A New Framework for Recognition of Mandarin Syllables With Tones Using Sub-syllabic Units

Chih-Heng Lin† ‡, Lin-Shan Lee‡ and Pei-Yih Ting†

† Telecommunication Labs, Ministry of Transportation and Communications,
‡ Department of Electrical Engineering, National Taiwan University,
Taiwan, Republic of China

ABSTRACT

The recognition of all the 1302 Mandarin syllables is the key problem in large vocabulary Mandarin speech recognition. Because every Mandarin syllable is assigned a tone, when the differences in tones are disregarded, the total number of different syllables is reduced from 1302 to 408. They are referred to as the 408 base syllables(each can bear different tones) here in this paper. A conventional approach of Mandarin syllable recognition has been widely accepted, in which the tones and the 408 base syllables are separately recognized in parallel by two sub-sytems. In this paper, on the other hand, three classes of sub-syllabic units for Mandarin syllables are defined, i.e., the Initials, the Finals, and the Transitions, and a new structure for Mandarin syllable recogniton is developed, in which the tones and base syllables are recognized jointly and a total of 574 sub-syllabic unit models will be enough to provide improved recognition performance.

1. INTRODUCTION

In Mandarin Chinese every character is pronounced as a mono-syllable and there exist a total of 1302 syllables. Furthermore, Mandarin is a tonal language and there are in general 4 lexical tones and 1 neutral tone in it. Some typical example waveforms for the syllables "jian3" and "ian3" are shown in parts (a) of Figs. 1 and 2 respectively, where the last digits "3" indicate the tones of the syllables.

When the differences in tones are disregared, for example, the syllables "ba1", "ba2", "ba3", "ba4" and "ba5" are considered as a single base syllable "ba" which can bear different tones, the total number of different syllables is reduced from 1302 to 408. Based on the assumption that the pitch or excitation behavior which basically differentiates the tones is essentially independent of the vocal tract parameters which differentiate the base syllables, a general structure of Mandarin syllable recognition [1] as shown in Fig. 3(a) has been widely accepted and used. In this structure the tones and 408 base syllables are seperately recognized in parallel by two sub-systems, one with tone models primarily based on pitch behavior and one with 408 base syllable models primarily based on vocal tract parameters. The basic idea of such a structure is to divide a difficult problem into smaller problems. It seems easier to distinguish 408 base syllables than 1302 syllables. Smaller searching space makes real-time recognition achievable, and sharing of training utterances can significantly reduce the required training data. However, such a structure is in fact not necessarily the best.

In this paper, a new approach for Mandarin syllable recognition based on sub-syllabic units is proposed in section 2, and how the sub-syllabic units are chosen is explained in section 3. The training and the recognition procedures are prsented in section 4, while the speech database and feature parameters employed are described in section 5. finally Section 6 then presents the experiment results, and concluding remarks are finally made in section 7.

2. THE PROPOSED NEW APPROACH

The basic problem with the conventional approach in Fig 3(a) is that there always exist correlation between excitation and vocal tract parameters[2] and such correlation is not negligible. The different tones in fact give slightly different vocal tract behavior, thus the modeling of base syllables to include different tone behavior becomes rather difficult. This is one of the major reasons why the base syllable recognition rate can't be very high[1]. Similarly, different base syllables also produce slightly different pitch behavior even in syllables with the same tone. This makes the modeling of tones difficult, and is one of the reasons why the tone recognition rate can't be very high even if there is only five different tones. Furthermore, an error in any of the two subsytems shown in Fig. 3(a) will definitely lead to an error in the finally recognized syllable. In other words, the *a posterior* probability of a tone T^i and a base syllable S^j for a given test utterance A has to be determined by those of a tone T^i and a syllable S^j respectively. i.e. $P(T^i, S^j|A) = P(T^i|A) * P(S^j|A)$. This makes the syllable recognition rate even lower.

On the other hand, considering the fact that the characteristics of the excitation and the vocal tract behaviors are naturally correlated with each other, a new structure for Mandarin syllable recognition as shown in Fig.3(b) is proposed in this paper, so that the base syllables and tones can be jointly recognized. Due to the special characteristics of

0-7803-0946-4/93 $3.00

Mandarin syllables, three classes of sub-syllabic units are defined, i.e., the Initials, the Finals, and the Transitions. The detailed discussions in the following will show that a total of 574 sub-syllabic unit models will be enough to discriminate all the 1302 syllables including tones.

3. THE CATEGORIZATION OF MANDARIN SUB-SYLLABIC UNITS

Each Mandarin syllable is conventionally decomposed into Initial part and Final part very similar to consonant/vowel parts in other languages. The Initial part is the initial consonant and the Final part includes the vowel or diphthong part but including possible medial or nasal ending. In general 38 Final's and 22 Initial's (including a null Initial) are defined in Mandarin. All these 22 Initials and 38 Finals are listed respectively in Tables 1 and 2. In table 2, the 38 Finals are further grouped into 7 categories, where all members in the same category approximately start with the same phoneme. This will be helpful in assigning the sub-syllabic units later. Because the Initials are in general shorter and unstable, the 408 base syllables can be grouped into 38 confusing sets, each consisting of syllables sharing the same Final but with different Initials. Previous experiments show that syllables within these confusing sets cause most of the errors of base syllable recognition.

If we choose the Initials and Finals as the sub-syllabic units, a syllable can be represented by either of the two types in parts (b) of Fig. 1 and 2 respectively, in which the shaded area represents the transition region[3], which is represented by a third type of sub-syllabic units, the Transition. There are two different kinds of Transition units, one is those between an Initial and a Final, and the other is those between the silence and a Final, just as in the sub-syllabic structures for the syllables "jian3" and "ian3" respectively. In figure 1(b), the first sub-syllabic units corresponds to the Transition between the silence and the Final "ian", and the second sub-syllabic unit corresponds to the Final "ian". In figure 2(b), the first sub-syllabic unit corresponds to the Initial "j", the second sub-syllabic unit corresponds to the Transition, and the third sub-syllabic unit listed corresponds to the Final "ian".

In our case, we make the models for all Initials and Finals context independent and the models for Transitions context dependent , according to the Initials of Table 2 and the categorization of the Finals at Table 1. In other words, the Finals in the same category of Table 1 don't make any difference for the context-dependence of the Transitions. Moreover, the models for Initials are tone independent while the models for Transitions and for Finals are tone dependent such that they keep the capabilities for tone classification. [4]. As in Fig. 1(b) and 2(b), the Transition of Fig. 1(b) is contextual dependent on the phoneme "j" and "i", while the sub-syllabic units "ian" in both figures correspond to the same Final model which carries the third tone.

4. THE REPRESENTATION BY HMM

Using a left to right state sequence with equal probabilities for the 2 transitions leaving from a state, each sub-syllabic unit mentioned above is characterized by a hidden Markov model(HMM), whose state observation functions are described by the continuous probabilities of a mixture of Gaussian distributions. The HMM of a syllable is then formed by the concatenation of HMM's for the corresponding component sub-syllablic units. In the experiments, each Initial and Transition unit has 1 state, while each Final unit has 3 states. Thus different syllables can have different numbers of states, depending on their composition of sub-syllabic units, as can be found in the syllables shown in part (c) of Figs 1 and 2. To estimate the parameters of the HMM's, a training procedure [5] based on the segmental k-mean algorithm is employed, which can be used to obtain all the parameters for the HMM's for all the sub-syllablic units without having to segment the training utterances for each sub-syllablic unit beforehand. To initialize the training procedure, the training utterances are first uniformly segmented. The sharing of training utterances by the sub-syllabic units actually significantly reduced the required training data, in addition to the improved robustness of the models.

During the rcognition stage, the recognition process is based on the Viterbi algorithm[6]. Because less computation are needed for less number of states with this sub-syllablic unit approach, the recognition speed is in fact significantly improved as compared to those based on whole syllable units. Meanwhile, the search space for the 1302 syllables can be easily reduced by the structure of the sub-syllabic unit combined with carefully designed n-best level-building algorithm[7].

5. THE SPEECH DATABASE AND THE FEATURES

The speech database used in the experiments includes syllabic utterances pronounced in isolation by two male speakers, and each speaker produced 6 utterances for each of the 1302 syllables. All the utterances have been endpoint detected and sampled at 10 k Hz. Cepstral coefficients of order 10 and the corresponding 10 delta cepstral coefficients are derived from the LPC coefficients with 0.95 preemphasis. The Hamming window width is 20 msec with 7 msec frame shift. The normalized energy and corresponding delta energy are also included so that a feature vector of dimension 22 is used.

In order to see the discriminating capability of this feature vector on a phoneme with different tones, the feature vectors of all the frames belonging to a HMM state of the Final model for vowel 'u' with tones 2 and 3 respectively, after performing the principal component anlysis[8], are plotted in Fig. 4. Clearly two clusters are formed by their tones in the figure. Note that in these features the pitch or excitation information is NOT used at all. In other words, it is possible to classify the tones using the feature vectors here without pitch or excitation information.

6. THE RESULT OF EXPERIMENTS

The recognition results are listed in Table 3, in which 5 utterances for each syllable were used as training data and

1 utterance as testing data. The top 5 recognition rates for the 1302 syllables are listed in the first row, while the recognition rates for 408 base syllables and tones respectively in the same experiment are listed in rows 2 and 3. For comparison, the results of a similar experiment [9] using the traditional structure in Fig.3(a) on the same speech database for the 1302 syllable recognition is listed in the last row, and the improvements are obvious. Note that without the pitch or excitation information, the tone recognition rate can already achieve 94.85%. An analysis on the errors indicates that about 54% and 35% of the recognition errors are due to the tone recognition errors only and the base syllable recognition errors only respectively. It is therefore believed when the pitch or excitation information is included, further improvements can be achieved. Table 4 listed the seperate recognition rate for each tone, and apparently the recognition of the third tone was the worst. This is similar to the tone recognition results previously obtained based on the pitch contour. As for the base syllable recognition, it was found that most of the errors still occur in the confusion sets just as the conventional recognition approaches.

7. CONCLUSION

For Mandarin speech recognition, the approach proposed suggests the base syllable and the tone can be recognized simutaneously. The increased computation can be solved by careful design of the search algorithm and better arrangement of units. This approach has also the potential to be extended to continuous Mandarin speech recognition, athough the preliminary experiments demonstrated here in this paper are based on isolated syllables only.

8. ACKNOWLEDGEMENTS

The authors would like to thank Dr. S.C. Lu, Director of Telecommunication Laboratories, Dr. C.H. Lee, Dr. F.K. Soong of AT&T Bell Laboratories, and Dr. B.S. Jeng, Mr. E.F. Huang and Mr. C.S. Liu for their invaluable advice and timely encouragement.

References

[1] L.S. Lee et al. A Real-Time Mandarin Dictation Machine For Chinese Language with Unlimited Texts and Very Large Vocabulary *Proc. ICASSP-90,* Albuquerque, New Mexico, pp.65-68, Apr 1990.

[2] Harald Singer et al. Pitch Dependent Phone Modelling For HMM Based Speech Recognition *Proc. ICASSP-92,* San Francisco, California, pp.I273-I276, Mar 1992.

[3] C.H. Lin et al. An Initial Study On Mandarin Syllable Recognition Based On Sub-syllable Units *Proc. ICCPCOL-91,* Taipei, Taiwan, pp.302-306, Aug. 1991.

[4] Hsiao-Wuen Hon Vocabulary-Independent Speech Recognition: The VOCIND System, *Doctoral Thesis, School of Computer Science, Carnegie Mellon University* Pittsburgh, PA, Mar. 1992

[5] C.H. Lee, et al. Acoustic Modeling of Subword Units for Speech Recognition, *Computer Speech and Language,* Vol.4, pp.127-165, Apr. 1990.

[6] L.R. Rabiner et al. An Introduction To Hidden Markov Models. *IEEE ASSP Magazine,,* Vol.3, No.1, pp.4-16, Jan 1986.

[7] F.K. Soong et al. A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition. *Proc. ICASSP-90,* Albuquerque, USA, pp.709-712, Apr. 1990.

[8] R.O. Duda et al. Pattern Classification and Scene Analysis, John Wiley & Sons, Inc., 1973

[9] Yumin Lee et al. Isollated Mandarin Syllable Recognition Specially Considering The Effect Of Tones. *Proc. ICCPCOL-92,* Florida, USA, Dec. 1992.

Category	Member
1	ㄢ
2	i, ia, ie, iai, iau, iou, ian, in, iang, ing
3	u, ua, uo, uai, uei, uan, uen, uang, ueng
4	iu, iue, iuan, iun, iung
5	a, ai, au, an, ang
6	e, ei, eh, en, eng, er
7	o, ou

Table 1: the 38 Finals in Mandarin

1	2	3	4	5	6	7	8	9	10	11
	j	ch	sh	r	tz	ts	s	g	k	h
12	13	14	15	16	17	18	19	20	21	22
ji	chi	shi	d	t	n	l	b	p	m	f

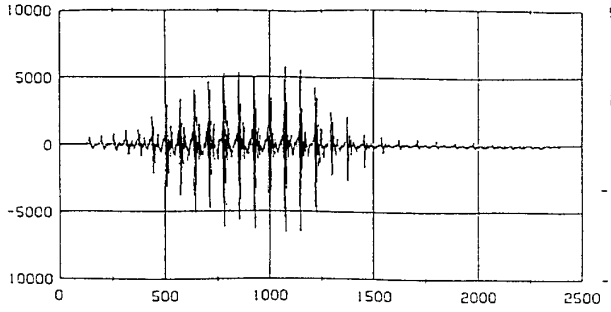
Table 2: the 22 Initials in Mandarin

	top 1	top 2	top 3	top 4	top 5
rate for 1302 syllables	92.0	96.7	98.0	99.0	99.3
rate for 408 base syllables	96.38	99.31	99.54	99.69	99.77
rate for the tones	94.85	97.54	98.69	99.30	99.61
rate for the previous structure	87.20	94.75	96.60	97.68	98.46

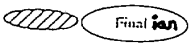
Table 3: The top-5 recognition rates for the experiments

	tone 1	tone 2	tone 3	tone 4
tone 1	97.329	0.009	0.000	0.1780
tone 2	0.0564	93.2331	0.0075	0.0038
tone 3	0.0270	0.0150	91.617	0.0419
tone 4	0.0247	0.0055	0.0028	96.703

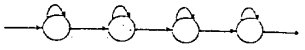
Table 4: The recognition rates of the tones



(a) waveform

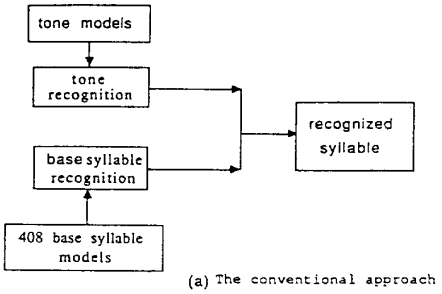


(b) sub-syllabic unit representation

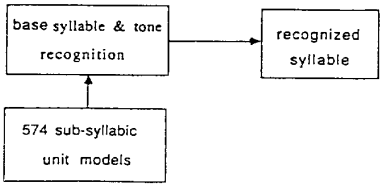


(c) HMM representation

Fig. 1 The syllable "ian3"

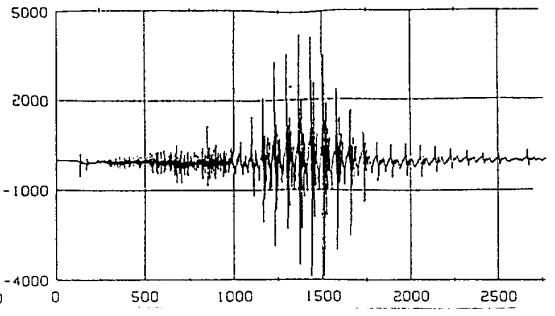


(a) The conventional approach



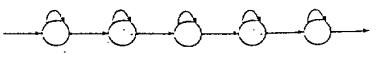
(b) The proposed approach

Fig. 3 The conventional and proposed approaches for Mandarin syllable recognition.



(a) waveform

(b) sub-syllabic unit representation



(c) HMM representation

Fig. 2 The syllable "jian3"

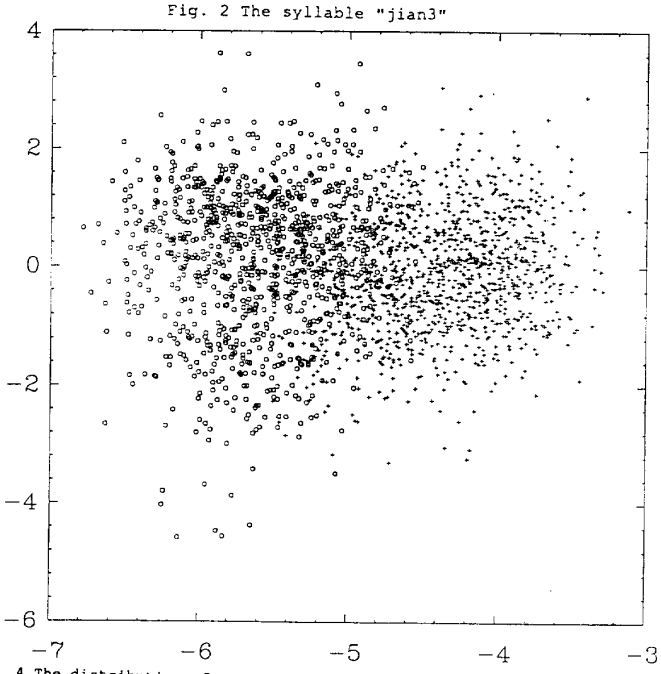


Fig. 4 The distribution of typical feature vectors for phoneme "u" with two different tones. "o" representing tone 2, while "+" representing tone 3.