Refining Phoneme Segmentations Using Speaker-Adaptive Context Dependent Boundary Models

Yong ZHAO[1], Lijuan WANG[2], Min CHU[1], Frank K. SOONG[1], and Zhigang CAO[2]

[1]Microsoft Research Asia, Beijing, China
[2]Department of Electrical Engineering, Tsinghua Univ. China
{yzhao, minchu, frankkps}@microsoft.com

Abstract

Consistent phoneme segmentation is essential in building high quality Text-to-Speech (TTS) voice fonts. In this paper we propose to adapt an existing well-trained Context Dependent Boundary Model (CDBM) for refining segment boundaries to a new speaker with limited, manually segmented data. Three adaptation approaches: MLLR, MAP, and a combination of the two, are studied. The combined one, MLLR+MAP, delivers the best boundary refinement performance. In comparison with other boundary segmentation methods, the adapted CDBM yields better results, especially with a limited amount of adaptation data. Given 400 manually segmented boundary tokens in about 20 sentences as a development set, the segmentation precision can reach 90% of human labeled boundaries within a tolerance of 20 ms.

1. Introduction

State-of-the-art text-to-speech (TTS) synthesis systems are predominantly database driven, concatenation based. It is fairly straightforward to port such systems to a new voice in a well-defined, systematic procedure, without hand tuning various parameters [1]. However, the segmentation precision obtained by an automatic HMM-based forced-alignment procedure is still not good enough to warrant high quality synthesis. A post-refinement, manual or automatic, is always needed to adjust unit segmentations [2][3][4].

Most studies on segmentation refinement are based upon a large single speaker TTS corpus. The speaker adaptation aspects of segmentation refinement have not been well studied. Furthermore, in these approaches, segmentation accuracy is generally improved at the expense of creating a significant amount of manually labeled segment boundaries for training. To prepare many such manually labeled boundaries is time consuming and poses a bottleneck when rapid prototyping of a new voice font. To facilitate a fast and high quality personalized TTS, we need to minimize the manual segmentation effort.

We have investigated the issue on how to make use of only a small set of manually segmented and labeled boundaries to improve segmentation accuracy in a speaker-dependent mode. In this paper we extend our work of speaker-dependent Context Dependent Boundary Model (CDBM) [5] to a speaker-adaptive one. In this extended approach there is no need of training a new CDBM from scratch, so the work of manual segmentation and checking is greatly reduced while a high performance of boundary refinement is maintained.

The rest of the paper is organized as follows: In Section 2 we review the CDBM model. In Section 3 we present how to adapt a CDBM to perform high-precision automatic segmentation refinement. In Section 4 we present our experiments and corresponding results. In Section 5 we give our analysis of results and discussions.

2. Context Dependent Boundary Model (CDBM)

In an earlier paper [5], we proposed to construct CDBMs for automatic refinement of segment boundaries. The approach was motivated by observing that spectral dynamics across a segmental boundary point is conditioned upon its left and right phoneme contexts. Thus, to build separate boundary models by efficiently clustering these boundaries into subgroups would be beneficial for segmentation accuracy.

A context-dependent boundary is denoted as X-B+Y, where B denotes the boundary; X, its immediate left phoneme; and Y, its immediate right phoneme. In order to characterize a specific boundary point, acoustic spectral features are collected from frames around the labeled segmentation points and a sequence of GMMs is trained.

In order to make robust parameter estimates, Classification and Regression Tree (CART) [7] are used to cluster acoustically similar GMMs into broad classes. The clustering procedure and the node-splitting question sets raised in building the regression tree are similar to triphone building in speech recognition [6].

Once such CDBMs are trained, the segmentations are refined by finding a point, in the vicinity of the tentative boundary (obtained by HMM forced alignment), which yields the maximum likelihood by measuring a long window of speech data against the trained CDBM. The performance of CDBM has been evaluated on a Mandarin speaker-dependent corpus. The agreement reached more than 90.0% to the human labeled boundaries within a tolerance of 20 ms, when a large amount of manual segmentations of about 250 sentences are provided to train CDBMs.

3. CDBM adaptation

The CDBM boundary refinement method uses a large speech database of a single speaker with some manually labeled boundaries. However, whether or not it works under a speaker adaptive mode has not been well studied. Furthermore, compared to other state-of-the-art segmentation methods, CDBM possibly requires more training data to obtain a decent performance, though it is believed to deliver superior performance as training data increases.

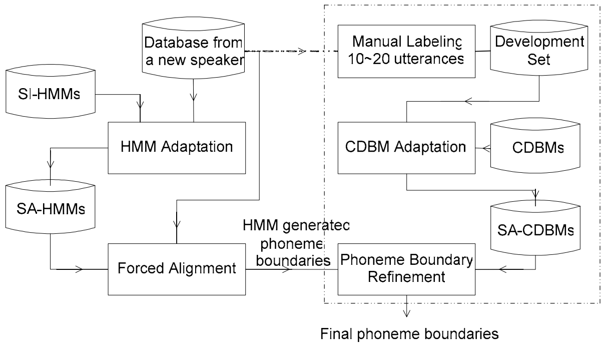
Database from a new speaker
Manual Labeling 10~20 utterances
Development Set
SI-HMMs
HMM Adaptation
CDBM Adaptation
CDBMs
SA-HMMs
SA-CDBMs
HMM generated phoneme boundaries
Forced Alignment
Phoneme Boundary Refinement
Final phoneme boundaries

Fig.1. Scheme for the CDBM adaptation

To reduce manual labeling effort in prototyping a new voice font, we investigate to extend the speaker-dependent CDBM (SD-CDBM) to a speaker-adaptive (SA-CDBM) one. The procedures for CDBM adaptation in a two-step segmentation refinement framework are illustrated in Fig.1. In the 1st step of coarse segmentation, HMM models are trained with the entire speech data. Then, according to the given phonetic transcriptions, HMM sequences are aligned with the corresponding speech to generate the tentative segmentation boundaries. In the 2nd step of boundary refinement, a segment boundary is refined by finding a point, in the vicinity of the tentative boundary, which yields the maximum likelihood by matching a long window of speech data against the trained CDBM.

The refinement process is similar to the originally proposed CDBM construction, except that it involves two additional adaptation phases to bridge the acoustic differences between the original model and the new speaker. The first one is to update HMM parameters for forced alignment; the second, to modify CDBM parameters for phoneme boundary post-refinement. With respect to the former issue, we follow the speaker adaptation used in speech recognition. The entire speech corpus from the target speaker is used for HMM adaptation, thus, the adapted models are believed to be reliable. In this paper, how to make CDBM adaptation will be focused.

As proposed in [5], a boundary was modeled by extracting (2N+1) frames of features from a time span centered at the labeled boundary points, see Fig. 2. Each frame is modeled by a GMM. Overall, (2N+1) GMMs are constructed to make one CDBM. The sequence of GMMs can be viewed as a (2N+1)-state HMM, where each state corresponds to one frame and the transition coefficients between adjacent neighboring states are set to 1, i.e., neither looping nor skipping is allowed. With such a model structure, CDBM adaptation can be easily extended from the standard speaker adaptation algorithms.

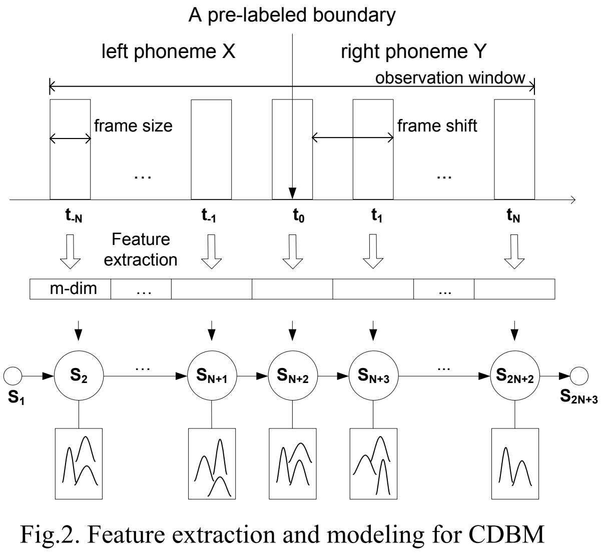

Fig.2. Feature extraction and modeling for CDBM

Three speaker adaptation approaches are investigated. They are Maximum a Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR), and a combination of the two, denoted as MLLR+MAP.

For MAP adaptation, the re-estimation formula for the Gaussian mean is a weighted sum of the prior mean with the ML mean estimate [7]:

$$\hat{\mu}_{ik} = \frac{\tau_{ik}\mu_{nw_{ik}} + \sum_{t=1}^{T}\zeta_t(i,k)x_t}{\tau_{ik} + \sum_{t=1}^{T}\zeta_t(i,k)} \qquad (1)$$

where τ_{ik} is the weighting parameter for the kth Gaussian component in the corresponding frame i, or state i. $\zeta_t(i,k)$ is the occupation likelihood of the observed adaptation data t.

For MLLR estimation, the kth mean vector μ_{ik} for each frame i can be transformed using the following equation [7]:

$$\tilde{\mu}_{ik} = A_c \mu_{ik} + b_c \qquad (2)$$

where A_c is a regression matrix and b_c is an additive bias vector associated with the broad class c.

When MLLR method is combined with MAP, we can benefit from both the compact MLLR transformations for rapid adaptation when only limited data is available and the asymptotically efficient properties of MAP adaptation as training data increases. There are a number of different ways to combine MLLR and MAP to improve performance. We found that using MLLR to transform frame means first and using MAP to locally modify the parameters that are observed in the adaptation data yields the best result.

4. Experiments and results

4.1. Speech corpora

Two Mandarin Chinese TTS speech corpora, CH-DB1 and CH-DB2, are used in our experiments. Both corpora are read by professional female speakers and contain roughly 12,000 sentences, or a total of 180,000 syllables. The syllable segmentation boundaries have been checked manually by experienced annotators with consistent guidelines. 20,000 syllable boundaries from CH-DB1 serve as the starting point for speaker adaptation. 20,000 syllable boundaries from CH-DB2 are used as the development set for adaptation. Additional 10,000 boundaries are used for testing.

In addition to the above two, four small corpora, db1, db2, db3, and db4, recorded by four non-professional speakers are also used for examining whether the algorithm is applicable to ordinary speakers. Each of the 4 databases contains of 200 utterances, about 4,500 syllables in total. All syllable boundaries in the four corpora have been manually checked with the same criteria as above. 2,500 boundaries from each are used for development and the other 2,000 boundaries for testing.

Refined boundary points are compared with hand-labeled boundaries. If the difference is within a pre-defined threshold, the auto-generated boundary is counted as correct. The correct percentage is used for measuring segmentation accuracy. Here, we use a tolerance of 20ms.

4.2. Experimental results

The original CDBM models trained from 20,000 boundaries in CH-DB1 are used as the initial models for adaptation. They are adapted to CH-DB2 with the corresponding developing set. Three adaptation schemes, MLLR, MAP and MLLR+MAP, are investigated and the corresponding adaptation performances are shown in Fig. 3. When adapted with ~200 boundary tokens, all three speaker-adaptive models without adaptation outperform the original baseline system.

As shown in the same figure, MLLR adapts faster than MAP when the developing set is small, whereas MAP becomes asymptotically more accurate than MLLR when the size of developing set increases. The crossover point is around 600 adaptation tokens. When MLLR is combined with MAP, not only faster adaptation is obtained, but also better performance over either MLLR or MAP is achieved. We use MLLR+MAP in the following experiments. Also, it is interesting to note that the trends of all three adaptations are very similar to what have been observed in the speech recognition [7].

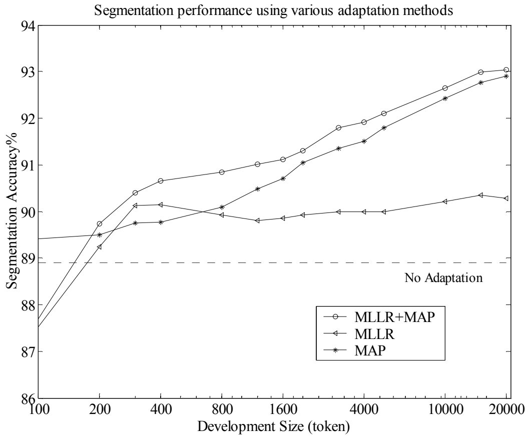

Fig.3. Performance comparison of MLLR, MAP and MLLR+MAP

To compare the performance of speaker-adaptive and speaker-dependent CDBMs, we have tested both models on CH-DB2. The speaker adaptive models are adapted as described above, and the speaker-dependent ones are trained with the development set in CH-DB2. The performances of both models on refining CH-DB2 are shown in Fig. 4. Also included as references are results from HMM forced-alignment, and the un-adapted system trained on CH-DB1.

The figure indicates that when the size of training set is limited, say below 2,000 tokens, speaker-adaptive system performs better than speaker-dependent one. The experiment is repeated on CH-DB1 and similar results are obtained.

To examine the generalization capability of the algorithm, we repeat the above experiments on the four small corpora db1, db2, db3, and db4. The original CDBM is trained on CH-DB1. Four speaker-adaptive CDBMs are obtained by adapting the original CDBM to the four small corpora individually. The segmentation accuracies of four corpora are averaged and presented in Fig. 5. As shown in the figure, similar results are observed, with accuracy only slightly inferior to that of the two professional corpora. With 400 manually labeled boundary tokens, or ~20 utterances, the speaker-adaptive CDBMs achieve 90% of segmentation accuracy in average, on the other hand, the speaker-dependent system needs about 10 times of training data to achieve a similar level of segmentation accuracy.

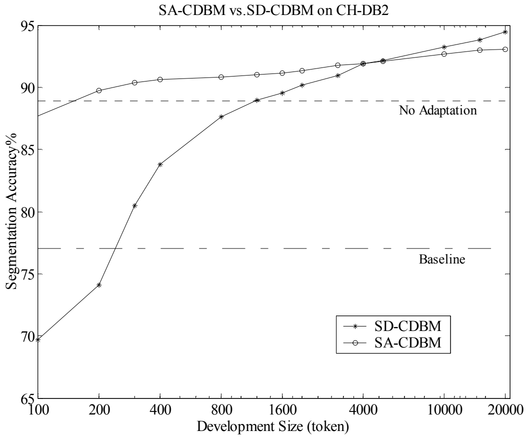

Fig.4. Performance comparison between SD-CDBM and SA-CDBM on CH-DB1

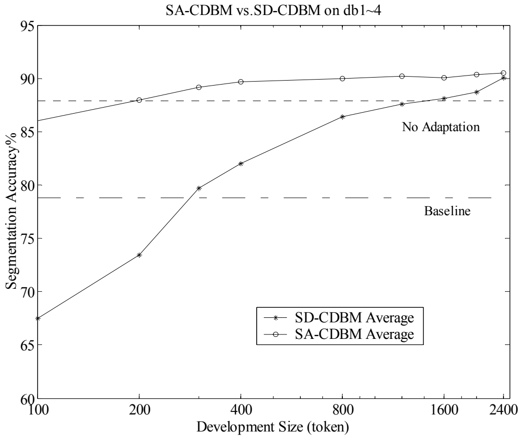

Fig.5. Average refinement performance comparison between SA-CDBM and SD-CDBM.

5. Analysis and Discussion

In this paper we propose to adapt a well-trained context dependent boundary model to segment the TTS speech database of a new speaker, by using only a very small set of hand-labeled segment tokens. The speaker adaptive CDBM trained with only 1/10 of manual data needed in a speaker-dependent system can achieve 90% segmentation accuracy.

Detailed analysis of the refined boundaries has shown that boundaries with sonorant phonemes on either side achieve a noticeable segmentation improvement after adaptation. This may be due to the fact that sonorant sounds are more speaker-dependent and can be adapted more effectively to a new speaker. Also, boundaries with stops or fricatives can achieve high refinement accuracy even without adaptation. The proposed adaptation approach requires boundaries to be labeled consistently across different corpora, since it only takes into account mismatch between speaker characteristics, but not the annotation criteria. Future work will be focused on detecting gross segmentation or labeling errors in the TTS unit inventory with a reliable statistical confidence measure.

6. References

[1] M. Chu, H. Peng, H.Y. Yang, and E. Chang, “Selecting non-uniform units from a very large corpus for concatenative speech synthesizer”, in *Proc. ICASSP-2001*, 2001. Salt Lake City, USA, 2001.

[2] A. Sethy and S. Narayanam, “Refined speech segmentation for concatenative speech synthesis,” in *Proc. ICSLP-2002*, pp.145-148, Denver, Colorado, USA, September 2002.

[3] D. T. Toledano and Luis A. Hernández Gómez, “Automatic phonetic segmentation”, IEEE *Trans. Speech and Audio Processing*, vol. 11, pp.617-625, November 2003.

[4] J. Adell and A. Bonafonte, “Towards phone segmentation for concatenative speech synthesis,” in *Proc. 5th ISCA Speech Synthesis Workshop*, Carnegie Mellon University, June 2004.

[5] L.J. Wang, Y. Zhao, M. Chu, J.L. Zhou and Z.G. Cao, “Refining segmental boundaries for TTS database using fine contextual-dependent boundary models,” in *Proc. ICASSP-2004*, pp.641-644, Quebec, Canada, May 2004.

[6] J. Odell, D. Ollason, P. Woodland, S. Young and J. Jansen, “The HTK Book for HTK V3.0,” Cambridge University Press, Cambridge, UK, 2001.

[7] X. D. Huang, A. Acero and H. Hon, Spoken Language Processing, Prentice Hall, New Jersey, 2001.