AUTOMATIC SPEECH SEGMENTATION WITH HMM

John Dines, Sridha Sridharan and Miles Moody
Speech Research Laboratory, Queensland University of Technology

ABSTRACT: In this paper we review aspects of our automatic speech segmentation system that has been utilised in conjunction with our speech synthesis research. The speech segmentation system is based on a hidden Markov model phone recogniser using training strategies optimised for the segmentation task. Our research includes an analysis of the various aspects of the phone recogniser’s design and identifying the distinctions between paradigms of parameter estimation for speech segmentation and recognition. We also look at the limitations of HMM based segmentation and techniques for overcoming these limitations. The system evaluation demonstrates the ability of our system to provide high reliability speech segmentation that is comparable in performance to other state of the art systems.

INTRODUCTION

Segmentation of speech according to its phonetic transcription has application in the fields of speech and speaker recognition, speech synthesis and speech coding. In our work (Dines et al., 2001) we have applied the speech segmentation task to the automatic phonetic annotation of speech waveforms for use as speech synthesis unit inventories. Normally this is carried out manually due to the need for the reliable annotation of the inventory, but this is a time consuming task that could greatly benefit from an automated analysis procedure such as the one described in this paper.

Various methods have been employed for the phonetic segmentation of speech, the most prevalent being techniques that rely on the hidden Markov model (HMM) (Donovan, 1996; Ljolje et al., 1996; Nefti and Bœffard, 2001; Pellom, 1998; Toledano and Gómez, 2002). The HMM currently forms the backbone of most modern speech recognition systems and, intuitively, it also seems well suited to the task of speech segmentation. Our research is also motivated by the search for a correspondence between high accuracy speech segmentation and the quality of speech reconstruction obtained from directly using the segmentation models for synthesis. This paper focuses on aspects of HMM system design for speech segmentation that diverge from ‘classical’ speech recognition approaches and attempts to explain why such divergences are necessary. We also look at methods for the correction of systematic segmentation error and the underlying reasons for the occurrence of such errors. The findings of this research show that highly reliable speech segmentation can be achieved with relative ease and also shows avenues for future improvements that can be made to the system.

An outline of the paper is as follows: Section 2 gives an overview of the HMM based speech segmentation system followed by Section 3 which gives a description of the various evaluations of the system that were carried out. Section 4 describes some techniques used for the correction of systematic alignment errors and finally Section 5 gives an overview of our research findings and points out areas where the system may be improved.

PHONETIC SEGMENTATION OF SPEECH USING HMM

Phone segmentation using HMM can be viewed as being similar in its implementation to the phone recognition problem except that the phone transitions are constrained by the transcription (ie. prior knowledge of what is being said is assumed) and instead of assessing performance by the recognition accuracy we are measuring the discrepancies between the manual segmentation(s) and the automatic segmentation. The system comprises the following critical elements:

- Feature extraction module providing an \( M \)-dimensional representation of successive quasi-stationary windows of speech data;
- HMM phone models whose parameters are estimated from speech features generated from the training data;
- Word-to-phone transcription module which predicts multiple or a single pronunciation of the word level transcription.

Proceedings of the 9th Australian International Conference on Speech Science & Technology

Accepted after full review
Viterbi decoding of test utterance(s) from the transition network determined by the provided pronunciation(s).

The TIMIT speech corpus (Garofolo et al., 1990) provides an ideal resource for the training and testing of our segmentation system as accurate time aligned transcriptions are provided for all of the utterances in the database. In our experiments we use only the male speakers of the TIMIT database which constitutes 3260 and 1120 utterances for training and testing respectively (making the system gender dependent). The phone model set comprises 48 left-right HMMs trained on 22nd order mel-warped cepstral features (Fukada et al., 1992) extracted every 5 msec with a 25 msec Hanning window from the 16 kHz speech data (including the 0th and delta features giving a 46 dimensional feature space)\(^1\). Figure 1 shows a block diagram of the baseline system.

![Block Diagram of Baseline HMM Segmentation Tool](image)

There are two types of error that arise from the phonetic transcription and alignment process, these being transcription errors and segmentation errors. In our work we assume that a 'correct' transcription is available (manual and hypothesis transcriptions are the same) and we only assess the segmentation accuracy. There are several means to gauge the segmentation accuracy, in this paper we calculate the percentage of segment boundaries that are located within a certain time range of the manually located boundaries. This measure provides a useful insight into the distribution of errors, where the percentage of segments located within small time range indicates the ability of the system to precisely locate segment boundaries and the percentage of segments located within a large time range indicates the incidence of catastrophic segmentation errors.

SYSTEM EVALUATION

Various topological and training strategies were tested to ascertain the most appropriate baseline system. A number of interesting discoveries were made with regard to the best performing model topology and training strategy. This section summarises the various experiments that were carried out and any notable observations that were made.

\(^1\)A lower order feature analysis could be conducted with similar results. We use a high order analysis to facilitate higher quality synthesis from these models.
Topological Experiments.

The first phase in the design of the HMM segmentation system requires the choice of model topology. A number of experiments were run to determine the optimal choice of states and mixture components for the given training data. Early experiments revealed that choosing a heterogenous set of model topologies over a homogenous set did not provide performance improvements and only complicated the design process. As a result further experiments were conducted with the same topology for all models, excepting the plosive and affricates which were best modelled as separate closure/release portions, each comprising three states. This facilitates the flexible modelling of these phones, allowing for the optional burst release of plosives, optional closure that can precede affricate release and also the substitution of the tap model [dx] for the plosives [t] and [d] which can occur frequently in American English.

Table 1 summarises the alignment accuracy results for the model topology experiments. It was found that modelling the phones with a greater number of states than has been reported in previous research produces better alignment results. We postulate that an increased number of states in the phone model achieves higher accuracy modelling of the segment boundaries by more closely fitting the trajectories of the speech features in a piecewise linear manner. More accurate modelling of feature trajectories does not contribute to speech recognition performance, hence, fewer states tend to be used in speech recognition applications. Increasing the number of mixtures appears to reduce the number of large segmentation errors (≥ 20 msec), but reduces precision (≤ 5 msec).

<table>
<thead>
<tr>
<th># Mixtures</th>
<th># States</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&lt; 5</td>
<td>&lt; 20</td>
<td>&lt; 5</td>
<td>&lt; 20</td>
</tr>
<tr>
<td>1</td>
<td>41.87</td>
<td>83.50</td>
<td>44.69</td>
<td>85.62</td>
<td>45.19</td>
</tr>
<tr>
<td>8</td>
<td>38.30</td>
<td>83.98</td>
<td>42.45</td>
<td>86.49</td>
<td>43.53</td>
</tr>
</tbody>
</table>

Table 1: Percentage of segment boundaries located within the given time range for various model topologies (trained using isolated parameter estimation).

Training Experiments

The training strategy adopted can also influence the alignment results for the HMM segmentation system. There were three training paradigms that were investigated:

**Case 1.** Model parameters are estimated using only the isolated training algorithms (Viterbi and Baum-Welch) from manual segmentations of the training data.

**Case 2.** Following the isolated training of model parameters (bootstrapping) parameter estimates are updated using the embedded training procedure.

**Case 3.** Context dependent models are estimated from embedded trained models (from Case 2) with tying of states using decision trees to ensure sufficient data coverage.

The results obtained for the TIMIT test set are shown in Table 2. The best performance was achieved using isolated training which is in contrast with speech recognition approaches to parameter estimation which rely heavily on embedded training. As expected the context dependent models performed poorly as their distributions tend to lose correspondence with the original manual segmentations as observed by Ljolje et al. (1996).

The concept of losing correspondence with the underlying phone sequence can also be extended to explain the improved performance of the isolated training scheme over the embedded estimation scheme. The isolated training scheme uses the manual segmentations as located by expert labellers to estimate the phone model parameters, each in isolation from the other phone feature statistics, whereas the embedded estimation of model parameters is carried out simultaneously for the entire model set, finding the model parameters that best represent the entire set of training data, regardless of the human labelling of the underlying phone sequence, although this effect is not as severe when the models are context independent. Furthermore, this phenomenon is made more distinct when state emission
<table>
<thead>
<tr>
<th>Training Condition</th>
<th># Mixtures</th>
<th>≤ 5</th>
<th>≤ 10</th>
<th>≤ 20</th>
<th>≤ 40</th>
<th>≤ 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>1</td>
<td>45.19</td>
<td>69.81</td>
<td>86.85</td>
<td>96.21</td>
<td>98.57</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>43.53</td>
<td>69.81</td>
<td>87.73</td>
<td>96.89</td>
<td>98.83</td>
</tr>
<tr>
<td>Case 2</td>
<td>1</td>
<td>38.64</td>
<td>63.50</td>
<td>83.63</td>
<td>95.04</td>
<td>98.12</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>40.31</td>
<td>66.50</td>
<td>86.25</td>
<td>96.43</td>
<td>98.70</td>
</tr>
<tr>
<td>Case 3</td>
<td>1</td>
<td>40.84</td>
<td>66.97</td>
<td>85.57</td>
<td>96.22</td>
<td>98.61</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>35.04</td>
<td>61.04</td>
<td>85.08</td>
<td>96.22</td>
<td>98.72</td>
</tr>
</tbody>
</table>

Table 2: Percentage of segment boundaries located within the given time range for various training strategies (7-state topology was used).

distributions comprising multiple mixture components are estimated, as this can exacerbate this loss of correspondence.

**SEGMENTATION POST-PROCESSING**

An analysis of the statistical trends of segmentation errors reveals that they can occur in a systematic fashion, bringing to light a number of possible means of improving segmentation accuracy. These errors are inherent to the alignment process itself and can not be easily incorporated into the HMM system, rather, the segmentation can be post-processed to produce the final alignment. Drawing on previous research (Ljolje et al., 1996; Toledano and Gómez, 2002) systematic errors in the speech segmentation results were analysed and identified as being responsible for three types of bias:

- Constant bias resulting from the speech parameterisation process due to the averaging of non-stationary frames by the analysis window.
- Variable bias resulting from states’ emission pdf’s overlapping with the feature statistics of adjacent phones. This bias is proportional to the durations of the overlapping states.
- A strong correlation between the the manually labelled durations of some adjacent phone segments which manifests itself in a correlation between the automatically determined duration and the alignment error.

Bias statistics are collected for each of the boundary state clusters determined from the decision tree state tying process. The decision tree for the context dependent models was built using phonological context questions with an outlier threshold ensuring sufficient training tokens in each leaf node to ensure stability of the bias statistics (a threshold of 100 was used in our experiments).

Toledano and Gómez (2002) propose a means for correcting the pdf overlap bias by calculating the average state overlap for each boundary cluster. We use a modified version of this measure as shown in Equation (1).

\[
\begin{align*}
\epsilon_L(p_{CL}^{th}(c_L)) &= \sum_{n=0}^{N-1} \max \left\{ 0, \min \left\{ 1, \frac{\sum_{i=1}^{M_{Su}} - t(n)^{ASu}}{t(n+1)^{ASu} - t(n)^{ASu}} \right\} \right\} \\
\epsilon_R(p_{CR}^{th}(c_R)) &= \sum_{n=1}^{N} \max \left\{ 0, \min \left\{ 1, \frac{t(n)^{ASu} - \sum_{i=1}^{M_{Su}}}{t(n)^{ASu} - t(n-1)^{ASu}} \right\} \right\}
\end{align*}
\]  

(1)

where \( \epsilon_L \) and \( \epsilon_R \) are the left and right hand state overlap with the alignment error and \( p_{CL}^{th}(c_L) \) and \( p_{CR}^{th}(c_R) \) denote the \( i \)th phone segment in the \( u \)th training utterance with left and right boundary state clusters, \( c_L \) and \( c_R \) respectively. For the \( u \)th utterance we have \( M_{Su} = \{ t_{i}^{ASu} \}_{i=0,...,p_{u}} \) as the manual segmentation and \( ASu = \{ t(0)^{ASu}, t(1)^{ASu}, \ldots, t(N)^{ASu} \}_{i=1,...,p_{u}} \) as the automatic state segmentation for an \( N \) state HMM topology.
The new segment boundary times are then calculated as:

\[
\hat{t}(0)_{i}^{ASu} = t(0)_{i}^{ASu} + \sum_{n=1}^{N} \max \left\{ 0, \min \left\{ 1, \hat{E}^{cL} + n - N \right\} \right\} \cdot (t(n)_{i+1}^{ASu} - t(n-1)_{i+1}^{ASu}) \\
- \sum_{n=1}^{N} \max \left\{ 0, \min \left\{ 1, \hat{E}^{cR} - n \right\} \right\} \cdot (t(n)_{i}^{ASu} - t(n-1)_{i}^{ASu})
\]  

(2)

where \( \hat{E}^{cL} \) and \( \hat{E}^{cR} \) are the average state overlap for the clusters \( cL \) and \( cR \) respectively.

In addition to this state overlap bias it is also recognised that the constant bias and duration correlated bias also need to be accounted for. We do this by carrying out a multi-variate linear regression between the residual error, Equation (3), and the durations of the phone segments that comprise that boundary.

\[
e^{su}_{i} = \hat{t}^{MU}_{i} - \hat{t}(0)_{i}^{ASu}
\]  

(3)

The multivariate linear regression for clusters \( cL \) and \( cR \) is:

\[
M \cdot \begin{bmatrix}
a_{c} \\
b_{c}
\end{bmatrix} = E
\]

(4)

where:

\[
M = \begin{bmatrix}
\hat{t}(N)_{i}^{ASu} - \hat{t}(0)_{i}^{ASu} & \hat{t}(N)_{i+d_{c}}^{ASu} - \hat{t}(0)_{i+d_{c}}^{ASu} \\
\vdots & \vdots \\
\end{bmatrix}_{\forall \{ u \in \mathcal{U}, p_{u}^{p} = p(c) \}}
\]

\[
E = \begin{bmatrix}
e_{i}^{u} \\
\vdots
\end{bmatrix}_{\forall \{ u \in \mathcal{U}, p_{u}^{p} = p(c) \}}
\]

\[
C = \{ CL, CR \}
\]

\[
d_{c} = \begin{cases}
+1, & \text{for } c = CL \\
-1, & \text{for } c = CR
\end{cases}
\]

Such that the final boundary placement is determined as:

\[
\hat{t}(0)_{i}^{ASu} = \hat{t}(0)_{i}^{ASu} + 0.5 \cdot \left\{ a_{cL} \cdot (\hat{t}(N)_{i}^{ASu} - \hat{t}(0)_{i}^{ASu}) + b_{cL} \cdot (\hat{t}(N)_{i+d_{c}}^{ASu} - \hat{t}(0)_{i+d_{c}}^{ASu}) + c_{CL} \\
+ a_{cR} \cdot (\hat{t}(N)_{i}^{ASu} - \hat{t}(0)_{i}^{ASu}) + b_{cR} \cdot (\hat{t}(N)_{i+d_{c}}^{ASu} - \hat{t}(0)_{i+d_{c}}^{ASu}) + c_{CR} \right\}
\]

(5)

Table 3 shows the results obtained for various model sets. Comparing these results with the context dependent results shown previously in Table 2, significant performance improvement is obtained for all error tolerances using both compensation methods. We also make note that the single mixture models outperform the multiple mixture models, most likely due to the average overlap bias varying between dominant mixture components. More effective correction of overlap bias may be achievable by calculating bias statistics depending on the relative likelihood contributions of each mixture component for each occurrence of the tied state in the training data.

<table>
<thead>
<tr>
<th>Post-processing</th>
<th># Mixtures</th>
<th>( \leq 5 )</th>
<th>( \leq 10 )</th>
<th>( \leq 20 )</th>
<th>( \leq 40 )</th>
<th>( \leq 60 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap Only, ( t(0)_{i}^{ASu} )</td>
<td>1</td>
<td>49.04</td>
<td>73.07</td>
<td>89.64</td>
<td>97.35</td>
<td>98.97</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>47.06</td>
<td>71.74</td>
<td>88.87</td>
<td>97.36</td>
<td>99.03</td>
</tr>
<tr>
<td>Full, ( \hat{t}(0)_{i}^{ASu} )</td>
<td>1</td>
<td>50.38</td>
<td>75.52</td>
<td>92.00</td>
<td>98.39</td>
<td>99.48</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>48.45</td>
<td>74.03</td>
<td>91.59</td>
<td>98.39</td>
<td>99.48</td>
</tr>
</tbody>
</table>

Table 3: Percentage of segment boundaries located within the given time range after bias compensation for the context dependent models.
CONCLUSIONS

In this paper we have detailed the development of a HMM based speech segmentation tool. This tool provides reliable alignment of a given transcription suitable for use in speech synthesis inventory creation. Our research has found a number of design aspects which contribute to improved segmentation performance:

- Modelling with a greater number of states to achieve better trajectory modelling gives better alignment precision;
- Isolated training gives better modelling of phone boundary statistics compared to embedded estimation where manually transcribed training data is available;
- Systematic alignment error can be corrected to give greatly improved performance, especially for the context dependent models which can be made to give better overall performance than context independent models.
- Single mixture models perform best when compensation for systematic errors are employed, while multiple mixture models provide better robustness of the segmentation performance in isolated training (although precision is still degraded).

In comparing our results with that of other research, in particular the research of Toledano and Gómez (2002) we note two areas in which performance has been significantly improved. The results for the speaker independent tests using context dependent models plus boundary correction show that our system displays greater precision (50.38% versus 41.32% for errors ≤ 5 msec) which we attribute to the topological and training design paradigms which we have used. In the larger error tolerances we also see significant improvement (92.00% versus 88.90% for errors ≤ 20 msec) which we attribute to the linear regression based post-processing technique. Research has shown that these results can be further improved through boundary refinement techniques such as those used in Toledano's work.

These observation provide direction for future research, including the investigation of the performance of smooth trajectory modelling techniques to speech segmentation, assessing the relative contributions of mixture components for the calculation of pdf overlap error, and the adaptation of the alignment system for language/accent independence.

REFERENCES


Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N. and Zue, V. (1990), TIMIT Acoustic-Phonetic Continuous Speech Corpus, National Institute of Standards and Technology, Gaithersburg, Maryland, USA.


