TOWARDS AUTOMATIC PHONETIC SEGMENTATION FOR TTS

Asaf Rendel[1], Alexander Sorin[1], Ron Hoory[1], Andrew Breen[2]

[1] Speech Technologies, IBM Haifa Research Lab, Haifa, Israel
[2] Text-To-Speech Research, Nuance Communications, Norwich, UK

ABSTRACT

Phonetic segmentation is an important step in the development of a concatenative TTS voice. This paper introduces a segmentation process consisting of two phases. First, forced alignment is performed using an HMM-GMM model. The resulting segmentation is then locally refined using an SVM based boundary model. Both the models are derived from multi-speaker data using a speaker adaptive training procedure. Evaluation results are obtained on the TIMIT corpus and on a proprietary single-speaker TTS corpus.

Index Terms— Phonetic segmentation, Phoneme alignment, Text to speech

1. INTRODUCTION

The development of a data-driven TTS voice is a complex process that starts with corpus design and recording of the target speaker. The recorded data then goes through a semi-automatic annotation process, including phonetic transcription and segmentation, followed by training and tuning of the voice models.

Generally, the annotation accuracy is an important factor in the synthesized speech quality. Therefore, voice development usually includes supervision and correction processes that require the effort of skilled personnel. Concatenative TTS systems are especially sensitive to phonetic transcription and segmentation errors. Such errors may result in an audible glitch or unintelligible speech each time an erroneous segment is selected.

The aim of our work is to improve the accuracy of the automatic phonetic segmentation process in order to eventually reduce the scale or even eliminate the manual effort without affecting the final voice quality. The phonetic segmentation approach presented in this paper includes HMM-based forced alignment followed by SVM-based local boundary refinement. Though some previous works [1][2] used the same major processing blocks, our method contains certain novel elements including HMM topology tuning, HMM adaptation strategy and local refinement features and goal function.

The rest of the paper is organized as follows. Section 2 contains a detailed description of the proposed phonetic segmentation method. Objective evaluation results are presented in section 3 with the goal of comparing the accuracy achievable by the proposed automatic process to that achieved manually. Finally in section 4, conclusions are drawn from this work.

2. METHOD

The proposed method performs phonetic segmentation by means of HMM-GMM forced-alignment, followed by SVM-based local refinement of phone boundaries. Both the HMM system and the SVM models are trained on a manually segmented and phonetically labeled corpora.

2.1. HMM alignment

The HMM alignment is carried out using the Attila speech recognition toolkit developed by IBM Research [3]. The default acoustic model configuration offered by the toolkit contains a constant number of HMM states for each phone and does not control the HMM state duration. The requirements to the acoustic model imposed by the phonetic segmentation task are not identical to those of the speech transcription task.

Duration control has been shown to improve segmentation accuracy in previous works [2][4][5]. Particularly, large segmentation errors can occur at slowly varying phone transitions due to a likelihood domination of one phone's lateral state over the other's. To reduce this phenomenon, we restrict the number of frames a lateral state can occupy. This is easily implemented by duplicating the lateral states, keeping the emission model shared between them, and replacing the self-transition loop by a transition to the next state. Figure 1 shows a modified 5-state left-to-right topology that limits the number of frames in lateral states to 4. This approach generalizes the idea exploited in [5], where the lateral states duration is limited to a single frame.

Apart from the restriction on the duration of lateral states, the number of states per phone and the frame rate imply minimal phone duration. The configuration used in this work enforces a minimal duration of 25 ms for most phones, including plosive closures, and a minimal duration

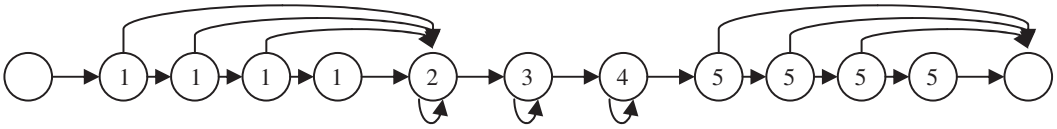

Figure 1: Modified 5-state topology example

of 15 ms for plosive bursts. However, there are phonetic contexts, in which this constraint has to be relaxed to allow shorter phones. In this work, we remove the duration constraint on plosive closures when following a stop consonant – either oral or nasal. This is achieved by allowing state skipping in the relevant phone units.

Once HMM topology is set up, a triphone-context dependent HMM-GMM model is trained. As initial features for the HMM alignment, we use 13-dimensional normalized PLP cepstra vectors extracted from 20 ms analysis window with 5 ms frame shift. The feature normalization includes cepstral mean subtraction (CMS) and cepstral variance normalization (CVN). In case of a multi-speaker training corpus, the CMS and CVN are performed on a per-speaker basis, and speaker-wise vocal tract length normalization (VTLN) frequency warping function is estimated and applied to the feature vectors. Then, at each frame, feature vectors associated with 17 consecutive frames surrounding it are concatenated, and an LDA transform is applied to project the concatenated vector down to 40 dimensions. Finally, features are adapted per speaker using feature space maximum likelihood linear regression (fMLLR).

The segmentation is determined by the usual Viterbi alignment between the given phonetic transcript and the acoustic observations sequence using the HMM-GMM acoustic model. Model space MLLR adaptation is performed prior to the alignment.

2.2. Local boundary refinement

In order to improve the precision of phone boundaries, the segmentation obtained by the forced-alignment is refined by locally adjusting phone boundary positions.

The local boundary refinement task can benefit from acoustic features with a high temporal resolution. Hence we augment the 40-dimensional basic feature vectors used in the HMM alignment with two 20-dimensional MFCC vectors extracted at a distance of ±5 ms. The MFCC vectors are CMS and CVN normalized and in the case of multi-speaker training corpus the VTLN warping transform available from the HMM alignment step is applied to them.

The final position of each boundary is set to the most likely position, in the vicinity of the original boundary, based on a phone-boundary acoustic model described below.

The phone transitions, or diphones, are clustered by a regression tree, adopting the same phonetic classes used in the construction of the HMM context tree. A multiclass SVM model [6] with a probability model [7] is trained per leaf for the following three classes: boundary, left-phone and right-phone. Training feature vectors for the negative classes (left-phone and right-phone) are extracted at offsets of ±10 ms and ±20 ms from the true boundary. Training feature vectors for the positive class (boundary) are extracted at the true boundary as well as at offsets of ±1 ms around it.

In run-time, each boundary is processed independently of the others. The search interval S is centered at the boundary and has duration of 40 ms or the minimum duration of the two surrounding phones. Feature vectors are extracted within the search interval at frames associated with time moments $\{t_k\}$. 1 ms frame shift (t_k-t_{k-1}=1ms) was used in this work. Each feature vector is fed to the SVM, and the positive class scores $P_+(t_k)$ emitted by the SVM [7] are used to calculate an error function (1) for each hypothetic boundary position t. The error function (1) may be viewed as the expected boundary estimate error if the normalized SVM confidence measures $Q(t_k)$ are considered as the probability distribution of the boundary position over the search interval. The final boundary is set at the position minimizing the error function.

$$err(t) = \sum_{t_k \in S} Q(t_k) \cdot | t - t_k |$$
$$Q(t) = \frac{P_+(t)}{\sum_{t_k \in S} P_+(t_k)} \quad (1)$$

2.3. Adaptation of the HMM-GMM model for TTS corpus

As the final step of the HMM-GMM acoustic modeling (prior to the local refinement), the HMM-GMM model can be further adapted to the target speaker data, making use of the large amount of single-speaker data available. We deal with two cases: 1) a multi-speaker manually segmented corpus is available; 2) only a limited amount of manually segmented data of the target speaker is available. The second case may be relevant to building a TTS system in a new language. In such case, a small amount of data may be manually segmented (or corrected) for the purpose of bootstrapping the acoustic model training.

In the first case, an initial HMM-GMM model is trained on the multi-speaker manually segmented corpus in a speaker adaptive manner as described above. Then, alignment is performed on the entire corpus, and the resulting segmentation is used for training of a new single-

speaker model, hereafter referred to as full target data (FTD) model.

In the second case, the initial model is trained on the limited amount of the manually segmented target data available. Then, FTD model is trained on the entire target data as in the first case. It should be noted that the portion of the manually segmented data in the training of that model is small. Hence we use the maximum mutual information (MMI) based discriminative training with I-smoothing to update the model parameters based on the manually segmented data. An approximation of the MMI objective function is used, where the denominator lattice consists of only a single path - the alignment generated by the FTD model.

3. EXPERIMENTAL RESULTS

3.1. Setup

The experiments in this paper are conducted on the TIMIT continuous speech corpus [8] and on a proprietary male US English TTS corpus. The TIMIT corpus contains 6,300 phonetically rich utterances read by 630 speakers. The standard NIST training set consists of 3,696 sentences and the test set includes 1344 utterances uttered by 168 speakers. The TIMIT phoneme set is mapped to a smaller set of 49 phones as in [9] (including the glottal stop).

The TTS corpus contains about 10K sentences, out of which 1000 sentences are manually segmented and are used for training and testing.

Segmentation accuracy is calculated by comparing the estimated boundary positions to the respective positions in the manually segmented data. Two accuracy measures are reported: 1) the percentage of the correctly estimated boundary positions within tolerance of 10 ms, 20 ms, 30 ms and 40 ms; 2) the mean absolute error (MAE) of boundary positions in milliseconds.

In the following sub-sections, we report results of three experiments and analyze the relative impact of the individual processing steps on the segmentation accuracy

3.2. Training and testing on TIMIT

The goal of this initial experiment is to assess the basic performance of our segmentation method on the TIMIT data which is widely used within the scientific community. No adaptation to the target speaker data, as described in section 2.3, is performed due to the small amount of test sentences available per speaker. The accuracy results obtained are presented in Table 1 where the first row represents the accuracy achieved by the first HMM alignment step and the second row shows the final result after the local boundary refinement.

System	Tolerance				MAE
	10ms	20ms	30ms	40ms	
HMM	80.9	94.2	97.6	98.9	7.1
+SVM	85.2	94.9	97.8	98.9	6.0

Table 1. Results on the TIMIT test set (training on TIMIT)

3.3. Training on TIMIT and testing on TTS data

The goal of this experiment is to simulate the scenario when a multi-speaker training corpus is available and no effort is invested in manual segmentation of the target TTS data. The initial training is performed using the TIMIT training set and the FTD model is built on the entire TTS data set of 10K sentences. The evaluation results obtained on the 1000 sentences TTS test set are presented in Table 2.

System	Tolerance				MAE
	10ms	20ms	30ms	40ms	
HMM	76.2	92.3	96.8	98.5	7.9
FTD HMM	75.9	93.6	97.7	99.1	7.5
+SVM	80.5	93.7	97.6	99.0	6.7

Table 2. Results on the TTS test set (training on TIMIT)

The rows of the table show the accuracy achieved by the initial speaker adaptive HMM model, the FTD HMM model and after the local boundary refinement. The results reveal a positive effect of the full target data HMM adaptation and local refinement. The first improves the accuracy at the large tolerances, and the second improves the accuracy at the small tolerances, together leading to overall improvement at all tolerances.

3.4. Training and testing on TTS data

The goal of this experiment is to simulate the scenario when no multi-speaker data is available for training, and to study the relationship between the amount of manually segmented target data and the segmentation accuracy. The available manually segmented TTS data is split into a training set and a test set, each consisting of 500 sentences. Table 3 presents the HMM alignment accuracy achieved with the initial model as a function of the training set size.

# of training sentences	Tolerance				MAE
	10ms	20ms	30ms	40ms	
50	71.3	88.3	93.5	95.7	10.7
100	77.1	91.9	96.6	98.2	7.8
200	80.1	93.5	97.5	98.9	6.8
500	81.1	94.7	98.0	99.2	6.4

Table 3. HMM alignment on TTS data with initial model trained on TTS data

Tables 4 and 5 present the HMM alignment accuracy figures achieved after the FTD and MMI adaptations respectively.

# of training sentences	Tolerance				MAE
	10ms	20ms	30ms	40ms	
50	71.8	90.4	95.8	97.8	8.8
100	76.2	92.8	97.3	98.8	7.5
200	78.4	94.1	98.0	99.3	7.0
500	79.6	94.7	98.3	99.4	6.6

Table 4. HMM alignment on TTS data with FTD adapted model

# of training sentences	Tolerance				MAE
	10ms	20ms	30ms	40ms	
50	75.2	91.7	96.6	98.4	8.0
100	78.3	93.6	97.8	99.0	7.0
200	81.1	94.7	98.2	99.3	6.4
500	81.3	95.0	98.4	99.4	6.4

Table 5. HMM alignment on TTS data with MMI adapted model

Finally we apply the local boundary refinement. The local refinement features are common to the different training set sizes as the LDA transform of the smallest training set is used in all the experiments. The accuracy figures are presented in Table 6.

# of training sentences	Tolerance				MAE
	10ms	20ms	30ms	40ms	
50	75.7	91.7	96.6	98.3	7.8
100	79.5	93.7	97.8	99.0	6.7
200	82.4	94.8	98.2	99.3	6.0
500	84.6	95.4	98.5	99.4	5.5

Table 6. Final segmentation (MMI HMM+SVM) on TTS data

The experimental results obtained under the "only target data" scenario demonstrate consistent reduction of the segmentation error level achieved by the final system compared to the HMM alignment with the initial model over all the tested training set sizes. Also the results reveal the relative influence of the consecutive intermediate processing steps which is positive in most cases.

4. CONCLUSIONS

The average segmentation accuracy achieved by the proposed system at 20 ms tolerance is within the range of inter-labeler agreement rates cited in [4] for various corpora, and is approaching the best cited inter-labeler agreement rate of 96%. The accuracy achieved on TIMIT at 20 ms tolerance is higher than previously reported in [2][4][10].

The evaluations conducted on the TTS data have revealed the contribution of the HMM adaption and local refinement to the overall accuracy.

Another important observation is that the availability of only 100-200 manually segmented sentences of the target speaker resulted in segmentation of the same average accuracy of a system trained on a large amount of multi-speaker data. This may have significant practical implication for building TTS voices for new languages.

5. ACKNOWLEDGEMENTS

The authors would like to thank Tara Sainath and Bhuvana Ramabhadran for their help with the IBM Attila speech recognition toolkit and the TIMIT database. Special thanks to Linda Thibault for her help with the TTS and manual segmentation.

6. REFERENCES

[1] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, "Automatic phonetic segmentation, " *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617- 625, 2003.

[2] J.-W. Kuo, H.-Y. Lo, and H.-M. Wang, "Improved HMM/SVM methods for automatic phoneme segmentation," *Interspeech*, pp. 2057-2060, 2007.

[3] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," *IEEE Workshop on Spoken Language Technology*, pp. 97-102, 2010.

[4] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, issue 4, pp. 352-368, 2009.

[5] K.U. Ogbureke and J. Carson-Berndsen, "Improving initial boundary estimation for HMM-based automatic phonetic segmentation," *Interspeech*, pp. 884-887, 2009.

[6] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, pp. 2:27:1-27:27, 2011.

[7] T.-F. Wu, C.-J. Lin, and R.C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *Machine Learning Research*, vol. 5, pp. 975-1005, 2004.

[8] J.S. Garofolo, L.F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM," *NIST*, 1993.

[9] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transacations on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1641–1648, 1989.

[10] D. McAllester, T. Hazan, and J. Keshet, "Direct loss minimization for structured prediction," *Neural Information Processing Systems*, pp. 1594-1602, 2010.