USING BOTH GLOBAL AND LOCAL HIDDEN MARKOV MODELS FOR AUTOMATIC SPEECH UNIT SEGMENTATION

Zheng Hong¹, Zu Yiqing²

¹ CASCO (ALSTOM) Signal Ltd., China
zhenghong@casco.com.cn

² Motorola China Research Center, China
Yiqing.Zu@motorola.com

ABSTRACT

This paper describes an effective method for automatic speech unit segmentation. Based on hidden Markov models (HMM), an initial estimation of segmentation from the explicit phonetic transcription are processed by our local HMM training algorithm. With reliable silence boundaries obtained by a silence detector, this algorithm tries different training methods to overcome the insufficient training data problem. The results show that using this method, a 14.98% improvement is achieved in the boundary detection error rate (deviating larger than 20 ms).

1. INTRODUCTION

Speech segmentation [1] on a large speech corpus is very important for a corpus-based concatenated text-to-speech (TTS) system. The accuracy of segmentation boundary is crucial to the quality of synthesis speech. The speech unit can be phone, diphone, polyphone and so on.

Researchers have developed many methods used for speech segmentation. Phonetic experts can generate the segmentation based on speech analysis, such as spectrograms, intonation, energy curves, and other techniques. This method is called manual segmentation and assures a very good result in the segmentation. However, it is a tedious and time consuming task to manually segment speech data. Moreover, the results of manual segmentation lack reproducibility because of the human’s subjective decisions involved. In manual segmentation, consistency problems are also serious. Therefore manual segmentation is non-ideal.

In so-called trainable TTS [2,3], there are great amount of segmentation tasks, and automatic segmentation becomes more important. There are different tasks in automatic speech unit segmentation on a speech corpus: (1) Automatic segmentation, which aligns a known speech unit sequence to speech waveform. (2) Automatic correct phonetic transcription. In fact, it’s a difficult task to obtain a correct phonetic transcription. In continuous speech, even when we get a correct words sequence, some phoneme elision and assimilation occurred. The automatic correct phonetic transcription is to spot those phonetic phenomena and correct the phonetic transcription.

This study focuses on the automatic segmentation problem. In correspondence, a great number of methods for automatic segmentation have been proposed, for example, hidden Markov models (HMM) [4], neural networks [5], statistical modeling [6], and parametric filtering [7]. These automatic methods can roughly be classified into two groups: implicit segmentation and explicit segmentation [8]. An implicit segmentation method splits up the utterance into segments without explicit information, such as the phonetic transcription. An explicit segmentation method splits up the incoming utterance into segments that are explicitly defined by the phonetic transcription. An implicit segmentation method does not use any phonetic information on speech waveform. Therefore an explicit method could perform better in principle. An explicit method uses reference patterns to describe each segment. Such patterns do not always fit well the utterance to be segmented and do not account for all the variability occurring in natural speech. It indicates that a combination of these two methods is expected to get better results than each method separately. This paper proposes a mixed method to improve the final segmentation result, in which the explicit and implicit ways are used together. The block diagram of this method is shown in Figure 1.

Figure 1. Automatic Speech Segmentation
In this paper we introduce the initial speech segmentation based on HMM in section 2. In section 3, we describe the improved segmentation, which includes improved silence detection and a local HMM segmentation algorithm. The training and testing data used in this study is a Mandarin TTS speech corpus.

2. INITIAL SEGMENTATION

Currently, the HMM techniques give the best result for automatic recognition. Upon applying HMM to automatic speech recognition, there exists an implicit segmentation process called model alignment, and with some modifications to reduce the computational cost of a complete recognition, it is possible to use them stand-alone for speech segmentation. Most of the current speech segmentation techniques are based on modeling each one of the phonetic units, usually with HMM.

In this study, the initial speech analysis comprises 12th order Mel frequency cepstral coefficients (MFCC) parameter and energy, plus delta and acceleration coefficients (totally 39 coefficients). The speech signal has first order pre-emphasis applied. The acoustic vector is formed each 10 ms by analyzing speech frames of 20 ms using a Hamming window. A set of 3-emitting-state left-to-right no-skip context-independent HMM with continuous (M Gaussian Mixture) PDF is trained using Baum-Welch re-estimate algorithm from the orthographic transcription. A pronunciation dictionary is used to create the acoustic models for Mandarin automatic speech recognition. Upon applying HMM to automatic recognition, there exists an implicit segmentation process called model alignment, and with some modifications to reduce the computational cost of a complete recognition, it is possible to use them stand-alone for speech segmentation. Most of the current speech segmentation techniques are based on modeling each one of the phonetic units, usually with HMM.

To distinguish from the local HMMs (described in section 3.2), these models are called global HMMs. After training by global HMM, a Viterbi-based alignment of the acoustic data with the phonetic transcription is performed to obtain an optimal sequence of speech unit boundaries.

3. IMPROVED SEGMENTATION

3.1 Local HMM training

In our speech recognition procedure, parameters of each HMM are re-estimated from one speaker’s great number of training data of variant contexts. This reliability makes the HMM be able to represent accurately most of recognition situations. However, in segmentation tasks, as the phonetic transcription is known, there are more complex factor influence the segmentation results. It would be better to have specific models such as phone/syllable dependent models. These models can be adapted to the local properties of the speech. Therefore, we suppose that the frames of acoustic features of a phone are more similar to that phone than to the context phones [10]. We can use these local HMMs with single Gaussian PDF to refine the segmentation on the Mandarin TTS speech corpus.

There is often no enough training data to train these models. Especially for the left-right models, the major problem is that one cannot use a single observation sequence to train the model. This is because the transient nature of the states within the model allows only a small number of observations for any state, until a transition is made to a successor state. There are some possible solutions to the problem of insufficient training data [11]:

- Increase the size of the training data.
- Reduce the size of the model.
- Smooth parameters.
- Tie parameters.

In this study, we adopt the parameter smoothing method. For continuous mixture HMM, we must pay extra attention to smoothing the covariance matrices. Three methods are used in this study:

- **VF**: Set variance floor, which can be used effectively for parameter interpolation to maintain adequate amount of training data;
- **GAU**: the covariance matrix of the models is assumed to be diagonal. For each segment the mean and covariance are estimated from the data of that segment;
- **EUC**: the covariance matrix of the models is assumed to be the identity.

3.2 Silence detector

By comparison with manual segmentation, the result of initial segmentation indicates that the accuracy is imperfect with many silence detection errors, though we construct silence models with different topologies in which there is only one emitting state. The rough boundaries of silence also lead other phonetic segmentation errors. Therefore, it is necessary to detect reliable silence boundaries.

To solve this problem, we use Rabiner and Sambur (R.S.) algorithm to detect silence from speech [9]. This algorithm is based on the features of magnitude and zero crossing rate and sets up a threshold for discriminating silence from speech. The

<table>
<thead>
<tr>
<th>Table 1: Initial/Final units in Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial</strong></td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>z</td>
</tr>
<tr>
<td>zh</td>
</tr>
<tr>
<td>l</td>
</tr>
<tr>
<td>s</td>
</tr>
<tr>
<td><strong>Final</strong></td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>e</td>
</tr>
<tr>
<td>i</td>
</tr>
<tr>
<td>ao</td>
</tr>
<tr>
<td>ang</td>
</tr>
<tr>
<td>uai</td>
</tr>
<tr>
<td>ao</td>
</tr>
<tr>
<td><strong>Nasal coda</strong></td>
</tr>
<tr>
<td>n</td>
</tr>
</tbody>
</table>
modified R.S. algorithm used for this study can be described by the following equations,

$$IZCT = \min(4IT, 2\sigma_{IZC} + 2\sigma_{IZC})$$  \hspace{1cm} (1)
$$I1 = 0.03 \times (IMX - IMN) + IMN$$  \hspace{1cm} (2)
$$I2 = 4 \times IMN$$  \hspace{1cm} (3)
$$ITL = \min(I1, I2)$$  \hspace{1cm} (4)
$$ITU = 5 \times ITL$$  \hspace{1cm} (5)

Where,

- $IZC$ is the zero crossing threshold;
- $ITL$ is the lower threshold of energy;
- $ITU$ is the upper threshold of energy;
- $IZC$ is the sum of the mean zero crossing rate during silence;
- $\sigma_{IZC}$ is the standard deviation of the zero crossing rate;
- $IMN$ represents the minimum energy value;
- $IMX$ represents the maximum energy value.

For each utterance in speech corpus, we assume that the first 60 ms interval may contain silence. During this interval, a statistical characteristic of silence or background noise is calculated by 

$$r_{j} = p(O_1 \cdots O_N | \lambda_1 \cdots \lambda_N)$$

For each boundary $b_j$, $j = 1 \cdots N - 1$

- For each movement $k$, $k = 1 \cdots M$, $M$ is the number of boundary moving to one direction.
- Move $b_j$ to $b_j - k$
- Compute $r_j = p(O_1 \cdots O_N | \lambda_1 \cdots \lambda_N)$
- Move $b_j$ to $b_j + k$
- Compute $r_j = p(O_1 \cdots O_N | \lambda_1 \cdots \lambda_N)$
- If $Max(r_0, r_1, r_2) = r_1$
  - $r_0 = r_1$
  - Move $b_j$ to $b_j - k$
- If $Max(r_0, r_1, r_2) = r_2$
  - $r_0 = r_2$
  - Move $b_j$ to $b_j + k$

For each utterance

Train local model $l_j$ of each segment in the utterance,

$$j = 1 \cdots N, N$$ is the number of segments

Calculate the initial reference probability,

$$r_0 = p(O_1 \cdots O_N | \lambda_1 \cdots \lambda_N)$$

Table 2: Fricatives, stops and nasals in Mandarin

<table>
<thead>
<tr>
<th>Fricative stops</th>
<th>z zh ch ch j q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stops</td>
<td>b d g k p t</td>
</tr>
</tbody>
</table>

3.3 Segmentation algorithm

As described in Figure 1, our segmentation algorithm produces initial segments by using methods presented in section 2 at first. Then the speech signal of each utterance is input to the silence detector. The output is used to locate silence in the speech. The locational information of silence is unchanged in next procedures. After detecting silence, local HMM mentioned in section 3.1 are trained. For each segment of the utterance, a local model is estimated. If some speech unit appears immediately before stops or fricative stops is difficult to detect. Based on the initial detected boundaries, we subtract 20ms from original silence’s right boundary that lies immediately before stops or fricative stops. All stops and fricative stops in Mandarin are shown in Table 2.

Finally, the silence detector corrects the position of silence in the utterance automatically. The redundant silence is discarded, while silence not existing in the automatic result is preserved and its boundary keeps unchanged.

Table 3 shows the segmentation result of the initial segmentation only using global HMM, improved segmentation

4. RESULTS

To evaluate our segmentation method, we select speech data from the Mandarin TTS speech corpus, which is uttered by a professional female speaker. This speech corpus also includes orthographic transcriptions and a pronunciation dictionary. All utterances are used for training the global HMMs. 200 sentences selected from this speech corpus are used for testing the segmentation method of local HMMs. These sentences have been labeled manually to each segment. In this study, acoustic models consist of 33 initial/tonal-final context-independent phones. Each model has 3-emitting-state Bakis topology without any skip arcs, plus 2 non-emitting state (an entry state and an exist state). The number of Gaussian mixtures is set to 19. The number ($M$) of boundary frames moving to each direction is set to 2 (or 20ms). The minimum length of segments is 30ms. This value assures that the local HMMs training can be carried on properly. The configuration information about frame length, frame rate, acoustic vector representation and the initial segmentation algorithm are described in section 2.

Table 3 shows the segmentation result of the initial segmentation only using global HMM, improved segmentation...
without silence detection and improved segmentation with silence detection, in comparison with manual segmentation.

Table 3: Segmentation results on Mandarin TTS speech corpus

<table>
<thead>
<tr>
<th>Time Diff (ms)</th>
<th>IS (%)</th>
<th>ISNSD (%)</th>
<th>ISSD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VF (GAU)</td>
<td>GAU</td>
<td>EUC</td>
</tr>
<tr>
<td>30</td>
<td>86.44</td>
<td>89.89</td>
<td>88.47</td>
</tr>
<tr>
<td>20</td>
<td>69.56</td>
<td>75.98</td>
<td>73.60</td>
</tr>
<tr>
<td>10</td>
<td>39.42</td>
<td>42.87</td>
<td>41.91</td>
</tr>
</tbody>
</table>

Note:
IS: Initial segmentation only using global HMM;
ISNSD: Improved segmentation without silence detection;
ISSD: Improved segmentation with silence detection;
In ISNSD and ISSD method, there are three types: applying VF (GAU), GAU and EUC. In VF, the variance floor is set to 6.
Time Diff.: the difference between automatic segmentation boundary and manual boundary.

As a result, the EUC method has the best performance whenever it is used in ISNSD or ISSD. The average boundary error (less than 20ms) in the ISSD (EUC) decreases 14.98% from the IS. We can see that the ISSD (VF (GAU)), ISSD (GAU) and ISSD (EUC) decrease the error rate by 7.02%, 8.33% and 2.26% from the ISNSD (VF (GAU)), ISNSD (GAU) and ISNSD (EUC) respectively. It proves that the silence detector also play an effective role in refining the boundaries of phonetic units.

5. CONCLUSION

In this paper, a new method has been proposed to improve the performance of speech segmentation. This method combines the explicit and implicit phonetic information to produce final segments. It constructs global HMM for each speech unit from explicit phonetic transcription and then processes speech signal implicitly, The Baum-Welch algorithm is used to train these models, and the Viterbi algorithm is used to align the speech units in each utterance. Silence boundaries in the result of Viterbi alignment are replaced by those detected by the silence detector. Finally, local HMMs for each segment in the utterance are trained by different methods aiming at reducing the effect of insufficient training data. Using the local HMMs, we refine the boundaries and make a comparison between various segmentation methods. It shows that our new segmentation method can reduce 14.98% the number of boundaries deviating larger than 20ms of manually placed boundary. In summary, we can conclude as follows:

1. Local HMM can improve the automatic segmentation effectively and practically;
2. Effect silence detection plays an important role in local HMM training;
3. By experience, some phones need special process. For example, we need to put extra 20 ms silence before stops and fricative stops.

6. REFERENCES