

AUTOMATIC SEGMENTATION COMBINING AN HMM-BASED APPROACH AND SPECTRAL BOUNDARY CORRECTION

Yeon-Jun Kim, Alistair Conkie

AT&T Labs – Research and AT&T Natural Voices
180 Park Ave., Florham Park, NJ 07932-0971, USA

ABSTRACT

Currently, AT&T Labs' *Natural Voices* multilingual TTS system produces high-quality synthetic speech with a large-scale speech corpus [1]. In the development of such systems, automatic segmentation constitutes a major component technology.

The prevalent approach for automatic segmentation in speech synthesis is *Hidden Markov Model* (HMM) - based. Even though an HMM-based approach is the most automatic and reliable, there are still several limitations, such as mismatches between hand-labeled transcriptions and HMM alignment labels which can lead to discontinuities in the synthetic speech, or the need for hand-labeled bootstrap data in HMM initialization. This paper introduces a new approach to automatic segmentation which aims both to minimize human intervention and to achieve a higher segmental quality of synthetic speech in unit-concatenative speech synthesis, by combining a conventional HMM-based approach and spectral boundary correction. A preference test demonstrates the proposed method is effective in reducing discontinuities in synthetic speech.

1. INTRODUCTION

As automatic segmentation saves a lot of human effort and time in training and building speech inventories, it has become quite important for both speech recognition and speech synthesis research, where the amount of speech data to be processed is becoming larger and larger. This paper describes automatic segmentation for speech synthesis in its function as a component technology of AT&T Labs' *Natural Voices* TTS systems.

HMM-based approaches adopted from automatic speech recognition (ASR) are most widely used for automatic segmentation in speech synthesis, providing a *consistent* and *accurate* phone labeling scheme [2]. Both consistency and accuracy are critical for building a speech inventory for a TTS system based on the principles of unit-selection and concatenative synthesis that produces intelligible and natural sounding speech. Even though the effort to adapt an HMM-based approach for the purpose of speech synthesis

has been successful, there is still room for improvement regarding degree of *automation* and *accuracy*.

Firstly, a major focus in TTS research currently is to reduce the time and cost for building an inventory of speech units, especially with respect to the increasing demand for more synthetic voices, including customized ones. Much of that work has had to be done manually so far, obviously slowing down the process of building synthesis inventories. For example, hand-labeled bootstrapping may require a month of labeling by a phonetic expert to prepare training data for speaker dependent HMMs which provide quite accurate phone segmentation results. On the other hand, automatic segmentation using speaker-independent HMMs bootstrapping reduces the manual workload considerably while at the same time keeping HMMs stable (see section 2).

Secondly, the accuracy of automatic segmentation determines to a large degree the quality of speech in synthesis using unit selection and concatenation. As stated, e.g., in [3], an HMM-based approach is somewhat limited in its ability to remove discontinuities at concatenation points, because the Viterbi alignment used in an HMM-based approach tries to find the best HMM sequence when given a phone transcription and a sequence of HMM parameters, not the optimal boundaries between adjacent phones. The system, therefore, may locate a phone boundary at a different position than expected, causing mismatches at unit concatenation points and resulting in discontinuities. In section 3, we propose a method combining spectral boundary correction with an existing HMM-based approach to reduce the chance of such discontinuities.

2. AUTOMATIC SEGMENTATION USING AN HMM-BASED APPROACH

An HMM-based approach generally consists of two phases; training HMMs, and unit segmentation using the Viterbi alignment. As in ASR, each phone is defined as an HMM prior to unit segmentation, and then trained with a given phonetic transcription and its corresponding feature vector sequence. The biggest difference between automatic segmentation for ASR and for TTS concerns the degree of re-

quired accuracy.

HMM tuning methods for improving accuracy, such as considering phonetic context (context-dependent HMMs) or applying a multiple mixture Gaussian density, have already been studied in numerous works, e.g., [4] and are now generally accepted. However, one important remaining question is how to choose initial estimates of the HMM parameters such that the local maximum is as close as possible to the global maximum of the likelihood function.

2.1. Bootstrap with speaker-independent HMMs

As shown in Fig.1, there are three possible ways to achieve the initialization of a set of HMMs, 1) *hand-labeled bootstrap*, 2) *speaker-independent (SI) HMM bootstrap*, and 3) *flat start*. Of these, hand-labeled bootstrap, which is used to segment a specific speaker’s hand-labeled speech data, results in the most accurate HMM modeling (so called speaker-dependent HMMs - SD HMMs). SD HMMs are generally used for automatic segmentation in speech synthesis, but do have the disadvantage of being quite time-consuming to prepare. If hand-labeled speech data is available for a particular language, but not for the intended speaker, bootstrapping with SI HMMs alignment is the best alternative.

In the work described in this paper, SI HMMs for American English, trained with the TIMIT speech corpus, were used in the preparation of seed phone labels. With the resulting labels, SD HMMs for an American male speaker were trained to provide the segmentation for building an inventory of synthesis units (Fig. 1). An advantage of bootstrapping with SI HMMs is that all the available speech data can be used as training data, if desired.

Our automatic segmentation system for American English consists of ARPA phone HMMs that are commonly using three-state left-to-right models with multiple mixture of Gaussian density. We use standard HMM input parameters; 12 MFCCs (Mel frequency cepstral coefficients) and normalized energy, and their first and second order delta coefficients.

As tested in 100 randomly chosen sentences, the SD HMMs bootstrapped with SI HMMs lead to phones being labeled with an accuracy of 87.3% (<20ms, compared to hand labeling). Many errors are caused by differences between the speaker’s actual pronunciations and the given pronunciation lexicon, i.e., errors by the speaker or the lexicon or effects of spoken language such as contractions. Therefore, speaker-individual pronunciation variations have to be added to the lexicon.

2.2. Iterative HMM training

The motivation for iterative HMM training is that more accurate initial estimates of the HMM parameters produce more accurate segmentation results, as shown for example in [5].

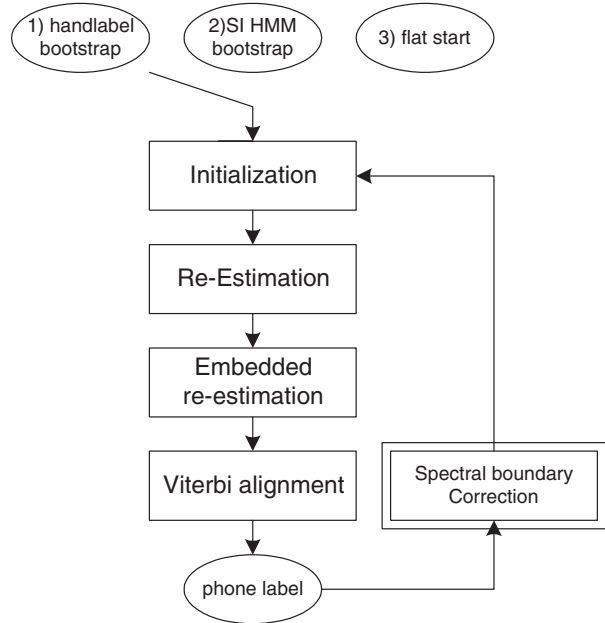


Fig. 1. Automatic segmentation procedure which combines an HMM-based approach with iterative training and spectral boundary correction

The phone labels that result from bootstrapping with SI HMM are more accurate than the original input (seed phone labels). For this reason, for tuning the SD HMMs to produce the best results, the phone labels resulting from the previous iteration are used as the input for HMM initialization and re-estimation, as shown in Fig. 1. This procedure is iterated to fine-tune the SD HMMs.

After several rounds of iterative training, mismatches between manual labels and phone labels assigned by an HMM-based approach were shown to be considerably reduced in the 100 test sentences. For example, when the HMM training procedure is iterated 5 times, an accuracy of 93.1% was achieved (<20ms, compared to hand labeling), yielding a noticeable improvement in synthesis quality.

The main question concerning this procedure is how many iterations are needed to achieve high-quality synthetic speech. The accuracy of phone labeling in a few speech samples alone cannot predict synthetic quality itself. The stop condition for our iterative training, therefore, is defined as the point when no more perceptual improvement of synthesis quality can be observed.

3. SPECTRAL BOUNDARY CORRECTION

In the previous section, it was shown that an HMM-based approach produces consistent phone labeling results, and that feedback from the previous training results in the iterative training procedure reduces the mismatches between

hand labeling and labels assigned by an HMM-based approach.

As a result, we can infer that a reduction of mismatches is to be expected when the temporal alignment of the feedback labeling is corrected. Phone boundary corrections can be done manually or by rule-based approaches, however, this conflicts with our goal to minimize human intervention. On the other hand, *assuming that phone labels assigned by an HMM-based approach are relatively accurate*, automatic phone boundary correction concerning spectral features is a promising alternative.

The procedure proposed in this paper combines an HMM-based approach and spectral boundary correction. (See Fig.1)

3.1. Spectral boundary detection measure

Our research goal is to minimize audible signal discontinuities caused by spectral mismatches between two successive concatenated units. In unit-concatenative speech synthesis, a phone boundary can be defined as the position where the maximal concatenation cost concerning spectral distortion, i.e. *spectral boundary*, is located.

The Euclidean distance between MFCCs is most widely used to calculate spectral distortions. As we already used MFCC in the HMM-based segmentation, we adopted instead the *weighted slope metric* (see Eq.(1)), as proposed by Klatt [4], to catch possible mistakes overlooked by the MFCCs.

$$d(S^L, S^R) = u_E |E_{S^L} - E_{S^R}| + \sum_{i=1}^K u(i) [\Delta_{S^L}(i) - \Delta_{S^R}(i)]^2 \quad (1)$$

In our implementation, S^L , S^R are 256 point FFTs divided into K critical bands. The S^L and S^R vectors represent the spectrum to the left and the right of the boundary, respectively. E_{S^L} , E_{S^R} are spectral energy, $\Delta_{S^L}(i)$ and $\Delta_{S^R}(i)$ are the i th critical band spectral slopes of S^L and

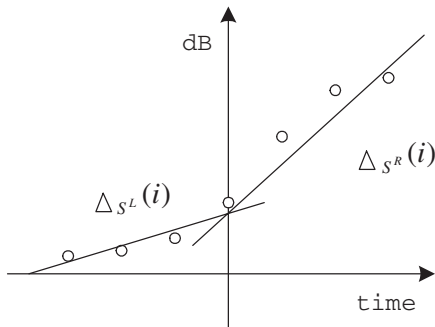


Fig. 2. Phone boundary detection using spectral transition measure

S^R (see Fig. 2), and u_E , $u(i)$ are weighting factors for the spectral energy difference and the i th spectral transition.

Spectral transitions are also believed to play an important role in human speech perception. The local maximum of $\sum_{i=1}^K u(i) [\Delta_{S^L}(i) - \Delta_{S^R}(i)]^2$, i.e., the *bending point* of spectral transition, often coincides with a phone boundary.

In this study, $|E_{S^L} - E_{S^R}|$, which is the absolute energy difference in Eq.(1), is modified to distinguish K critical bands, as in Eq.(2). This is because each phone boundary is characterized by energy changes in different bands of the spectrum.

$$|E_{S^L} - E_{S^R}| = \sum_{j=1}^K w(j) * |E_{S^L}(j) - E_{S^R}(j)| \quad (2)$$

where $w(j)$ is the weight of the j th critical band.

3.2. Phone class-dependent time window

Although there is a strong tendency for the largest peak to occur at the correct phone boundary, the automatic boundary detector described in 3.1 is likely to produce a number of spurious peaks. For this reason, the decision on the location of phone boundary is somewhat risky without any human intervention.

To minimize this chance of mistakes in the automatic correction, we use a *phone class-dependent time window* in which the optimal phone boundary is more likely to be found. The phone boundary is checked only within the specified time window.

As shown in [2] and [6], temporal misalignment, as compared to hand labeled segmentation, tends to vary in time (ms) depending on phone class of the two successive phones. Therefore, the time window for finding the local maximum of spectral boundary distortion is empirically determined by the adjacent phones (Table 1).

By means of this spectral boundary correction, we have achieved an accuracy of 94.8% (<20ms, as compared to hand labeling) and less noticeable discontinuities in synthetic speech.

BND	time window	BND	time window
V-V	-4.5 ± 50	P-V	-1.6 ± 30
V-N	-4.8 ± 30	N-V	0 ± 30
V-B	-13.9 ± 30	B-V	0 ± 20
V-L	-23.2 ± 40	L-V	11.1 ± 30
V-P	2.2 ± 20	S-V	2.7 ± 20
V-Z	-15.8 ± 30	Z-V	15.4 ± 40

Table 1. Phone-class dependent time window (in ms) for spectral boundary correction (V: Vowel, P: Unvoiced stop, B: Voiced stop, S: Unvoiced fricative, Z: Voiced fricative, L: Liquid, N: Nasal)

4. PERCEPTUAL EXPERIMENT

The proposed approach was evaluated using the AT&T Labs' *Natural Voices* TTS system based on the unit selection approach. A preference test was conducted with stimuli of speech synthesized with different approaches to automatic segmentation:

iterative training only The traditional HMM training approach using an iterative training procedure.

iterative training + \bar{X} correction The labels assigned by an HMM-based approach are shifted by the average offset (\bar{X}) between hand labeling and labels assigned automatically. These shifted labels are then used as feedback for iterative training.

spectral correction after HMM-based segmentation After a speech corpus has been labeled by HMM-based segmentation, the labeled data are re-aligned by spectral boundary correction, but *not* used for iterative training.

iterative training + spectral correction The labels assigned by an HMM-based approach are re-aligned by means of spectral boundary correction. These re-aligned labels are then used as feedback for iterative training.

All the approaches listed above were trained and used for automatic segmentation with the same 8 hours of recorded speech from one male speaker. Seventeen sentences originally designed for the purpose of speaker selection [7] were used as test stimuli. They include difficult pronunciations that challenge not only the synthesis system, but human readers as well. The 10 listeners who participated in the informal test were all employees or contractors working at AT&T Labs Research, and included both native and non-native speakers of *American English*. Listeners were asked to listen to the test sentences and rank order the versions for each sentence according to their subjective preferences. The ranked preference test result for comparing various approaches is shown in Table 2.

Segmentation Methods	Rankings		
	1st	2nd	3rd
Iterative training only	7	12	46
Iterative training + \bar{X} correction	48	69	43
Spectral correction after HMM-based segmentation	11	52	42
Iterative training + Spectral correction	104	37	29

Table 2. Result of the preference test with various correction approaches (The number of iteration was kept constant at 5.)

5. SUMMARY AND CONCLUSIONS

This paper has presented an approach to automatic segmentation that combines two different methods of segmentation, one HMM-based, the other using spectral features.

As the observation of phone labeling in the randomly sampled sentences shows, the proposed approach helps to reduce the misalignments between target phone boundaries and boundaries assigned by automatic segmentation. The preference test also shows that the proposed approach produces more natural synthetic speech, i.e. concatenation points are less noticeable.

The improvements achieved by this approach are very encouraging and a useful addition to HMM-based segmentation. The proposed approach is now being evaluated for other speakers and other languages and results will be reported in the near future.

Synthesis samples using this approach can be found at http://www.naturalvoices.att.com/demos/rcvd_us_m.html#rich

6. REFERENCES

- [1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," in *Proc. Joint Meeting of ASA, EAA and DEGA*, 1999.
- [2] Matthew J. Makashay, Colin W. Wightman, Ann K. Syrdal, and Alistair Conkie, "Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis," in *Proc. ICSLP 2000, Beijing*, 2000, pp. 431–434.
- [3] Jan P. H. van Santen and Richard W. Sproat, "High-Accuracy Automatic Segmentation," in *Proc. of EUROSPEECH'99*. Budapest, Hungary, 1999.
- [4] Lawrence Rabiner and Bing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [5] Pedro Carvalho, Isabel Trancoso, and Luís Oliveira, "Automatic Segmentation Alignment for Concatenative Speech Synthesis in Portuguese," in *Proc. RECPAD'98 - 10th Portuguese Conference on Pattern Recognition*. Lisboa, 1998.
- [6] Andrej Ljolje, Julia Hirschberg, and Jan P.H. van Santen, *Automatic Speech Segmentation for Concatenative Inventory Selection*, chapter 24, pp. 305–311, Springer, 1996.
- [7] A. Syrdal, A. Conkie, and Y. Stylianou, "Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis.," in *Proceedings of ICSLP '98*, 1998.