Use of Prosodic Information for Mandarin Word Verification

CHUNG-HSIEN WU, YEOU-JIUNN CHEN, AND GWO-LANG YAN

Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan, R.O.C.

(Received April 22, 1998; Accepted January 5, 1999)

ABSTRACT

In this paper, prosodic information, a very special and important feature of Mandarin speech, is used for Mandarin word verification. A two-stage strategy, with recognition followed by verification, is adopted in word recognition of telephone speech. For word recognition, 94 right context-dependent INITIALs, 37 context-independent FINALs in Mandarin speech and one silence model are used as the basic recognition units. For word verification, 15 anti-subsyllable Hidden Markov Models (HMMs), 175 context-dependent prosodic HMM’s, and five anti-prosodic HMM’s, are constructed. A word verification function combining phonemic-phase and prosodic-phase verification is investigated. Using a test set of 2200 word utterances from 22 speakers (14 males and 8 females), at 1.9% false rejection, the proposed verification method obtained a 9.0% false alarm rate. Comparison with a baseline system without prosodic-phase verification shows that, using prosodic information, the proposed system yields a false alarm rate reduction of 22.5% and a false rejection rate reduction of 20.8%, respectively.

Key Words: Mandarin speech, prosodic information, word verification, false alarm, false rejection

I. Introduction

Rapid advances in speech recognition technology have been achieved in recent years. This has enabled speech recognition systems to migrate from the laboratory to actual applications (Cole et al., 1995; Huang and Lee, 1993). However, the performance in speech recognition via a telephone channel is often degraded due to unknown adverse conditions, such as ambient noise, channel distortion, and different variations of telephone handsets. Cepstrum mean subtraction or signal bias removal is generally adopted to deal with this problem (Rahim et al., 1996; Rahim and Juang, 1996). On the other hand, for telephony-based applications, users may respond with speech that does not include any of the vocabulary words. The recognizer should be endowed with the capability of rejecting out-of-vocabulary speech.

Word recognition and verification are two related research areas. Good recognition capability necessarily implies good rejection performance. For word verification, a large number of approaches have been proposed (Rahim et al., 1997; Sukkar and Lee, 1996). These approaches employ some type of a likelihood ratio distance to verify whether or not a given word exists within an input speech segment. Anti-word models were generally constructed to provide an anti-word scoring in computing likelihood ratio statistics. Rahim et al. (1997) used the scores of anti-keyword models and a general acoustic filler model for digit string utterance verification. Sukkar and Lee (1996) used two-stage verification: subword-level verification followed by string-level verification.

Chinese is a tonal language, in which the same phonemic syllable when pronounced using different tones has quite different meanings. The five tones, four lexical tones and one neutral tone, of Mandarin Chinese have lexical meaning. Conventionally, there are 408 Mandarin base syllables, regardless of tones, which are composed of 21 INITIALs and 37 FINALs. In recent years, the most popular configuration of Mandarin Chinese speech recognition has consisted of two subrecognizers. One is a tone recognizer and the other is a phoneme or syllable recognizer (Lee et al., 1993). Some of these efforts have focused on a combination of phonemic and prosodic features. The phonemic features, including cepstral coefficients, and log energy, are generally used for phoneme or syllable recognition. Prosodic information, such as pitch and spectral energy at the fundamental frequency, plays an important role.
Prosodic Information for Word Verification

in Mandarin speech recognition. In our approach, it is highly desirable to include the expected or correct words in a higher rank. In addition, homonym words with different tone combinations need prosodic information for further discrimination. Therefore, prosodic information is adopted in Mandarin word verification.

Figure 1 shows the block diagram of the word recognition system.

II. Feature Extraction

Two sets of features, phonemic and prosodic features, are used in our system. For a phonemic feature, a 26-dimension feature vector is extracted. The 12 Mel-Frequency Cepstrum Coefficient (MFCC), 12 delta MFCC, delta log energy, and delta delta log energy are adopted. A 20-channel filter bank design, in which each filter has a triangle bandpass frequency response with bandwidth and spacing determined by a constant mel frequency interval, is used to extract MFCC. Using this 20-channel filter bank, the spectral energy at the fundamental frequency and the spectral energy at the formant frequency can be estimated.

From the viewpoint of a language, prosodic information pertains to strings of segments, from the smallest units, morae, through syllables, words and phrases, to utterances and paragraphs. It encompasses syllable tones, word accents and stress, pause, and intonation. Prosodic variations in human speech result from different speaking styles or different emphasis at different frequencies of the same utterance. These prosodic features are always encoded in duration, intensity, pitch contour, and spectral energy at the fundamental frequency. A suitable representation of prosodic information is proposed to adequately model the different dynamic articulatory characteristics of that sequence when it is produced in contexts with different meanings.

Many prosodic parameters, which can be used for speech recognition, are listed below:

1. duration,
2. pitch period,
3. waveform envelope amplitude,
4. spectral energy at the fundamental frequency,
5. spectral energy at the formant frequency,
6. harmonic ratio,
7. degree of spectral tilt.

However, to avoid computation complexity and based on results of some preliminary experiments, only four features are adopted in the prosodic feature vector in our approach.

In prosodic feature extraction for the $i$th analysis frame of the $j$th FINAL part, four parameters in the prosodic feature vector are defined as follows:

$$V_j^o = [v_j^o(1), v_j^o(2), v_j^o(3), v_j^o(4)]$$  \hspace{1cm} (1)

(1) Normalized pitch period
\[ v_j^{(1)} = \begin{cases} \frac{P_j - \overline{P}_j}{r}, & \text{Pitch period \( \neq 0 \)} \\ r, & \text{Otherwise} \end{cases} \]  

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.

(2) Delta logarithmic pitch period

\[ v_j^{(2)} = \Delta v_j^{(1)} = \mu \sum_{k=0}^{N-1} k v_j^{(1)}, \]  

where \( N \) is the window size and \( \mu \) is a normalized constant. In our system, \( N \) is chosen as 2, and \( \mu \) is set to 0.0375.

(3) Spectral energy at the fundamental frequency

\[ v_j^{(3)} = \begin{cases} \log \left( \frac{S_j^{(F)}}{S_j^{(0)}} \right), & \text{Pitch period \( \neq 0 \)} \\ -\log (\overline{S}_j^{(0)}), & \text{Otherwise} \end{cases} \]  

where \( S_j^{(F)} \) is the spectral energy at the fundamental frequency, and \( \overline{S}_j^{(0)} \) is the average spectral energy.

(4) Spectral energy at the formant frequency

\[ v_j^{(4)} = \begin{cases} \log \left( \frac{S_j^{(\text{max})}}{S_j^{(0)}} \right), & \text{Pitch period \( \neq 0 \)} \\ -\log (\overline{S}_j^{(0)}), & \text{Otherwise} \end{cases} \]  

where \( S_j^{(\text{max})} \) is the spectral energy at the first formant frequency.

III. Construction of Anti-Word Models

In the word recognizer, 94 right context-dependent INITIALs, 37 context-independent FINALs and one silence model 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 130 subsyllables should be used. This renders the anti-subsyllable very general and ineffective. In our approach, the INITIALs and FINALs in Mandarin speech are treated separately. The 94 right context-dependent INITIALs and 37 context-independent FINALs are clustered into 9 groups and 6 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 9 INITIAL anti-subsyllable HMMs and 6 FINAL anti-subsyllable HMMs. For INITIAL and FINAL anti-subsyllable HMMs, 8 and 16 Gaussian nodes are used, respectively.

Since the lexical tone is the most important feature of the prosodic information, the 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 that every possible different tone combination needs a context-dependent model; therefore, a total of 175 prosodic HMMs are needed and listed as follows:

1. At the beginning of a sentence: 4×5 models.
2. In the middle of a sentence: 5×5×5 models.
3. At the end of a sentence: 5×5 models.
4. Isolated syllables: 5 models.

For the construction of anti-prosodic models, the training data are divided into five groups according to their corresponding lexical tones. Five anti-prosodic HMMs, each corresponding to one context-independent lexical tone, are constructed to enhance discriminiability among prosodic HMMs. An anti-prosodic HMM can be considered a lexical-tone-specific model. It is based on a concept similar to others in speaker verification (Rosenberg et al., 1992). An anti-prosodic HMM is generally trained on the training data with all lexical tones except the corresponding lexical tone. Each prosodic HMM has 4 states and 6 mixtures.
IV. Two-Stage Recognition

1. Word Recognition

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

$$W_k = \arg \max_j L(O|W_j),$$

and $L(O|W_j)$ 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

$$W_k = s_{1}^{(k)} s_{2}^{(k)} \cdots s_{2N}^{(k)},$$

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

$$W_k = i_1^{(k)} i_2^{(k)} \cdots i_N^{(k)} f_1^{(k)} f_2^{(k)} \cdots f_N^{(k)},$$

where the subsyllable string $i_1^{(k)} i_2^{(k)} \cdots i_N^{(k)}$ is the subsyllable lexical representation of word $W_k$.

2. Word Verification

Word verification can be treated as a problem of statistical hypothesis testing. Two types of errors can occur: false rejection (Type I) and false acceptance or false alarm (Type II) errors (Fukunaga, 1990). In this verification process, a two-phase verification scheme is employed.

A. Phonemic-Phase Verification

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

$$LR(I_n^{(t_n-1)}; s_n^{(k)}) = \frac{1}{F_n^{(k)}} \log L(O_{t_n-1}^{I_n} | s_n^{(k)}) - \frac{1}{F_n^{(k)}} \log L(O_{t_n-1}^{I_n} | \overline{s}_n^{(k)}),$$

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

$$D(O; W_k) = \log \left[ \frac{1}{2N} \sum_{n=1}^{2N} \alpha_n^{(k)} \exp \left( -\eta \cdot LR(I_n^{(t_n-1)}; s_n^{(k)}) \right) \right]^{\frac{1}{2}},$$

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

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

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

B. Prosodic-Phase Verification

In prosodic-phase verification, the corresponding lexical tone string $T_{W_k}$ with respect to the word $W_k$ is obtained and can be written as

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

Since most of the prosodic information is embedded in the FINAL part, 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

$$CM(P_j; t_j) = \log [G(p_{j,t_j})] - \log [G(p_{\overline{t}_j})],$$

where $P_j = [p_{j,t_j}; p_{j,\overline{t}_j}]$ represents the verification feature vector, and $G(*)$ is a Gaussian distribution of the verification feature vector. The parameters of the feature vectors $p_{j,t_j}$ and $p_{j,\overline{t}_j}$ are obtained by processing the prosodic feature vectors of the segmented FINAL part using prosodic model $t_j$ and anti-prosodic model $\overline{t}_j$, respectively. Therefore, $p_{j,t_j}$ 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 the discrete Legendre polynomial (Chen and Wang, 1990). The number coefficients used in this polynomial can be as high as 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.

(2) Four parameters representing the state durations in terms of the number of frames normalized by the total frame duration of the segmented FINAL part.

(3) The prosodic HMM likelihood $L(V_j|t_j)$.

Similarly, $p_{T_j}$ is formed by processing $V_j$ using the anti-prosodic model $T_j$ and by 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:

$$D(P;G_{w_k}) = \log \left[ \frac{1}{N} \exp \left( -\kappa \cdot CM(V_j; t_j) \right) \right]$$

where $\kappa$ is a positive constant. The outputs of the prosodic and phonemic verification functions are then combined as follows:

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

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

V. Experimental Results

In order to assess the word recognition system performance, a query system to access telephone number information for a person in a directory was implemented. In our system, 1200 names of faculty and students in National Cheng Kung University were selected as vocabulary words. This vocabulary contains 328 two-syllable names, 627 three-syllable names, and 245 four-syllable names. Most of the two- and four-syllable names were student names in order to balance the number of names with different lengths in the vocabulary. 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 provided by 295 speakers (192 males and 103 females). All the speech data were recorded via public telephone lines at 8 KHz using a Dialogic D/41D telephone card and a 16-bit Soundblaster card. We also recorded 2200 randomly chosen name utterances for testing, spoken by a different group of 22 speakers (14 males and 8 females). All the test utterances were assigned to one of the following categories. The percentage for each category in the testing database is also listed below:

1. in-vocabulary names with two syllables (InC1): 22%;
2. in-vocabulary names with three syllables (InC2): 41%;
3. in-vocabulary names with four syllables (InC3): 15%;
4. out-of-vocabulary names with two syllables (OutC1): 6%;
5. out-of-vocabulary names with three syllables (OutC2): 12%;
6. out-of-vocabulary names with four syllables (OutC3): 4%.

In this database, only 78% of the users provided an in-vocabulary name. 22% of the user responses had no names in the vocabulary. These responses had to be rejected. In our experiments, two types of errors, namely, false rejection and false alarm errors, were used to evaluate the system performance. Several experiments were conducted to determine factors necessary to achieve the best performance.

1. Effect of the Weighting Parameter $\beta$

In the first experiment, the variation of the total Type I and Type II errors was evaluated as a function of the weighting parameter $\beta$. Figure 2 shows that the combination of phonemic and prosodic information can improve the word recognition rate for $0.125 \leq \beta \leq 0.50$. When $\beta=0.375$, the system can achieve the best recognition performance.

2. Experiments on the Effects of Prosodic Information

The 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 phonemic-phase verification was established. Figure 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, under 1.9% false rejection, use of the proposed system resulted in a 9.0% false alarm rate. In comparison, use of the baseline system resulted in an 11.6% false alarm rate and a 3.1% false rejection rate. This implies that the proposed system using prosodic-phase verification yields a false alarm rate reduction of 22.5% and a false rejection rate reduction of 20.8%, respectively. When false rejection is equal to zero, the false alarm rate is the same as the word recognition error rate. In these experiments, the word error rates without rejection were 12.6% and 14.9% for systems with and without prosodic information, respectively. From Fig. 3, it is obvious that the word recognition error can be reduced by using prosodic information.

3. Experiments on the Lengths of Words

The speech utterances divided into six categories were experimented upon to evaluate the effects of the lengths of the words in an utterance. The experimental results are listed in Table 1. According to Table 1(a), it is obvious that words with fewer syllables achieved better improvement in the false alarm rate. That is because the grammar tree can give satisfactory performance for words with more syllables, and the improvement in the recognition rate becomes insignificant compared to the improvement for words with fewer syllables.

The false rejection experiments results listed in Table 1(b) reveal that the false rejection rate for two-syllable names is the highest. That is because short utterances contain less information. On the other hand, four-syllable names may have the lowest false rejection rate. On average, the false rejection rate can be improved from 2.4% to 1.9% by using prosodic information.

### Table 1. (a) False Alarm Rates and (b) False Rejection Rates (%) for Six Speech Utterance Categories

<table>
<thead>
<tr>
<th>Speech utterance category</th>
<th>OutC1(6%)</th>
<th>OutC2(12%)</th>
<th>OutC3(4%)</th>
<th>InC1+InC2+InC3(78%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False alarms (%) without prosodic-phase verification</td>
<td>3.0</td>
<td>2.3</td>
<td>1.9</td>
<td>2.4</td>
</tr>
<tr>
<td>False alarms (%) with prosodic-phase verification</td>
<td>2.6</td>
<td>1.8</td>
<td>1.7</td>
<td>1.9</td>
</tr>
</tbody>
</table>

VI. Conclusions

In this paper, we have described some advances in continuous Mandarin speech word recognition and verification. In this system, 94 right context-dependent INITIALs, 37 context-independent FINALs and one silence model are used as the basic recognition units. A two-stage strategy, with recognition followed by verification, is adopted for word recognition of telephone speech. For utterance verification, 15 anti-subsyllable HMMs, 175 context-dependent prosodic HMMs and five anti-prosodic HMM’s, are constructed. A word verification function combining phonemic-phase and prosodic-phase verification has been investigated. Experimental results show that utterance verification with prosodic information outperforms the baseline system without prosodic information.

Acknowledgment

The author would like to thank the National Science Council of the Republic of China for financial support of this work under contract No. NSC 86-2622-E-006-067.
References


應用語音訊息於詞彙驗證

吳宗憲  陳有圳  顏國郎

國立成功大學資訊工程研究所

摘要

在語音辨識中，語音訊息是一個特殊且重要的參數，因此在本論文中，我們應用語音訊息於詞彙驗證。此系統中採用兩階段的辨識方法。在詞彙辨識方面，我們建立94前後文相關篇，37前後文不相關篇及一個背景噪音的隱藏式馬可夫模組當作我們的辨識基本單位，另外建立15個反次音節隱藏式馬可夫模組以作場驗證用。而在詞彙音節驗證方面，157個前後文相關的音節隱藏式馬可夫模組和5個音節隱藏式馬可夫模組當作詞彙驗證的單位，最後利用一個驗證函數將語音和語音的資訊整合。在實驗方面，由22個人透過電話線取得2200個詞彙話音資料當作測試音檔，在19%的錯誤拒絕率的時候，我們提出的方法只有9%的錯誤產生率。在與沒有加入語音訊息的系統比較中，我們發現可以將誤差產生率22.5%，而錯誤拒絕率可以減少20.8%。