest that the BI-context may strike a better balance between high recognition accuracies and low computational complexities. In addition, even though the training and testing speakers may be overlapping in Mandarin corpus, the evaluation results for the Cantonese corpus is slightly better. This may be attributed to the read nature of the Cantonese corpus. Based on these two corpora, the sub-syllabic unit that is most desirable for *both* Cantonese dialects is the BI-IF unit.

Furthermore, this paper also presents the first published evaluation results based on the CUSENT corpus. Our results should serve as a useful benchmark for Chinese speech recognition across the two dialects.

8 ACKNOWLEDGEMENT

The CUSENT corpus is developed with support from the Industrial Support Fund of the Hong Kong SAR Government. We also thank Y. W. Wong for assistance with the training and testing procedures.

9 REFERENCES

[1] NG, K. *et. al.*, "Subword unit representations for spoken document retrieval," *Eurospeech-97*, Rhode Island, 1997

[2] CHEN, B. *et. al.*, "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics," *ICASSP2000*, Istanbul, 2000.

[3] MENG, H. *et. al.*, "A study on the use of syllables for Chinese spoken document retrieval," *Technical report SEEM99-11*, CUHK, Hong Kong, 1999.

[4] 徐世榮, "普通話語音常識," *語文出版社*, 北京, 1993.

[5] CHING, P. C., *et. al.*, "From phonology and acoustic properties to automatic recognition of Cantonese," *ISSIPN-94*, Hong Kong, 1994.

[6] GAO, S. *et. al.*,"Acoustic modeling for Chinese speech recognition : A comparative study of Mandarin and Cantonese," *ICASSP2000*, Istanbul, 2000.

[7] The 1998 Topic Detection and Tracking projects, `http://www.itl.nist.gov/iaui/894.01/tdt98/tdt98.htm`, NIST-ITL, U.S.A., 1998.

[8] MENG, H. *et. al.*, "Mandarin-English Information (MEI): Investigating translingual speech retrieval," *Embedded MT Workshop, NAACL*, Seattle, 2000.

[9] LO, W. K. *et. al.*, "Development of Cantonese spoken language corpora for speech applications," *ISCSLP98*, Singapore, 1998.

[10] Linguistics Data Consortium (LDC), `http://www.ldc.upenn.edu.`

[11] The 1997 Mandarin broadcast news evaluation, `http://www.itl.nist.gov`, NIST-ITL, U.S.A., 1997.