Improving Initial Boundary Estimation for HMM-based Automatic Phonetic Segmentation

Kalu U. Ogbureke, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
Kalu@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

This paper presents an approach to boundary estimation for automatic segmentation of speech given a phone (sound) sequence. The technique presented requires an extension to existing approaches to Hidden Markov Model based automatic segmentation which modifies the topology of the model to control for duration. An HMM system trained with this modified topology places 77.10%, 86.72% and 91.15% of the boundaries, on the TIMIT speech test corpus annotations, within 10, 15 and 20 ms respectively as compared with manual annotations. This represents an improvement over the baseline result of 70.99%, 83.50% and 89.18% for initial boundary estimation.

Index Terms: automatic phonetic segmentation, hidden markov models, gaussian mixture models

1. Introduction

Phonetic segmentation of speech is an area which has been well studied in speech science. This is because of its importance in the context of speech recognition and synthesis as well as in other areas of speech research. In stochastic approaches to speech recognition, accurate segmentation ensures appropriate initialization of model parameters during the bootstrapping process. Furthermore, concatenative text-to-speech (TTS) synthesis requires a very accurate phonetic segmentation in order to produce intelligible and natural speech.

Automatic approaches to speech segmentation have achieved widespread acceptance as an alternative to a manual approach. While a manual approach is considered by some to be more accurate [1], the inherent problems of cost, time and inter-transcriber agreement make it unsuitable for use with very large corpora and in a domain which is becoming ever more multilingual. Indeed, several expert annotators are unlikely to agree (to the millisecond) on exactly where a segment boundary should be placed. For this reason, it seems reasonable to assume that the best automatic annotation will only ever aim to be able to estimate such boundaries within a 10 to 15 millisecond time-frame. The aim of automatic segmentation is thus to model as accurately as possible the boundaries which have been manually annotated and thus the reference point for evaluation of the technique is standardly an existing manual annotation of some test data. This has led to a target which is somewhat arbitrary in that two manual annotations of the same data may differ considerably at the level of actual boundary points; this has been a problem for many developers of Hidden Markov Model (HMM) based automatic segmentation systems [2, 3, 4, 5, 6, 7, 8, 9]. Improved quality with respect to manual annotations has typically been achieved through a two-step process: initial boundary estimation followed by boundary refinement [1, 7, 8, 9, 10]. More recently in [11], a minimum boundary error framework and its hybrid used to train an HMM, is shown to improve segmentation accuracy compared to the manual annotation. This paper presents a further extension to the HMM-based approach based on a modified topology. The approach is applied at the phone level but it represents the first step towards the development of a finer grained automatic annotation system which uses phonetic features.

HMM-based systems perform Viterbi alignment on an existing acoustic model λ to determine the best state sequence q, given a phone sequence and an observation sequence O. The frames at the transitions from one phone to the other become the boundaries.

\[ P(O, q|\lambda) = P(O|q, \lambda)P(q|\lambda) \]

(1)

\[ q = q_1q_2q_3 \ldots q_N \]

(2)

\[ O = o_1o_2o_3 \ldots o_N \]

(3)

\[ \lambda = (A, B, \pi) \]

(4)

Where Λ is a matrix of state transition probabilities; B is a set of state observation probabilities; π is a special member of Λ: a set of initial state probabilities and N is the number of frames in the observation O. This addresses the second basic problem of HMM which deals with finding the best state sequence q giving an observation sequence O and a model λ, [12].

This contribution proposes an HMM topology for training of acoustic models which improves segmentation accuracy compared to manual annotation. The technique presented represents an extension to existing approaches to Hidden Markov Model based automatic segmentation which modifies the topology of the model to control for duration. An underlying feature is the use of a sufficient number of gaussian mixtures during model initialization process. The motivation for the research presented in this paper includes boundary estimation for both phonetic feature-based speech recognition and synthesis. This work is thus a first step towards a model which will support automatic annotation of a finer granularity at the level of phonetic features which should be more applicable across languages.

The rest of this paper is organized as follows. Section 2 provides an overview of the current state of the art. Section 3 describes the proposed architecture and its relevance to HMM parameter estimation. The training process, data and results are presented in section 4. Section 5 discusses results achieved and section 6 concludes with a summary of the contribution and highlights the future direction of this research.
2. State of the art

This section reviews some HMM-based segmentation systems reported in the literature. [1] presents some approaches to automatic segmentation and observes that context-independent HMMs are better for phonetic segmentation than their context-dependent counterparts. Thus, in order to compensate for the errors produced by context-dependent HMMs, a statistical correction procedure is proposed: Statistical Correction of Context Dependent Boundary Marks. An increase in segmentation precision is reported through the use of speaker adaptation techniques. The authors propose a framework for local refinement of the boundaries and report an increase in the performance of a baseline HMM tool. [2] describes an improved algorithm for automatic phonetic segmentation based on HMMs. It is demonstrated that forward-backward algorithms outperform Viterbi because the latter results in boundaries that reach maximum likelihood and is thus sub-optimal. Furthermore, the former provide confidence intervals on each of the generated boundaries which are used to remove biases between manual and automatic systems. The segmentation tool described is based on forward-backward algorithms. [3] combines an HMM-based approach and spectral boundary corrections using an iterative training and reports an improvement on the segmentation result compared to a manual annotation. The approach is aimed at reducing manual post-processing as well as achieving a better quality of synthetic speech as Viterbi approach only finds the most likely state sequence, not the optimal boundaries between adjacent segments. Spectral boundary correction uses spectral transition measures and phone class-dependent time window for boundary detection and correction.

[4] segments and refines the boundary using context-dependent HMM as well as predefined boundary models. The approach starts with initial estimate from a context-dependent phone based HMM system, followed by refinement with trainable statistical models that represent the phone boundaries. This is used for time mark refinement. It uses separate Gaussian Mixture Models (GMM) and HMM models for each boundary, based on the type of phonemes defining the boundary. [5] describes an extension to Baum-Welch algorithms for training HMMs that improves segmentation accuracy. This is achieved by the use of phoneme segmentation to constrain the forward and backward lattice of the algorithm. The approach also highlights that recognition and segmentation require different training techniques as well as confirms that for phonetic segmentation, context-independent models appear to be more effective than context-dependent models and state tying.

Other approaches include boundary refinement with Neural Networks [8] whereby a neural network is used as a phonetic transition probability estimator that estimates the probability of the frame within the estimated boundaries (refinement interval) being the transition frame. [9] presents a boundary refining method based on Classification and Regression Tree (CART) as well as Gaussian Mixture Models. It extracts features from frames within the refining interval which are used to train GMMs that model the boundaries. CART is used to automatically cluster phonetic boundaries based on their similarity with respect to acoustic features. More recently, [11] presents a minimum boundary error (MBE) framework and its hybrid used for HMM training as well as segmentation and shows an improvement in accuracy compared to a manual annotation. It uses MBE criterion which minimizes the boundary errors over a set of possible segmentations as well as improves the discriminability of HMMs for automatic speech segmentation. Finally, [13] investigates factors which affect HMM-based automatic segmentation systems. The effects of feature vector, parameterization, context dependency, topology as well as speaker and gender effects are considered. Boundary accuracies for classes of phonemes that are segmented well and those that are not are reported; nasals, fricatives and vowels are segmented more accurately than plosives and silences.

3. System overview

This section presents a system which represents an extension to existing approaches of automatic segmentation which modifies the topology of the HMM to control for duration. The system architecture is shown in figure 1. The prototype model for all phones is defined as a 5-state left-right topology with duration control states at either end, where states 1 and 7 are null states which do not emit observations. This is shown in figure 2. This topology improves segmentation accuracy by reducing the likelihood of remaining in the beginning and end states as these states model the boundaries between phones. The observation vectors at the transition from one phone to the other are clustered at these states; they model the transition from one phone to the other. An alternative topology of a 3-state model with duration control states is shown in figure 3. The common thing between figures 2 and 3 is that the probability of remaining at the states at either end is 0.

Each state in the model is split into 16 gaussian mixture components. Although the notion of mixture and its relation to HMM training is not new, conversion from single gaussian to multiple mixture component is usually one of the final steps in building a system [14]. However, mixture splitting is best done at initialization for phonetic segmentation using monophone models. The use of sufficient mixtures models the probability density function (PDF) of speech as speech PDF is not normally distributed. This is especially useful in speaker independent scenarios where the corpora consist of speech from a large number of different speakers. This ensures that the model parameters are close enough to the global maxima since Baum-Welch algorithm does not always converge to a global maxima.

Model initialization is done with the HTK tools HCompV and HInit [14]. HCompV assigns global mean and variance to all states and mixture components by uniformly segmenting the speech. This provides the seed HMMs for the initialization of the monophone models. HInit recomputes the mean and variance of each mixture and state using segmental k-means clustering and Viterbi segmentation to assign each observation vector...
Figure 2: Topology for the prototype HMM showing transition probabilities and duration control states.

to the most likely mixture component and state [14]. Initialization affects the final parameters of the models and thus the need to use a sufficient number of mixtures. The output of this stage is further refined using HRest tool which reestimates each monophone model using Baum-Welch algorithm.

In model refinement, the parameter of each monophone model is reestimated using the Baum-Welch algorithm. This addresses the third basic problem of HMMs, namely how to choose $\lambda$ such that $P(O|\lambda)$ is locally maximized [12]. The output of this stage is a set of monophone HMMs, each containing the model parameter $\lambda$. The monophone HMMs are used with HVite [14] to perform forced alignment on a speech signal given the phone sequence.

4. Training process and results

The TIMIT corpus of read speech was used for training of the acoustic models as well as testing [15]. It contains a total of 6300 sentences spoken by 630 speakers from 8 different dialects. 4620 (3.13hrs) sentences in the training set were used for training and the remaining 1680 (1.14hrs) sentences for testing. These are the standard training and test sets.

Table 1: Segmentation accuracy as a function of the number of mixtures for the 5-state HMM.

<table>
<thead>
<tr>
<th>No. of mixtures</th>
<th>% of boundaries within tolerance in ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>±5ms</td>
</tr>
<tr>
<td>ITA</td>
<td>60.38</td>
</tr>
<tr>
<td>Baseline</td>
<td>45.13</td>
</tr>
<tr>
<td>1</td>
<td>53.36</td>
</tr>
<tr>
<td>2</td>
<td>53.36</td>
</tr>
<tr>
<td>4</td>
<td>53.51</td>
</tr>
<tr>
<td>8</td>
<td>53.26</td>
</tr>
<tr>
<td>12</td>
<td>53.62</td>
</tr>
<tr>
<td>16</td>
<td>53.61</td>
</tr>
</tbody>
</table>

The acoustic model was trained using HTK [14]. It consists of 61 monophone models covering the complete TIMIT phone set. Each model is a left-right, 5-state HMM with duration control states at either end as shown in figure 2. The frame window size and period are 10 and 5 ms respectively, where each frame is parameterized as a 39-dimensional MFCCs consisting of the first 12 coefficients, with normalized energy, as well as their delta and delta-delta values.

The models were first trained with a 3-state HMM with a left-right topology and the number of gaussian mixtures increased from 1 to 16. The performance of the trained models with 16 mixtures was chosen as the baseline used in tables 1 and 2. Furthermore, the models were trained using a 3-state HMM with duration control states at either end. Finally, they were trained using a 5-state HMM with duration control states.

Table 2: Segmentation accuracy as a function of the number of mixtures for the 3-state HMM.

<table>
<thead>
<tr>
<th>No. of mixtures</th>
<th>% of boundaries within tolerance in ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>±5ms</td>
</tr>
<tr>
<td>ITA</td>
<td>60.38</td>
</tr>
<tr>
<td>Baseline</td>
<td>45.13</td>
</tr>
<tr>
<td>1</td>
<td>53.25</td>
</tr>
<tr>
<td>2</td>
<td>53.66</td>
</tr>
<tr>
<td>8</td>
<td>53.14</td>
</tr>
<tr>
<td>12</td>
<td>53.22</td>
</tr>
<tr>
<td>16</td>
<td>53.34</td>
</tr>
</tbody>
</table>

Besides accuracy within a tolerance, the Root Mean Squared Error (RMSE) criterion is also used as a measure of segmentation performance.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$$

where $n$ is the number of boundaries in the test set and $x_i$ is the difference between the reference boundary and the hypothesized boundary. Figure 4 shows the RMSE as a function of the number of mixture components. The short dash line is the baseline, the long dash line and straight line are for the 3-state and 5-state topologies with duration control states respectively.

5. Discussion

One of the suggested ways of improving the performance of HMM-based segmentation systems is to modify the topology by using duration control states at either end as shown in this paper. Tables 1 and 2 show improved segmentation accuracy, compared to manual annotation, when duration control states are used. This is reflected in both reduction in RMSE as well as increase in accuracy within a given tolerance. Duration control is essential in the current HMM framework to better adapt the sequence recognition framework to segmentation framework.

Furthermore, the use of sufficient number of mixture components is necessary, especially for speaker independent acoustic models where the training corpora come from a large number of different speakers. This models correctly the probability
density function of the speech. Mixture splitting is best done at initialization for monophone models.

6. Conclusions and future work

This paper presented an HMM topology for maximum likelihood training of acoustic models that improves phonetic segmentation accuracy compared with manual annotation. The purpose of this was to adapt the sequence recognition framework of HMMs to sequence segmentation. This improved segmentation accuracy by reducing the likelihood of remaining at the beginning and end states as these states modelled the boundaries between phones. Furthermore, mixture splitting for segmentation is best done at initialization. The use of sufficient mixtures models correctly the probability density function as corpora for training speaker independent acoustic models come from a large number of different speakers This ensures that model parameters are close enough to the global maxima at initialization.

Future work will involve a more detailed error analysis to determine the performance of each phone. Furthermore, a 5-state HMM has a longer minimum duration than a 3-state HMM and this can affect segmentation of very small phones which cannot be captured within the minimum duration of a 5-state HMM. This will be investigated. The modified topology serves as a basis for investigation of this approach at a finer level of granularity, namely with respect to articularatory and acoustic phonetic features in order to provide more extensive feature level annotations for feature-based speech recognition and synthesis in a multilingual domain.

7. Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

Peter Cahill suggested using the full TIMIT phone set as opposed to a reduced set. Finally, thanks to the reviewers for their comments on the previous version of this paper.

8. References