

PERCEPTUAL EVALUATION OF AUTOMATIC SEGMENTATION IN TEXT-TO-SPEECH SYNTHESIS

*Matthew J. Makashay*¹ *Colin W. Wightman*² *Ann K. Syrdal* *Alistair Conkie*

AT&T Labs - Research, Florham Park, NJ, U.S.A.

¹also Dept. of Linguistics, Ohio State University, Columbus, OH, USA

²also Dept. of Computer and Information Sciences, Minnesota State University, Mankato, MN, USA

ABSTRACT

The quality of a concatenative text-to-speech (TTS) system is directly related to the accuracy with which the underlying acoustic inventory is labeled. An open issue has been the tradeoffs between the speed of automatic labeling, the accuracy of hand labeling, and the quality of the resulting synthesis. The goal of our study was to perceptually compare the impact of different phonetic segmentation techniques on TTS synthesis.

A concatenative TTS system using an acoustic inventory phonetically segmented and time aligned by each of six different methods was subjectively evaluated in a listening test. A 90 minute acoustic inventory of speech read by a female professional speaker was segmented and aligned by each of four different methods: (1) expert hand segmentation (HS), and automatic HMM-based segmentation by means of either (2) speaker independent monophone (SIM) models, (3) speaker dependent monophone (SDM) models, or (4) speaker dependent cross-word triphone (CWT) models. Two additional systems evaluated used variations of CWT segmentation: (5) CWT-HS used an alignment half-way between the CWT and hand segmentations; (6) CWT-OJ used an algorithm to search within a range centered on the CWT phone boundary for an optimal join point between concatenated units[3]. Except for the segmentation and alignment of the acoustic inventory, the TTS system[1] used for synthesis was identical in all respects.

The three systems that used variations of CWT models were rated significantly higher than all other synthesizers, including the hand segmented system. Neither reducing alignment differences between CWT labels and manually assigned labels nor using an optimal join point algorithm improved upon unmodified CWT synthesis quality.

1. INTRODUCTION

Concatenative unit selection TTS synthesis, used by the AT&T NextGen system, requires very accurate labeling of the acoustic inventory from which units (such as phones, diphones, or half-phones) are selected and concatenated to

create synthetic speech. For example, the closure and release burst portions of stop consonants must be accurately labeled within a few milliseconds to avoid audible errors such as no closure, no burst, or double bursts.

The slow pace of manual labeling often creates a bottleneck in TTS development. Even well trained, experienced phonetic labelers working with a familiar voice using efficient speech display and editing tools on a modern workstation require about 200 times real time to segment and align speech utterances. A unit selection TTS inventory that contains two hours of speech, for example, would require 50 8-hour days to label by hand. Whenever manual labeling is involved, especially if several transcribers are used, there is also a problem of consistency.

Automatic phonetic classification using speech recognition (ASR) techniques is very fast and very consistent, and therefore attractive for use in labeling a TTS inventory. However, the accuracy of automatic segmentation and alignment is questionable for TTS purposes. The phone set used by the ASR system must match (or map to) the set used in synthesis. ASR models of breaths, lip smacks, and other nonspeech sounds (that are invariably present in long continuous utterances) are needed to label extraneous sounds and to avoid misclassifying them as speech. Even if the phones and nonspeech noises are correctly classified, their time alignment with the utterance must also be very precise. Because errors are audible, accuracy requirements are much more stringent for phonetically labeling a TTS inventory than they are for an ASR database.

1.1. Background

Numerous studies of automatic segmentation and alignment have been conducted over the past 20 years. Most of the research has been focused on labeling speech data for the training and testing of ASR systems, and some of the studies have used TTS modules to generate phonetic transcriptions or synthetic utterances as references. Techniques have included acoustic alignment to a synthetic reference utterance[7] and alignment directly to phonetic symbols using phoneme recognition[8][4]. These studies usually eval-

uated their results by objective comparisons of automatic segmentation and alignment to hand labeled references.

There are several recent studies in which an acoustic inventory for concatenative synthesis was phonetically segmented and aligned using HMMs. The Entropic “Aligner”[9] used context independent HMMs trained on 614 speakers from the TIMIT speech database plus an on-line dictionary to produce time-aligned transcriptions. Using context dependent HMMs trained on the multi-speaker TIMIT speech database, 80% accuracy within 17 ms of manually obtained boundaries was achieved[5] using automatically derived transcription during testing. For HMMs trained on a single speaker, the best results in [6] obtained 80% accuracy within 12 ms of manual boundaries. For the same corpus, manual boundaries assigned by two different labelers had 80% agreement within 8 ms, which serves as an index of human performance.

Some TTS systems rely entirely on automatic segmentation of their acoustic inventories, but they also use automatic error detection or correction methods. Microsoft’s trainable TTS system, Whistler[2], prunes statistical outliers from the acoustic inventory and then uses HMM scores to select a small number of optimal instances for each unit. ATR’s TTS system CHATR uses automatic methods to label its unit selection speech inventory, but it attempts to correct for possible alignment error by looking for the best join point in the vicinity of the unit boundary at run time[3].

Some TTS researchers strongly advocate the manual segmentation of TTS inventories because they doubt that automatic methods are currently capable of the precision required.

In the current study, we statistically compare different automatic segmentations with manual segmentation, but we use empirical perceptual results to identify which of several techniques including manual segmentation results in the best synthesis when applied to a TTS inventory.

2. EXPERIMENTAL METHODS

2.1. Segmental Labeling Conditions

A 90 minute acoustic inventory of speech read by a female professional speaker was segmented and aligned by each of four different methods: (1) expert hand segmentation (HS), and automatic HMM-based segmentation by means of either (2) speaker-independent monophone (SIM) models, (3) speaker-dependent monophone (SDM) models, or (4) speaker-dependent cross-word triphone (CWT) models. Two additional experimental systems used variants of CWT: (5) CWT-HS used an alignment half-way between the CWT and HS alignments, and (6) CWT-OJ used an algorithm to search within a range of the CWT phone boundary for an optimal join point between concatenated units[3]. Except for CWT-OJ, a synthesizer was constructed with the acoustic inventory segmented and aligned by each of the above methods, and test stimuli were generated by a method that

used the exact segment boundaries determined by its respective labeling method. The CWT-OJ system used the same acoustic inventory as CWT, but with an optimal join synthesis technique.

HS: Expert Hand Segmentation was conducted on the entire 90 minute speech database. This process consisted of taking the output from the Aligner program and manually correcting it. The Aligner segmentation for each utterance was examined using the ENTROPIC waves+ program, and where discrepancies in either the transcription or boundary alignment were found, they were corrected. Several labelers knowledgeable in acoustic phonetics corrected the data. Labeling procedures were adopted so that consistency across labelers was maintained as far as possible. A final check was carried out with one person examining all the data for errors and variations in segment boundary placement.

Three automatic aligners generated time-aligned phonetic transcriptions from text and speech input. All used an on-line dictionary to obtain possible pronunciations of words in the text, and selected the best pronunciation match for an utterance by using a Viterbi search.

SIM: The Speaker Independent Monophone HMM model was the Entropic’s “Aligner”[9]. It used a proprietary on-line dictionary and trained speaker independent and context independent HMMs on 614 speakers from the manually segmented TIMIT speech database.

SDM: The Speaker Dependent Monophone aligner used the AT&T TTS dictionary with alternative pronunciations, and its context independent HMMs were trained on the single speaker 90-min. manually segmented female speech corpus.

CWT: Speaker Dependent Cross Word Triphone HMMs were trained from the same speech database as for SDM, and again the AT&T TTS dictionary was used to generate pronunciations.

Two additional variations of CWT were included in the experimental conditions:

CWT-HS: Boundary alignment halfway between manual and CWT was produced by using two sets of labels (HS and CWT) for the same set of speech files. A third set of labels was constructed by combining the two input sets. Where the label sequence was identical (a very large proportion of the time), the boundary alignment in the CWT-HS version was made the midpoint of the boundary alignments in the HS and CWT versions. In cases where there were differences in labels, the original hand corrected labels were preserved. This condition was included to simulate a 50% reduction in CWT alignment errors relative to hand segmentation.

CWT-OJ: CWT with Optimal Joins used the identical segmentations and unit selection database as in the cross

word triphone condition. The difference here was that in the synthesis process, there was an extra step after units were selected, where an algorithm attempted to choose optimal join points[3]. This step consists of examining many candidate join points for each pair of units and selecting the one with the smallest discontinuity, according to a measure based on cepstral parameters.

2.2. Perceptual Test

Six experimental concatenative TTS systems, whose 90 minute acoustic inventories were phonetically segmented and time aligned by each of six different methods were perceptually evaluated. Except for the phonetic labeling of the acoustic inventory and the optimal join option used only for CWT-OJ, the TTS system[1] used for synthesis was identical in all respects.

Test stimuli were 30 (three randomly selected sets of 10) Harvard phonetically-balanced sentences synthesized by each of the six experimental TTS systems. Recordings of the test materials were not included in the TTS acoustic inventory. Two different speech representation methods were used by each synthesizer for each test sentence in order to expand the generality of the results. Recordings of the test sentences spoken by the same female speaker who recorded the TTS inventory were included as a natural speech reference condition. Stimuli were sampled at 16 kHz, energy normalized, 40-6500 Hz band-pass filtered, and presented to listeners over headphones.

Forty-one normal-hearing native speakers of American English participated in listening tests. Listeners rated each test sentence on a five-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). There were 15,990 observations in the perceptual experiment.

3. RESULTS

First, results from two autosegmentation methods, SDM and CWT, will be compared with manual segmentation for classification agreement and for boundary alignment differences. Second, results of the formal listening test will be presented.

3.1. Objective Comparisons of Automatic and Manual Segmentation and Alignment

Figure 1 displays cumulative, absolute time differences in the alignment of segment labels (measured with respect to the reference hand labeled model) for both the speaker-dependent monophone model (SDM) and the triphone model (CWT). The triphone segmentation more closely modeled hand labelers' alignment decisions than the monophone segmentation did: The average amount of displacement for CWT from the hand labeled segments was 9 ms (median

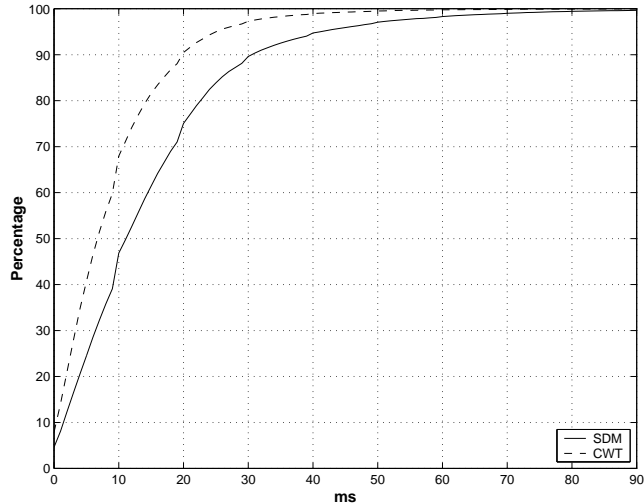


Figure 1: Cumulative errors in time alignment from reference hand labeled segment endpoints for SDM (solid line) and CWT (dashed line).

= 6 ms), while for SDM at 15 ms (median = 11 ms) it was significantly higher ($t = 70.7572$, $df = 84708$, $p < 0.0001$). For segment classification accuracy, CWT also outperformed SDM, with 95% agreement with labelers' transcriptions, vs. 88% for SDM.

Examining how the models performed on specific segment classes (vowels, other sonorants, stops, and fricatives) also shows the higher level of accuracy of CWT. As listed in Table 1¹, CWT consistently yielded results for speech segments closer to hand labeled values than did SDM. For each pair of measurements, CWT was significantly closer to the hand labeled values than SDM was.

Segment class	ONSET		OFFSET	
	SDM	CWT	SDM	CWT
vowel	-3	+2	-18	-8
sonorant	-18	-6	-8	-1
stop	-12	-7	-9	+1
fricative	-16	-11	-8	-3
\bar{X}	-11	-4	-12	-4

Table 1: Average label onset and offset misalignments in ms for SDM and CWT with respect to reference hand labeled values

3.2. Perceptual Test

As shown in Figure 2, the two HMM monophone models, SDM and SIM, fared significantly worse (each with average

¹ Overall onset and offset averages may not match since a value is obtained only if the segment was given the same transcription by both the manual and the automatic method.

ratings of 3.28) than all other voices. The hand segmented system (HS), although ranked significantly higher (with a 3.43 mean opinion score) than the monophone models, scored significantly lower than the three cross-word triphone conditions. The automatic triphone voices (CWT-HS, CWT-OJ, and CWT) had scores of 3.48, 3.50, and 3.52, respectively. The basic triphone model (CWT) performed significantly better than CWT-HS, with segment boundaries halfway between CWT and hand segmented values. The natural speech control (NC) had a mean rating of 4.53.

Listening test results were analyzed by way of a repeated measures ANOVA. For statistically significant main effects, Newman-Keuls post-hoc tests were performed with $\alpha = 0.05$.

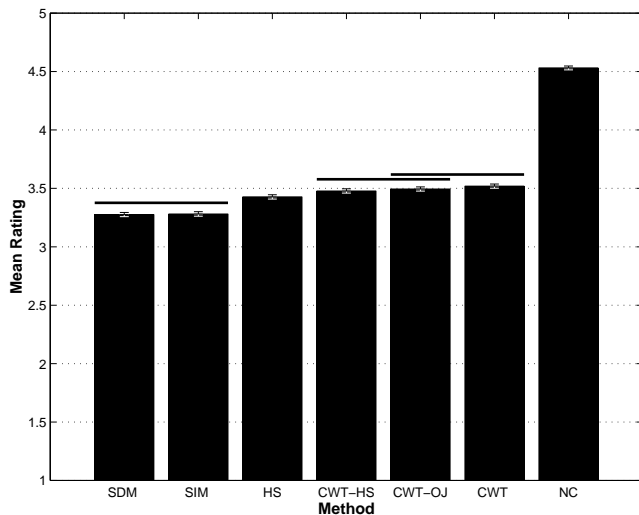


Figure 2: Mean Opinion Scores with standard error bars for the six segmentation methods and the natural speech control. Conditions whose ratings were not significantly different from each other are indicated by the same horizontal line above the bars.

4. CONCLUSIONS

CWT provides the best segmentation and concatenation method of all that were tested. We define better segmentation for our purposes as that which leads to better quality concatenative synthesis, as measured by the ratings in a formal listening test. Automatic HMM-based segmentation models using monophones (SDM and SIM) are worse than segmenting by hand, even if they are trained only on the speaker being modeled. The lack of context sensitivity of monophone models undoubtedly accounted for their poorer test results. It is well known that acoustic characteristics of segments (such as formant transitions, duration, and amplitude) can be affected drastically by the surrounding phones. The manual segmentation experts learn what these environments are, and segment accordingly. However, in this test, the triphone models outperformed the hand labelers at segmentation, probably because they are sufficiently context

sensitive and the automatic techniques are extremely consistent in placing segmental boundaries across different environments. For the cross-word triphone voices with different concatenative techniques, the optimal segmental join points for CWT-OJ did not distinguish that model from CWT in the perceptual ratings.

The practical implications of this study for TTS are very significant. Not only does the CWT method yield better segmentation and alignment than manual labeling of the same database, it is also several orders of magnitude faster. Since we have consistently achieved higher TTS quality whenever the speech inventory size has increased, the practical and perceptual impact of the CWT aligner is consequently magnified. We can now segment a larger TTS inventory better and much faster than could otherwise be produced after months of tedious manual labeling.

5. 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*, page SASC4_4, Berlin, March 1999. ASA, EAA, and DEGA.
2. X. Huang, A. Acero, J. Adcock, H-W. Hon, J. Goldsmith, J. Liu, and M. Plumpe. Whistler: A trainable text-to-speech system. In *Proc. Fourth Internat. Conf. Spoken Language Processing*, volume 4, pages 2387–2390, Philadelphia, October 1996. ICSLP.
3. A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP-96*, volume 1, pages 373–376, Atlanta, 1996.
4. H. C. Leung and V. W. Zue. A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *Proc. ICASSP-84*, volume 1, pages 1945–50, San Diego, 1984.
5. A. Ljolje and M. D. Riley. Automatic segmentation and labeling of speech. In *Proc. ICASSP-91*, pages S473–S476, Toronto, May 1991.
6. A. Ljolje and M. D. Riley. Automatic segmentation of speech for tts. In *Proc. 3rd European Conf. Speech Communication & Technology*, pages 1445–1448, Berlin, 1993. EUROSPEECH.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and T. M. Zampini. A bootstrapping training technique for obtaining demisyllable reference patterns. *J. Acoust. Soc. Amer.*, 71:1588–1595, 1982.
8. M. Wagner. Automatic labeling of continuous speech with a given phonetic transcription using dynamic programming algorithms. In *Proc. ICASSP-81*, volume 1, pages 1156–1159, Boston, 1981.
9. C. Wightman and D. Talkin. The Aligner: A system for automatic time alignment of English text and speech. Document version 1.7, Entropic Research Laboratory, Inc., 1994.