ACOUSTIC MODEL COMPARISONS FOR AN EMBEDDED PHONEME-BASED MANDARIN NAME DIALING SYSTEM

Xuan ZHU, Rui WANG, Yi-Ning CHEN, Jia LIU, Run-Sheng LIU
Department of Electronics Engineering, Tsinghua University, Beijing
zhuxuan98@mails.tsinghua.edu.cn

ABSTRACT
In this paper we put our attention on gaining a set of effective acoustic model bestowed in an embedded phoneme-base mandarin name dialing system. Different sets of sub-word units are tested in the diverse vocabulary lists. The emission probability varied with the amount of mixtures and the type of covariance matrices is adopted. The speech feature with various elements is employed. Ultimately plentiful experimental results are enumerated.

1. INTRODUCTION
In recent years great improvements have been achieved by applying speech recognition technologies in the practical systems, especially in some embedded systems, such as mobile phones, personal digital assistants and so on. But the SD (Speaker-Dependent) name dialing application in current systems is too inconvenient to use for its full-training and small vocabulary limitation. Therefore, a PBSI (phoneme-based Speaker-Independent) name dialing function is obviously more valuable in the embedded systems.

Despite that template matching algorithms and DHMM (Discrete Hidden Markov Model) both have achieved good performance in the their corresponding special applications, but for the phoneme-based speaker-independent (PBSI) speech recognition systems they are too rough to undertake the task. Thus, CDHMM (Continuous Density Hidden Markov Model), which is usually implemented in the LVCSR (Large-Vocabulary Continuous Speech Recognition), seems to be the best choice to obtain the high recognition accuracy. On the contrary, the limited hardware resources among a speech ASIC (Application Specific Integrated Circuit) apparently makes the recognition time unacceptable. Since the acoustic model is the foundation of the whole system, we will focus on simplifying the acoustic model to adapt the system’s requirement in this paper.

To find out the best selection in the PBSI isolated-word recognition, we mainly study three problems related with the acoustic model, which are:
- Deciding a set of sub-word units — bi-phone or mono-phone? Tone is distinguished or undistinguished?
- Simplify the emission probabilities — the amount of mixtures in each state? Full or diagonal covariance in each Gaussian distribution?
- Choosing the speech feature extracted in the front end — the type of feature, frame length & frame step, feature dimensions.

Abundant experiments are designed to get the most effective acoustic model under medium vocabulary isolated-word recognition task.

This paper is organized as follow. In the section 2 we recommend the acoustic layer of a LVCSR baseline system. The comparative study is described in the section 3. Section 4 illustrates the relevant experimental results. The conclusion is drawn in the section 5.

2. THE BASELINE SYSTEM
Indubitably CDHMM, which is implemented in the most of LVCSR systems, is the best acoustic model until now. To attain the highest performance, our PBSI name dialing system will be derived from the LVCSR system using CDHMM. Hence, we firstly look over the LVCSR baseline system.

According to the knowledge of mandarin phonetics, the mandarin phoneme, usually called as semi-syllable, is composed of 27 consonants and 38 vowels. Then, Mandarin Chinese is a tone-distinguished language and each vowel involves four tones. As the pronunciation of consonant is deeply affected by the following vowel, a set of clustering context-dependent sub-word units is adopted in our system. To obtain better models, each vowel’s tone is distinguished. In this way we get a sub-word units set, which is constituted by 102 consonants (initials), 146 vowels (finals), 1 silence model and 1 pause model [1].

State is the essential unit in an acoustic model set. In the baseline system, the number of state models in a semi-syllable is various differing with the semi-syllable’s type. Silence, pause, initial, final semi-syllables respectively include 1, 1, 2, 4 states.

The finite mixture of Gaussian distribution is introduced to measure the likelihood between a feature vector and a state model [2]. In our LVCSR system the emission probability model of each state is simulated by a Gaussian mixture with full covariance matrix.

The front-end information in the baseline system is as follows. The sample rate is 16 kHz and each sample is linearly quantified into 16 bits. The frame length is 20ms and the frame step is 10ms. A frame of feature vector has 45 dimensions, which include 14 PLPCs (Perceptual Linear Predictive Coefficient), normalized energy and their first and second order of differentials [3].
At the acoustic layer an entire syllables recognition decoding network is applied. Then, we acquire a fine tone-undistinguished syllable error rate that is 30.38%.

3. SIMPLIFYING ACOUSTIC MODELS

Considering that the PBSI name dialing system is not as complicated as LVCSR system, the acoustic models used in the baseline system may be too precise for our current system. Therefore, careful study must be done to cut down the hardware demanding.

3.1 Sub-word units

From 2.1 we know that the mandarin Chinese is a language with four tones in each vowel. Accordingly, the two key points need to be taken into account about the sub-word units employed in the current system.

3.1.1 Tones are distinguished or undistinguished?

In the LVCSR system, we usually adopt tone-distinguished models, but ignore the tones’ differences when checking the syllable recognition results. Thus, the tone-distinguished models perform better than tone-undistinguished ones [1].

![Fig1. Different state-nodes’ connections of digit “3”](image)

To test the performance of the tone-distinguished models and the tone-undistinguished models in the PBSI name dialing application, we respectively introduce them to our PBSI isolated-word recognition system, where fig1.a is the state node chain of digit “3” with tone-distinguished models and fig1.b is the one with tone-undistinguished models.

It is unlike the LVCSR systems that the recognition accuracy with tone-distinguished models is NOT better than tone-undistinguished, although the semi-syllables’ number of tone-distinguished models is much more. Tonal modification in the mandarin continuous speech is one of important reasons. Therefore, multi-tone state node chain is applied in our system, which not only bring a great deal of hardware resources demanding, but also lose a little recognition accuracy for its complicated topological structure. Fig1.c is a typical sample of multi-tone state node chain.

3.1.2 Bi-phone or mono-phone?

In the LVCSR system, a set of clustering bi-phone units has been adopted. If the tone of each vowel is not distinguished, the clustering bi-phone units include 358 state models and yet the mono-phone units just involve 208 state models [2]. Whether the mono-phone units can obtain the same performance as bi-phone just like tone-undistinguished units doing is in focus.

It is proved that the bi-phone model works better when the quantity of word list is larger and the confusion is higher for the bi-phone model is more accurate. Nevertheless, the difference between them becomes smaller and smaller with the vocabulary’s scale reducing. The corresponding experiments are described in the experimental results.

3.2 The emission probability

In our PBSI name dialing system, the finite mixtures with Gaussian distribution are still betaken as the emission probability of each state. The function of emission probability is as follows.

\[ b_s(x) = \sum_{m=1}^{M} \pi_m \phi_{m,s}(x) \]

where

\[ \phi_{m,s}(x) = \frac{1}{(2\pi)^{N/2} |\Sigma_m|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_m)^T (\Sigma_m)^{-1} (x - \mu_m) \right\} \]

in which \( s \) is the current state number, \( M \) is the number of Gaussian mixtures in a state, \( \mu_m \), \( \Sigma_m \) and \( c_m \) is the mean, the covariance matrix and the mixture weight of the \( m \)th mixture in the \( s \)th state respectively and \( x \) is a frame of feature vector [3].

In the practical system, the log-likelihood is adopted to limit the range of \( b_s(x) \). The function to compute the log-likelihood is:

\[ L_s(x) = \ln \left[ \sum_{m=1}^{M} \frac{c_m}{(2\pi)^{N/2} |\Sigma_m|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_m)^T (\Sigma_m)^{-1} (x - \mu_m) \right\} \right] \]

In respect that calculating the likelihood of state emission probabilities usually occupies about half recognition time, how to decrease the complexity of the state emission probabilities is a hot point in the speech recognition field.

There are two factors, the quantity of Gaussian mixtures in a state and the type of covariance matrices in each mixture, can directly affect the calculation time and the memory spending. Supposing that the feature vector’s dimension is \( N \) and each state makes up of \( M \) Gaussian mixtures, and then when the covariance matrix is full, it will take

\[ M \times (N^2 + N) \times [a \times b] + M \times N^2 \times [a + b] + (M - 1) \times \log(expa + expb) \]

to calculate emission probability of this state and

\[ 2M \times (N \times (N + 1)/2 + N + 1) \] bytes to save this state model; if diagonal covariance matrix is employed, the above items are

\[ M \times 2N \times [a \times b] + M \times N \times [a + b] + (M - 1) \times \log(expa + expb) \]

and \( 2M \times (2N + 1) \) bytes. \( \log(expa + expb) \) is equal to 2 additions, 11 multiplications, 1 comparison and 1 shift in our program.
It is evident that the model with full covariance matrices is much more complicated than with diagonal covariance matrices, and therefore multiple mixtures’ model with full covariance matrices are unbearable in our system despite that it has high recognition rate.

3.3 Features

In all the speech recognition systems, the analog speech signals are presented by the speech features extracted in the front end. It is the absolutely first step in the speech recognition. In fact, the features have tight relationship with a speech recognition system’s performance. Whatever are the feature’s type, the frame length, the frame step, the feature’s elements and their dimensions, any of them don’t make the exception.

3.3.1 Feature’s type

Although numerous types of features are once advanced in the speech recognition field, only a few features are actually proved to be successful in the practical systems. The most three typical features are LPCC (Linear Predictive Cepstrum Coefficient), MFCC (Mel-Frequency Cepstrum Coefficient) and PLPC. The extracting process of these three features can reference to fig2 [2].

Fig.2 Block diagram of LPCC/MFCC/PLPC extraction

In our embedded speech recognition system, the feature selecting must obey two rules:

- The recognition accuracy must be kept in a proper level;
- The operation of extracting a frame of feature must be accomplished in one frame step, for a sentence of speech is unable to be saved in the memory of the speech ASIC.

To synthetically consider the recognition rate and the characteristics of the speech ASIC, we introduce the MFCC as speech feature by reason of its high performance and medium quantity of computation.

3.3.2 Frame length & frame step

Since the speech is a quasi-periodical signal, the frame length is designed to be approximatively equal to the speech’s quasi-period, whose range is ordinarily from 20ms to 30ms.

The frame length is 20ms and the frame step is 10ms in the LVCSR system. When the sample rate is 16 kHz, the sample’s number in one frame is 320. Hence, in the following 512-point FFT (Fast Fourier Transform) 192 zeros must be filled to the vector in the time domain.

To utilize the potential computation ability of the speech ASIC, 32ms frame length and 16ms frame step are applied in current system, which bring on 512 samples in one frame. In this way the new frame number is just 62.5% comparing with the old. This means that the decoding network’s scale is also cut down to 62.5%. Under this circumstance the recognition rate is still no change.

3.3.3 Feature’s elements and their dimensions

In the baseline system feature’s elements are formed by MFCC, its first and second differentials (that are ΔMFCC and Δ²MFCC), the normalized energy E and its first and second differentials (shortening to ΔE and Δ²E). In these six elements mentioned above, the essentiality is arranged in this order [2]:

Tab1. The essentiality of feature’s elements

<table>
<thead>
<tr>
<th>Essentiality</th>
<th>Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st order</td>
<td>MFCC</td>
</tr>
<tr>
<td>2nd order</td>
<td>ΔMFCC &amp; E</td>
</tr>
<tr>
<td>3rd order</td>
<td>Δ²MFCC &amp; ΔE &amp; Δ²E</td>
</tr>
</tbody>
</table>

The Δ² MFCC in the 3rd order occupies 14 dimensions but cannot play such a significant role as its consumption, thus it is abandoned from feature’s elements. Consequently the feature comprises MFCC, ΔMFCC, E, ΔE and Δ²E, total of which is 31 dimensions.

We know that DCT (Discrete Cosine Transform), the last step when extracting MFCC, is one of sub-space transform, by which the feature vectors are ranked by their significances. From the experiment we can find that the MFCC dimensions can be compressed from 14 to 10 only with little accuracy’s loss in the current system.

4. EXPERIMENTAL RESULTS

Training set 1: National 863 standard Mandarin Speech Corpus. It involves 47 hours speech data spoken by 83 male speakers.

Training set 2: Microsoft Mandarin Continuous Speech Corpus. It includes 20 hours speech data spoken by 100 male speakers.

Recognizing set: Isolated-word Speech Corpus spoken by 10 male speakers. Each of them says 600 isolated words, which are respectively 200 place names, 200 person names and 200 stock names.

The above speech data are recorded in the ordinary office environment. The sample rate is 16 kHz and each sample is linearly quantified into 16 bits.

4.1 Tone-distinguished vs. tone-undistinguished

In this experiment the bi-phone model is employed. The emission probability in each state is simulated by one mixture with full covariance matrix.

This experiment indicates that the tone-undistinguished model is more applicable than tone-distinguished model. The subsequent experiments all base on this result.
4.2 Bi-phone vs. mono-phone
The target of this experiment is to compare these two models’ ability while different word numbers in the lists are betaken.

Although the models with full covariance matrix can obtain higher recognition rate, their QM and QC is unbearable in any embedded system. By this reason the finite Gaussian mixtures with diagonal covariance matrix.

4.4 Experiment about features
Both 20ms and 32ms frame lengths are adopted while the feature’s dimension is 31. Then, the recognition rates with 31, 27, 25, 23 dimensions and 32ms frame length are illustrated in the table 4. In this experiment 3 Gaussian mixtures with diagonal covariance matrix are employed.

In addition it is clear that the mono-phone model works better in recognizing the stock names than place names, for the stock name is not as confused as place name.

And reducing the dimension number seems no obvious effect on the system’s performance, 25 or 27 dimensions are applied in the practical system to keep robustness.

5. CONCLUSION
Comparing to the LVCSR system we can draw some conclusions as follows:

- Tone-undistinguished model is more efficient;
- While the word number in a list is beyond 50, bi-phone model is essential; otherwise, mono-phone is fine;
- It is the best choice to introduce the finite Gaussian mixture with diagonal covariance matrix computing likelihood;
- 32ms frame length and 27 dimensions feature are adopted.

6. ACKNOWLEDGEMENT
It is appreciated that Microsoft Corp. supplies the speech corpus. The research is supported by National Natural Science Funds of China (Item No. 69975007) and National High Technology R&D Program (863) of China (863-306ZD13-04-6).

7. REFERENCES