ICSLP 2002 (In Press)

REFINED SPEECH SEGMENTATION FOR CONCATENATIVE SPEECH SYNTHESIS

Abhinav Sethy, Shrikanth Narayanan

Speech Analysis and Interpretation Lab
Integrated Media Systems Center
Department of Electrical Engineering-Systems
University of Southern California
http://sail.usc.edu
[sethy,shri] @sipi.usc.edu

ABSTRACT

High accuracy phonetic segmentation is critical for achieving good quality in concatenative text to speech synthesis. Due to the shortcomings of current automated techniques based on HMM-based alignment or Dynamic Time Warping (DTW), manual verification and labeling are often required. In this paper we present a novel technique for automatic placement of phoneme boundaries in a speech waveform using explicit statistical models for phoneme boundaries. Thus we are able to cut down substantially on the labor and time intensive manual labeling process required to build a new voice. The phonetic speech segmentation is carried out using a two-step process, similar to the way a human expert would label the waveform. In the first step an initial estimate of the labeling is generated using an HMM based phoneme recognizer. The second step refines the boundary placements by searching for the best match in a region near the estimated boundaries with predefined boundary models generated from existing labeled speech corpora. The proposed method can be used in conjunction with any of the segmentation schemes used in practice. In the performance evaluations carried out the system is able to give time marks which are 30-40% better than the schemes currently used.

1. INTRODUCTION

Automatic labeling of speech waveforms is of great interest to many areas of speech research such as training speech models or constructing acoustic unit databases. It is especially important for concatenative text to speech synthesis which uses segmented corpora for construction of intonation, duration and synthesis components [1][2][3]. The quality of the synthesized speech hence depends critically on the accuracy of the labeling process. Currently the labeling step requires considerable human effort and is a long and painstaking task. This creates a big hurdle in rapidly generating new voices automatically from a set of utterances recorded from a given speaker.

Automatic segmentation methods rely on the knowledge of the underlying phone sequence. When the text (orthography) is available, phonetic transcriptions are generated by using dictionary lookup or letter to sound rules. A more challenging scenario is a completely unsupervised segmentation with the aid of a phoneme recognizer [4]. This work assumes the availability of the underlying phonetic transcription.

Once the phonetic transcription is available the standard approach for automated segmentation is to adapt an automatic speech recognition (ASR) system by restricting its language model to the transcription of the input sequence [5][6]. Results achieved by using these techniques are good but they are not sufficient for building high quality concatinative text to speech synthesis systems. Hence manual verification and labeling become necessary.

Text to speech systems require boundaries to obey standard phonetic conventions. Speech recognition techniques like HMMs on the other hand are more focused on correct identification of the sequence and not on accurately placing the phone boundaries. ASR systems do not have proximity to boundary positions as an optimality criterion in the training phase. So the underlying recognition system is not suitable for accurately segmenting the speech waveform.

In addition, the segmentation accuracy requirements for text to speech systems are very high. For example a 10 ms error in detection of closure-burst boundaries would lead to labeling which would give acoustic units that miss the burst part completely. ASR systems do not provide this level of accuracy for boundary placements.

In this paper we propose a system that attempts to address these shortcomings in the ASR based system. We will focus on the problem of accurately segmenting speech given an orthographic or phone level transcription. In such cases the segmentation problem is described as linguistically constrained segmentation or explicit segmentation [7]. It is possible that the underlying speech sequence differs from the phonetic transcription generated by the orthography to phoneme mapping (dictionary or letter to sound rules) or by using a phonetic recognizer. This occurs due to recognition errors or dialectal variations, which are not captured by a dictionary (or letter to sound rules). However we do not focus on the problem of correctly transcribing speech.

The paper is organized as follows. In section 2 we provide an overview of the proposed system. Section 3 describes the process for refining initial estimates of labeling. The training procedure and data are described in section 4. In section 5 we present results showing improved performance of this scheme. In the concluding section, we provide a summary of the scheme, our major findings and an outline for future research.

2. SYSTEM OVERVIEW

An ideal segmentation system should be able to work at different time resolutions. It should be able to reliably detect the phonemes and also detect speech changes with a high time resolution so that

the boundaries can be accurately located. However to reliably determine the phonemes we need to use high frequency resolution but the corresponding time resolution is poor.

To handle these contradicting requirements we take cues from speech segmentation performed by human experts, which is basically a two-step process. In the first step, the expert listens to put a rough set of marks to identify the sequence and then carries out a refinement process in which he uses his knowledge of phonetic convention and boundaries to locate the true boundary near the rough estimates.

In our system we first get an initial estimate of the boundaries from a context dependent phone based HMM(CDHMM) system, which is trained on an expert labeled generic speech database such as TIMIT. This system is constrained to operate on a language model specified in terms of the phonetic transcription of the speech utterance to be labeled. In cases where multiple pronunciations of the same word are possible the language model can be expanded to not only give the segmentation but the most probable transcription for the utterance, which is subsequently taken as the correct phonetic transcription for the utterance. There are alternatives to the context dependent HMM system such as DTW based alignment[8] and other HMM models which differ in terms of context information and the use of speaker adaptation techniques. These are well studied in the literature [9] and we will not discuss them further.

In the next stage we refine the boundary placement by moving each time mark to a position near the initial estimate where it matches maximally with predefined models for that phonetic boundary. The models are built so as to ensure that they follow the phonetic convention and waveform evolution properties for that particular phone boundary. The next section focuses on how the refinement process is carried out.

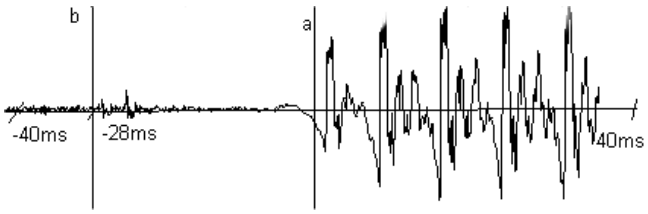

Fig. 1. 80 ms of speech at boundary between /f/ and /ea/. (a) corresponds to the crossover point labeled by a human and (b) is the label put by a CDHMM phone recognizer. (b) is 28 ms to left of (a)

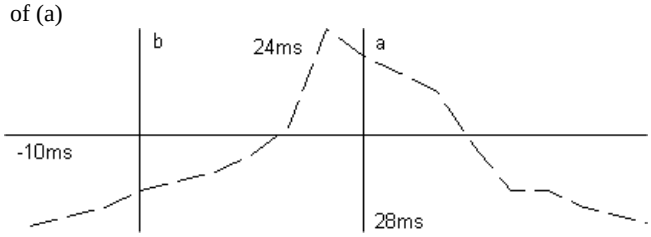

Fig. 2. Plot of log probability measure given by HMM model for boundary candidates centered at regular intervals across initial boundary(b). The peak is close to the actual boundary and is at 24 ms to right of b giving a error of 4 ms

3. TIME MARK REFINEMENT

Our goal while developing the refinement procedure was to build a data driven technique, which can be adapted to different speakers and language variations by automated training techniques. Refinement techniques such as [10][11] require manual modifications to adapt to these differences. In [11] the discriminant features for boundary edge detection are determined by studying system behavior for the different choices. In [10] a set of expert determined fuzzy logic rules are required to describe the boundary regions. In [7] a technique is described which considers the homogeneity of speech segments for a given phoneme in the same utterance to refine the boundaries. However this technique is useful only when there are multiple occurrences of the same phoneme in an utterance and there are no major prosodic variations within a single utterance. So we adopted trainable statistical models to represent the phone boundaries.

The evolution of the speech waveform near a phoneme boundary is determined by the context. The same phone boundary might have different signal characteristics for different contexts. Thus for the boundary between any two phones a and b a simple a-end b-start kind of detector will not work very accurately. Also the signal features that change sharply near the boundary depend on both the phonemes. For example, an unvoiced fricative-fricative boundary will not be accompanied with a sharp change in voicing nature unlike a boundary from a fricative to a vowel. Thus it seems logical to use separate models for each boundary based on both the phonemes defining the boundary. We model the boundary by taking signal features for equal number of small frames of speech data on both sides of the boundary location. Since the region near the boundary is rapidly evolving the frame size chosen is around 3 ms. Signal features are then calculated for these frames. We use mean energy, autocorrelation based voicing measure and MFCC's along with the respective delta measures to obtain the feature space representation for these frames.

We first used GMMs to model the features for each frame location. Thus with n frames of speech on either side of a given boundary we will have a total of $2n$ GMMs for that particular boundary model. These models are then trained using a pre-existing labeled speech corpus.

The boundary refinement is carried out by using these models to search near the approximate time marks generated by the initial segmentation process for the actual boundary point. For every initial boundary estimate we position candidate boundary points at equal spacing on either side of the boundary and calculate the signal features in the same way as the training process. Then probabilities are calculated for all frames using the corresponding GMM. The total weighted probability is taken as the match measure for that particular position. The candidate position with the maximal match measure is taken as the new boundary location.

Modeling using GMMs gave an improvement over the initial time marks but it was not very substantial because the evolution of the speech waveform near the boundary can occur at different rates i.e., sometimes the boundary is sharp and the region of transformation from one phone to another is short and in other cases it might be longer. So the one-to-one correspondence between GMMs de-

rived for the training data and the speech frames in the input waveform at the same position from the boundary is not always valid.

Thus we decided to model the boundary using HMMs, which can allow for time variability by skip state transitions. The boundary between every pair of phones was modeled as a HMM. The parameters for the HMM were estimated using a scheme similar to the GMM case. The number of states for the HMM was taken to be equal to the number of analysis frames taken in the GMM case. The models are then trained on a large phonetically labeled corpus to ensure proper coverage for different phoneme boundaries.

The refinement process is similar to the GMM case and we look for the best match (in terms of probability) in the vicinity of the initial time mark to get the refined boundary position. At each step in the search we calculate the probability of the boundary being centered at that position by matching the speech region around that position with the HMM. So if $\mathbf{P(i)}$ corresponds to log probability of the HMM model centered at $\mathbf{t^{initial} + s * i}$ matching the candidate boundary waveform with $\mathbf{t^{initial}}$ being the initial boundary estimate, $\mathbf{s}$ the search step size and $\mathbf{i}$ the frame index, then the boundary is moved to the position corresponding to the frame index

$$\mathbf{index_{boundary} = argmax(P(i))} \quad (1)$$

This is illustrated in figures 1 and 2, which show an actual example from the time refinement case. The initial estimate as generated by a phone recognizer was 28 ms off from the human labeled boundary. Log probability match between the waveform and the HMM boundary model, centered at fixed intervals of 2 ms around the initial estimate, gave a peak at 24 ms. Thus the refined boundary estimate is only 4ms off from the actual boundary.

4. SYSTEM TRAINING

The implementation was carried out using HTK[12]. A phone based CDHMM system was trained on TIMIT and was used to provide the initial labeling. For the time mark refinement experiments a corpus of 400 English utterances spoken by a professional speaker and labeled by human experts was used for training. For every phoneme boundary, we dumped a 70 ms region centered at the labeled crossover point to generate training instances for building the corresponding HMM model. Signal parameters were calculated using a small frame size(3ms) and the HMM models were trained.

Decision Tree clustering [12] driven state trying was used to ensure that the phoneme boundary models were getting properly trained. Phonemes were divided into phonetic classes such as Voiced stops, Voiced fricatives, Nasals, Liquids etc and questions were constructed so that phoneme boundary was very similar for that particular class. Since there are equal number of frames on either side of the boundary, and this number equals the states in the HMM model, the central state can be seen as dividing the HMM into left and right contexts similar to triphone clustering. States to the left of the central state are taken to represent the left phone context and the ones on the right represent, the right context.

5. RESULTS AND DISCUSSION

On the test data comprising 100 utterances by the same speaker as the training set we found the system to give a much better performance than the DTW aligner available with the festival package [8] or the commercial aligner software from ENTROPIC[6] and other phone based HMM systems built using TIMIT.

Tolerance (ms)	DTW	CDHMM	Entropic	Adapted CDHMM
4	11	14	18	19
8	30	35	41	44
16	76	78	79	83

Table 1. Percentage of time marks which lie within a tolerance region around the human labeled boundary.

Tolerance (ms)	DTW	CDHMM	Entropic	Adapted CDHMM
4	31	36	38	39
8	54	61	65	67
16	86	87	89	93

Table 2. Percentage of time marks which lie within a tolerance region around the human labeled boundary after the refinement stage.

The mismatch(in ms) from the hand labeled boundary can be taken as the error introduced by the automatic segmentation system for that particular label. The performance of an automated phoneme labeling scheme is evaluated by considering the percentage of boundary marks for which the error is less than some predefined limit. Thus if the segmentation scheme puts a certain percentage of time marks within a smaller error bound(say 4 ms) from the hand labeled boundary it performs better than a system which puts the same percentage of marks at a higher bound(say 8ms).

To evaluate the performance of the time mark refinement step, we tried some previously reported systems, which include the ENTROPIC aligner, DTW aligner from the festival package, CDHMM based on TIMIT and CDHMM adapted to the test speaker to provide different initial alignments. We tabulate the percentage of boundaries lying within a certain error bound (tolerance) in a cumulative fashion with the tolerance increasing from 4 to 16ms. Table 1 shows the error distributions obtained for different initial estimation methods. As can be seen from Table 1 the adapted CDHMM scheme provides the best performance. The percentage of boundaries with error less than 16ms is high (around 80%) and very similar for the different techniques indicating that these schemes perform well only at this high tolerance level.

In Table 2 we show the error distributions for each of these methods after refinement. From Table 2 it can be seen that the refined time marks have a much higher percentage of boundaries corresponding to the smaller tolerance limits of 4 and 8 ms. For the 16 ms tolerance limit, there is a significant improvement in performance after time mark refinement, though as can be expected it is not as pronounced.

Also, from Table 2 it can be seen that the time mark refinement stage is able to rectify initial boundary estimates with similar accuracy for the different initial labeling methods, although as can be expected an accurate initial method gives slightly better results. Figure 3 shows cumulative error distribution for the initial time marks as generated by the adapted CDHMM model and the final time marks after refinement for the same scheme.

To gain further insights on the post refinement results, we examined the cases with errors over 16ms. We found that most of the errors were due to incorrect transcription or coarticulation with the actual boundary being outside the search region for the time mark

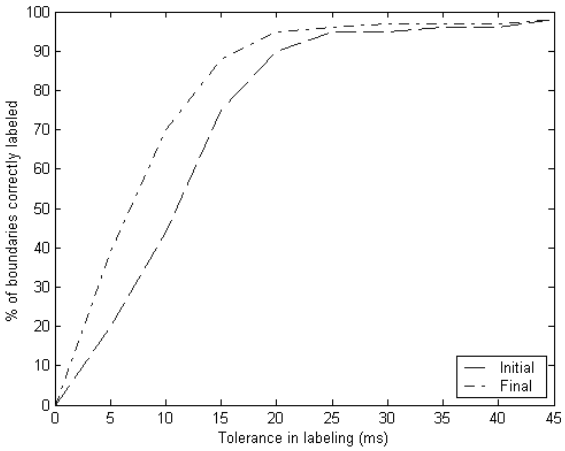

Fig. 3. Cumulative error histogram for adapted CDHMM with (dot-dashed) and without(dashed) time mark refinement.

refinement step. Thus, most of these boundary errors are extreme cases that can be corrected only if the underlying transcription is changed or the speaker takes care to speak very precisely albeit at the cost of compromising naturalness.

6. CONCLUSION

In this paper we described a two-stage speech segmentation scheme which uses HMM based boundary models to refine initial boundary estimates given by a phone recognizer. Performance analysis of this scheme indicates that this technique can give results that are very close to manual segmentation of speech. This would result in a substantial decrease in the time, and more importantly, the amount of manual intervention, required in building a new voice for a TTS system. This method is quite generic and makes no inherent assumptions on the language or speaker type.
However, further testing is required to evaluate the performance on different speakers and other languages (for multi lingual TTS). It should be interesting to evaluate the performance of time mark refinement models built from a corpus like TIMIT on other languages and also for speech corresponding to a different speaking style such as command voice, excited voice etc. It is expected that using different signal representations to build models for different boundary classes(Vowel-Vowel, Vowel-Fricative etc) should give better results. This system can be extended to include other speech features that have been developed recently to discriminate between phonemes[4].

The motivation behind developing techniques for accurate labeling of speech is to move towards automatic generation of new voices for concatenative TTS from (semi) casual speech. Another focus area for future research in this direction would be towards correcting phone level transcriptions generated from orthography. Incorrect transcriptions are the largest source of labeling errors after the time mark refinement process. This problem is more pronounced for speech from casual speakers. Casual speech has a high percentage of phone additions, deletion and substitutions along with coarticulation. We need techniques, which can not only detect mismatches from phonetic transcriptions but also attempt to correct the transcription by using restricted phoneme recognition or choosing the best alternative phoneme(s) as described by a set of rules generated from a corpus.

7. REFERENCES

[1] R.W. Sproat. "Multilingual Text-to-Speech Synthesis: The Bell Labs Approach", Kluwer, Boston, MA, 1997.

[2] A. Ljolje, J. Hirschberg and J.P.H van Santen, "Automatic Speech Segmentation for Concatenative Inventory Selection", *Progress in Speech Synthesis*, Springer 1997, pp 305-311.

[3] F. Malfrere and T. Dutiot, "High-Quality Speech Synthesis for Phonetic Speech Segmentation", *Proc. EUROSPEECH 1997*, pp 2631-2634.

[4] A. M. Abdelatty, J. Van der Spiegel, Gavin Haentjens, J. Berman and P. Mueller, "An Acoustic-Phonetic Feature-based System for Automatic Phoneme Recognition in Continuous Speech", *IEEE ISCAS*, May 1999, Proc. Vol. III, pp. 118-121.

[5] A. Ljolje and M.D. Riley, "Automatic segmentation of speech for TTS", *Proc EUROSPEECH 93* , volume 2, pp 1445-1448.

[6] C.W Wightman and D.T Talkin, "The Aligner: Text-to-speech alignment using Markov models", *Progress in Speech Synthesis*, Springer 1996, pp 313-324.

[7] Antonio Bonafonte, Albino Nogueiras and Antonio Rodriguez-Garrido, "Explicit segmentation of speech using Gaussian models", *Proc. ICSLP 96.*

[8] Alan W Black, Kevin A.Lenzo, "Building Voices in the Festival Speech Synthesis System. Processes and issues in building speech synthesis voices", *http:festvox.org/festvox/festvox_toc.html.*

[9] Matthew J Makashay, Colin W. Wightman, Ann K. Syrdal and Alistair Conkie, "Perceptual evaluation of automatic segmentation in Text-To-Speech Synthesis", *Proc. ICSLP 2000.*

[10] D. Torre Toledano, M. A. Rodrguez Crespo, J. G. Escalada Sardina, "Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules", *Proc. Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS, 1998.*

[11] Jan Ph.H van Santen, Richard W. Sproat, "High Accuracy Automatic Segmentation", *Proc. EUROSPEECH 99.*

[12] Odell J, Ollason D, Woodland P, Young S , Jansen J, "The HTK Book for HTK V2.0", Cambridge University Press, Cambridge, UK, 1995.