Utterance Verification Using Prosodic Information for Mandarin Telephone Speech Keyword Spotting

Yeou-Jiunn Chen, Chung-Hsien Wu, and Gwo-Lang Yan
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.

ABSTRACT

In this paper, the prosodic information, a very special and important feature in Mandarin speech, is used for Mandarin telephone speech utterance verification. A two-stage strategy, with recognition followed by verification, is adopted. For keyword recognition, 59 context-independent subsyllables, i.e., 22 INITIAL's and 37 FINAL's in Mandarin speech, and one background/silence model, are used as the basic recognition units. For utterance verification, 12 anti-subsyllable HMM's, 175 context-dependent prosodic HMM's, and five anti-prosodic HMM's, are constructed. A keyword verification function combining phonetic-phase and prosodic-phase verification is investigated. Using a test set of 2400 conversational speech utterances from 20 speakers (12 males and 8 females), at 8.5% false rejection, the proposed verification method resulted in 17.8% false alarm rate. Furthermore, this method was able to correctly reject 90.4% of nonkeywords. Comparison with a baseline system without prosodic-phase verification shows that the prosodic information can benefit the verification performance.

1. INTRODUCTION

Recently, research into algorithms that are able to spot keywords has been focused on constructing hidden Markov model (HMM) based speaker independent keyword spotting systems using either subword models or whole word models [1]-[2]. In utterance verification, a large number of research efforts have been proposed. These approaches employ some type of a likelihood ratio distance to verify whether or not a given keyword exists within an input speech segment. Anti-keyword models were generally constructed to provide an anti-keyword scoring in computing likelihood ratio statistics. Rahim *et al.* [1] used the scores of antikeyword models and a general acoustic filler model for digit string utterance verification. Sakkar *et al.* [2] used a two-stage verification: subword-level verification followed by string-level verification.

Chinese is a tonal language in which the same phonetic syllable when pronounced in different tones gives quite distinct meanings. The five tones, four lexical tones and one neutral tone, in Mandarin Chinese have lexical meaning. Conventionally, there are 408 Mandarin base syllables, regardless of tones, which are composed of 21 INITIAL's and 37 FINAL's. In recent years, the most popular configuration of Mandarin Chinese speech recognition consists of two subrecognizers. One is tone recognizer and the other is phoneme or syllable recognizer[3]. Some of these efforts focused on the combination of phonetic and prosodic features. Prosodic information, such as pitch and spectral energy in fundamental frequency, plays an important role in Mandarin speech recognition. In this paper, it is highly desirable to include the expected or correct keywords in a higher rank. In addition, homonym words with different tone combinations need prosodic information for discrimination. Therefore, the prosodic information is first adopted for Mandarin speech utterance verification.

Fig. 1 shows the block diagram of the keyword spotting system. First, the phonetic and prosodic features are extracted. Hidden Markov Models (HMM's) with continuous observation densities are adopted to model the phonetic and prosodic features. In the first stage, Viterbi algorithm is employed to find the scores of the N best keyword candidates and their corresponding subsyllable boundaries. Subsyllable boundaries are then used to extract the FINAL part of Mandarin syllables which contains the prosodic information. In the second stage, the phonetic features of these FINAL parts are fed to their corresponding anti-subsyllable HMM's to give a phonetic verification score. Similarly, the prosodic features of these FINAL parts are fed to the prosodic HMM's and anti-prosodic HMM's to output a prosodic verification score. A keyword verification function combining phonetic-phase and prosodic-phase verification functions is investigated and used to reorder the ranks of the N best keyword candidates.

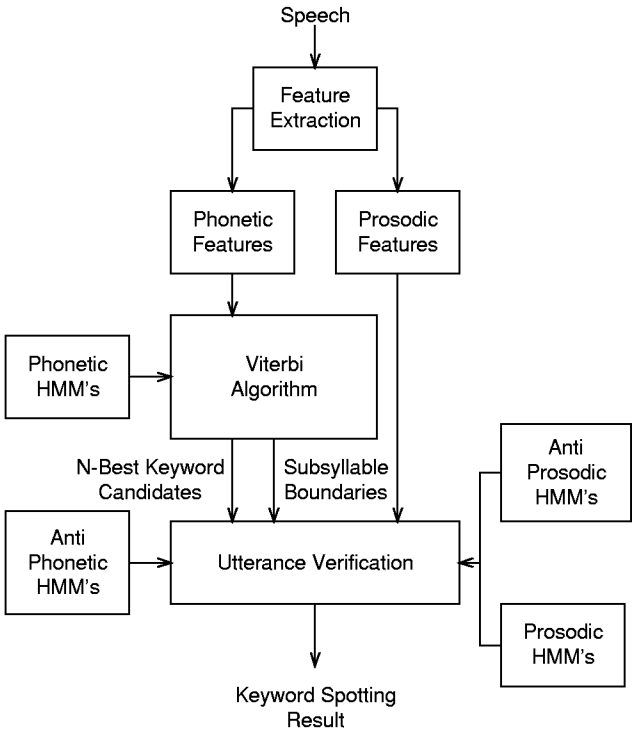

Fig. 1. The block diagram of the keyword spotting system.

2. FEATURE EXTRACTION

Two sets of features, phonetic and prosodic features, are used in this system. For phonetic features, 12 MFCC, 12 delta MFCC,

delta log energy, and delta delta log energy are used. In prosodic feature extraction for the *ith* analysis frame of the *jth* FINAL part, four parameters in the prosodic feature vector are defined as follows:

$$V_j^{(i)} = [v_j^{(i)}(1), v_j^{(i)}(2), v_j^{(i)}(3), v_j^{(i)}(4)] \quad (1)$$

where

$$v_j^{(i)}(1) = \begin{cases} P_j^{(i)} - \overline{P}_j^{(i)} & ,Pitch \quad period \neq 0 \\ r & ,Otherwise \end{cases} \quad (2)$$

$$v_j^{(i)}(2) = \Delta\left(P_j^{(i)} - \overline{P}_j^{(i)}\right) \quad (3)$$

$$v_j^{(i)}(3) = \begin{cases} \log\left(\dfrac{S_{j:k}^{(i)}}{\overline{S}_j^{(i)}}\right) & ,Pitch \quad period \neq 0 \\ -\log\left(\overline{S}_j^{(i)}\right) & ,Otherwise \end{cases} \quad (4)$$

$$v_j^{(i)}(4) = \begin{cases} \log\left(\dfrac{S_{j:\max}^{(i)}}{\overline{S}_j^{(i)}}\right) & ,Pitch \quad period \neq 0 \\ -\log\left(\overline{S}_j^{(i)}\right) & ,Otherwise \end{cases} \quad (5)$$

where P is the logarithmic value of the pitch period of a FINAL part, $\overline{P}$ is the average logarithmic value of the pitch period, and r is a small random value.S_k is the spectral energy at the fundamental frequency, $\overline{S}$ is the average spectral energy, and S_{max} is the spectral energy at the first formant frequency.

3. CONSTRUCTION OF ANTI-KEYWORD MODELS

In the keyword recognizer, 22 INITIAL (including one null INITIAL) and 37 FINAL context-independent HMM's are constructed. Each INITIAL HMM consists of 3 states and each FINAL HMM consists of 5 states, each with 10 Gaussian mixture densities. In general, for every subsyllable model in the model set, a corresponding anti-subsyllable model is trained specifically for the verification task. However, for every subsyllable model, the corresponding anti-subsyllable model should be trained using a wide range of sounds. For example, to train the anti-subsyllable /a/, all the training data of the other 58 subsyllables should be used. This renders the anti-subsyllable very general and ineffective. In this work, the INITIAL's and FINAL's in Mandarin speech are separately treated. The 22 INITIAL's and 37 FINAL's are clustered into 3 groups and 9 groups, respectively. The K-means clustering algorithm is used to cluster subsyllables based on minimizing the overall intersubsyllable group distance. For each subsyllable group, all speech segments corresponding to sounds that are not modeled by any of the subsyllables in that subsyllable group are used to train an anti-subsyllable HMM. In total, there are 3 INITIAL anti-subsyllable HMM's and 9 FINAL anti-subsyllable HMM's. For INITIAL and FINAL anti-subsyllable HMM's, 8 and 16 Gaussian nodes are used, respectively.

Since lexical tone is the most important feature of the prosodic information, prosodic model should be constructed based on lexical tone behavior. Earlier investigations showed that the tone behavior is very complicated in continuous Mandarin speech, although there are only 5 different tones in Mandarin. Therefore, we assume every kind of possible tone combination needs a context-dependent model, then a total of 175 prosodic HMM's will be needed. For the construction of anti-prosodic models, the training data are divided into five groups according to the five lexical tones. Five anti-prosodic HMM's, each corresponding to one context-independent lexical tone, are constructed to enhance the discriminability among prosodic HMM's. An anti-prosodic HMM can be considered as a lexical-tone-specific model. It is based on similar concept to the cohorts in speaker verification [5]. An anti-prosodic HMM is generally trained on the training data with all lexical tones but that with the corresponding lexical tone. Each prosodic HMM has 4 state and 6 mixtures.

4. TWO-STAGE RECOGNITION

4.1 Keyword Recognition

In this system, a two stage recognition scheme is used. In the first stage, Viterbi algorithm is employed to find the most likely keyword W_k, where

$$W_k = \arg\max_j L(O \mid W_j) \quad (6)$$

and $L(O \mid W_j)$ is the likelihood of the observation sequence O given word W_j. In the context of subsyllable recognition, W_k is a concatenation of subsyllable units that can be written as

$$W_k = s_1^{(k)} s_2^{(k)} ... s_{2N}^{(k)} \quad (7)$$

where 2N is the number of subsyllables. For detailed representation, W_k can be expressed as a concatenation of INITIAL and FINAL parts described as follows.

$$W_k = i_1^{(k)} f_1^{(k)} i_2^{(k)} f_2^{(k)} ... i_N^{(k)} f_N^{(k)} \quad (8)$$

where the subsyllable string $i_1^{(k)} f_1^{(k)} i_2^{(k)} f_2^{(k)} ... i_N^{(k)} f_N^{(k)}$ is the subsyllable lexical representation of keyword W_k.

4.2 Utterance Verification

Utterance verification can be treated as the problem of statistical hypothesis testing. Two types of errors can occur: false rejection (Type I) and false acceptance or false alarms (Type II) errors. In this verification process, a two-phase verification scheme is employed.

4.2.1 Phonetic-Phase Verification

Given a subsyllable $s_n^{(k)}$, the normalized confidence measure is defined as

$$LR(O_{t_{n-1}}^{t_n}; s_n^{(k)}) = \frac{1}{T_n^{(k)}} \log[L(O_{t_{n-1}}^{t_n} \mid s_n^{(k)})] - \frac{1}{T_n^{(k)}} \log[L(O_{t_{n-1}}^{t_n} \mid \overline{s}_n^{(k)})] \quad (9)$$

where $\overline{s}_n^{(k)}$ is the anti-subsyllable model of $s_n^{(k)}$, $T_n^{(k)}$ is the number of frames allocated for subsyllable $s_n^{(k)}$. For an N-syllable (or 2N subsyllable) string $s_1^{(k)} s_2^{(k)} ... s_{2N}^{(k)}$ corresponding to the most likely keyword W_k, the whole word phonetic verification function is defined as follows:

$$D(O;W_k)=\log[\frac{1}{2N}\sum_{n=1}^{2N}\alpha_n^{(k)}\exp[-\eta\cdot LR(O_{t_{n-1}}^{t_n};s_n^{(k)})]]^{\frac{1}{\eta}} \quad (10)$$

where η is a positive constant, and $\alpha_n^{(k)}$ is a subsyllable weighting empirically chosen as

$$\alpha_n^{(k)}=\begin{cases}0.75 & if\ s_n^{(k)}is\ an\ INITIAL\\ 1.0 & if\ s_n^{(k)}is\ a\ FINAL\end{cases} \quad (11)$$

The subsyllable weight for INITIAL is chosen smaller than that for FINAL. This is because that the INITIAL part in Mandarin syllable occupies just a short duration compared to the FINAL part and the recognition accuracy or reliability for INITIAL is lower than that for FINAL part.

4.2.2 Prosodic-Phase Verification

In the prosodic-phase verification, the corresponding lexical tone string T_{W_k} with respect to the keyword W_k is obtained using the sandhi rules [5] and written as

$$T_{W_k}=t_1^{(k)}t_2^{(k)}...t_N^{(k)} \quad (12)$$

Since most of the prosodic information is embedded in the FINAL part, the prosodic verification is only performed on the FINAL part. Given the prosodic feature vectors of a FINAL part corresponding to the lexical tone t_j, the prosodic confidence measure is written as

$$CM(P_j;t_j)=\log[G(p_{t_j};t_j)]-\log[G(p_{\bar{t}_j};\bar{t}_j)] \quad (13)$$

where $P_j=[p_{t_j},p_{\bar{t}_j}]$ represents the verification feature vector, and G(•) is a Gaussian distribution of the verification feature vector. The parameters of the feature vectors p_{t_j} and $p_{\bar{t}_j}$ are obtained by processing the prosodic feature vectors of the segmented FINAL part through prosodic model t_j and anti-prosodic model $\bar{t}_j$, respectively. Therefore, p_{t_j} forms a 21-dimensional vector consisting of the following:

- Coefficients representing the contour of the prosodic features of the segmented FINAL part. To be more specific, each prosodic feature in V_j is represented by a smooth curve formed by orthonormal expansion with discrete Legendre polynomials [6]. The number coefficients used in this polynomial is up to the third order. The zero-th order coefficient represents the mean of the prosodic feature contour and the other three coefficients represent its shape. Given a 4-dimensional prosodic feature vector, the number of parameters is 16.
- Four parameters representing the state durations in number of frames normalized by the total frame duration of segmented FINAL part.
- The prosodic HMM likelihood $L(V_j / t_j)$.

Similarly, $p_{\bar{t}_j}$ is formed by processing V_j using the anti-prosodic model $\bar{t}_j$ and computing the corresponding 21 parameters. For the whole word verification, the verification function can be decomposed into a series of FINAL part verification functions. Assuming independence, the whole word prosodic verification function is defined as follows:

$$D(P;G_{W_k})=\log[\frac{1}{N}\sum_{j=1}^{N}\exp[-\kappa\cdot CM(P_j;t_j)]]^{\frac{1}{\kappa}} \quad (14)$$

where κ is a positive constant. The outputs of the prosodic and phonetic verification functions are then combined as follows.

$$D(O,P;W_k)=\beta D(O;W_k)+(1-\beta)D(P;G_{W_k}) \quad (15)$$

where β is a weighting. Finally, the keyword rejection/ acceptance decision is made by comparing $D(O, P, W_k)$ with a predefined threshold.

5. EXPERIMENTAL RESULTS

In order to assess the keyword spotting system performance, a query system to access telephone number information for a person in the directory has been implemented. In our system, 200 faculty names in National Cheng Kung University were selected as the keywords. A continuous telephone-speech database was employed to train the system. The database is part of the MAT (Mandarin Speech Across Taiwan) speech database and is composed of short spontaneous speech, numbers, syllables, words, and sentences. The total number of files is 12,386. This database was pronounced by 295 speakers (192 males, 103 females). All speech data were recorded via public telephone lines in 8 KHz using a Dialogic D/41D telephone card and a 16-bit Soundblaster card. We also recorded 2400 utterances for testing spoken by a different group of 20 speakers (12 males, 8 females) responding to requests for a person's name in our vocabulary. All test utterances were assigned to one of the following categories. The percentage for each category in the testing database is also listed below.

- In-vocabulary names, spoken in isolation (K): 34%
- In-vocabulary names, embedded before a phrase (K+N): 17%
- In-vocabulary names, embedded after a phrase (N+K): 20%
- In-vocabulary names, embedded in a sentence (N+K+N): 19%
- Speech with no in-vocabulary names (N): 10%

In this database, only 90% of the users provided an isolated name or a name embedded in a phrase or a sentence. 10% of user responses included no names at all. All of these responses need to be rejected. In our experiments, two types of errors, namely, false rejection and false alarm are used to evaluate the system performance. Several experiments were conducted to determine factors necessary to achieve the best performance.

5.1 Effect of the weighting parameter β

In the first experiment, the variation of the total Type I and Type II errors as a function of the weighting parameter β was evaluated as a function of the weighting parameter β. Fig. 2 shows that the combination of phonetic and prosodic information can improve the keyword spotting rate for $0.25\le\beta\le0.50$. When β=0.375, the system can achieve the best recognition performance.

5.2 Experiments for the effects of prosodic information

Two experiments were conducted to test the performance of the proposed verification method. In order to benchmark the verification performance, a baseline system that employs only

phonetic-phase verification was established. Fig. 3 shows the verification performance of both the proposed and the baseline verification methods. It is clear that the proposed method outperforms the baseline system. For instance, at 8.5% false rejection, the proposed system resulted in 17.8% false alarm rate. This is compared with the baseline system, which results in 22.4% false alarm rate at the same 8.5% false rejection rate.

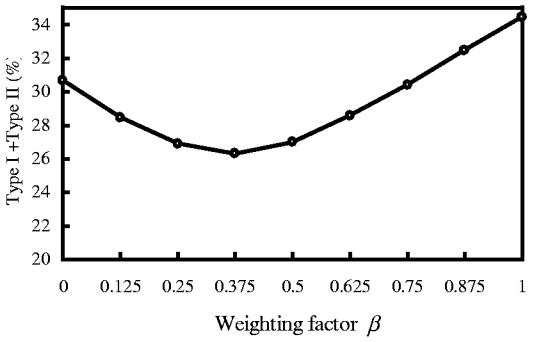

Fig.2 Combined Type I and Type II errors as a function of the weighting factor β

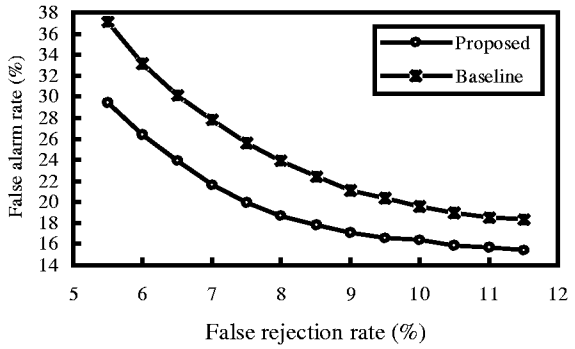

Fig. 3 Utterance verification performance comparison of the proposed and the baseline methods

5.3 Experiments for the locations of keywords

The speech utterances divided into five categories were experimented upon to evaluate the effects of the locations of the keywords in an utterance. The experimental results are listed in Table I. The false alarm rate for the first category (K) and the third category (N+K) was 11.4% and 19.8%, respectively, at a false rejection rate of 8.5%. They were better than that for other categories. This is because the FINAL part in these two categories can be easily detected. Consequently, we can obtain better performance in these two categories. It is reasonable that the first category (K) and the second category achieves the lowest and the highest false alarm rates, respectively. At 8.5% false rejection, the average false alarm rates were 11.4% and 24.6%, for isolated and embedded keywords, respectively. Furthermore, for the fifth category (N), the proposed method was able to correctly reject 90.4% of nonkeywords.

6. CONCLUSIONS

In this paper, we have demonstrated some achievements in continuous Mandarin speech keyword recognition and verification. In this system, 59 context-independent subsyllables are used as the basic recognition units. A two-stage strategy, with recognition followed by verification, is adopted. For utterance verification, 12 anti-subsyllable HMM's, 175 context-dependent prosodic HMM's and five anti-prosodic HMM's, are constructed. A keyword verification function combining phonetic-phase and prosodic-phase verification is investigated. Experimental results shows that utterance verification with prosodic information outperforms the baseline system without prosodic information.

Table I False alarm rates (%) for five speech utterance categories, at a false rejection rate of 8.5%

	Speech utterance category				
	Isolated	Embedded			No Keyword
	K (34%)	K+N (17%)	N+K (20%)	N+K+N (19%)	N (10%)
False alarms (%)	11.4	24.6	19.8	21.2	9.6
Average (%)	11.4	21.7			9.6

7. ACKNOWLEDGMENTS

The author would like to thank the National Science Council, Republic of China, for financial support of this work, under Contract No. NSC86-2622-E-006-003.

8. REFERENCES

[1] M.G. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance for connected digits recognition," *IEEE Transactions on Speech and Audio Processing*. VOL. 5, NO. 3, pp.266-277, May 1997.

[2] R.A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Transactions on Speech and Audio Processing*. VOL. 4, NO. 6, pp.420-429, November 1996.

[3] L-S. Lee, C-Y. Tseng, H-Y. Gu, F-H. Liu, C-H. Chang, Y-H. Lin, Y-M. Lee, S-L. Tu, S-H. Hsieh, and C-H Chen, "Golden Mandarin (I) - A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary," *IEEE Trans. Speech and Audio Processing*, 1(2), pp.158-179, 1993.

[4] A.E. Rosenberg, C.H. Lee, B.H. Juang, and F.K. Soong, "The use of cohort normalized scores for speaker verification," in Proc. 1992 Int. Conf. Spoken Language Processing, 1992, pp.599-602.

[5] L.S. Lee, C.Y. Tseng and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system", *IEEE Trans. Acoustic Speech, and Signal Processing*, Vol. 37, No. 9, pp. 1309-1319, September 1989.

[6] S.H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. Commun.*, 38, pp. 1317-1320, 1990.