AN EVALUATION OF AUTOMATIC PHONE SEGMENTATION FOR CONCATENATIVE SPEECH SYNTHESIS

Hisashi Kawai[†] *and Tomoki Toda*[‡]

[†]ATR Spoken Language Translation Research Laboratories
[‡]Graduate School of Engineering, Nagoya Institute of Technology
{hisashi.kawai,tomoki.toda}@atr.jp

ABSTRACT

This paper studies the performance of automatic phone segmentation from two viewpoints: (1) temporal precision and (2) effect on the naturalness of synthetic speech. The absolute error of the phone onset time for the best 90% and worst 10% were 4.6 ms and 25.9 ms, respectively. These values are comparable to discrepancies among human labelers. As the result of perception tests in which naturalness was pair-compared between synthetic speeches generated from hand-segmented data and from auto-segmented data, it was found that the latter is statistically inferior.

1. INTRODUCTION

Recent concatenative speech synthesizers rely on large amount of phone-segmented speech corpora to realize high-quality synthetic speech. If it has to be done by humans, phone segmentation is labor-intensive work that requires large costs and a long time. According to the authors' experiences, the speed of phone segmentation by humans is almost 130 times real-time. For recent TTS systems requiring a speech corpus of 10 or more hours, hand-segmenting the entire corpus is economically impractical. Therefore, we should suppose a situation where only fully automatically phone-segmented corpora are available. Although automatic speech recognition techniques such as the HMM (Hidden Markov Models) enable precise automatic phone segmentation, it is believed that the precision is inferior to manual segmentation. This lack in precision can become a source of naturalness degradation.

Numerous studies have been made on automatic segmentation of phones. Almost all of the recent studies are based on forced phoneme recognition using the HMM, with a few exceptions that are based on the DTW technique [1]. A typical performance reported by Ljolje et al. [2] is that more than 80% of automatically segmented phone boundaries fall within 5 ms error when speaker-dependent HMMs are used. They also reported that context-independent models outperformed context-dependent models and that embedded training degrades the temporal precision. Their references for segmentation errors were manually segmented phone labels. However, the relationship between temporal errors and naturalness of synthesized speech is not always clear. In this respect, Makashay et al. [3] reported an encouraging result: they conducted a perception test to find that automatic segmentation can produce more natural synthetic speech. However, it is not always apparent that their result is applicable to any kind of concatenative TTS since the quality of synthetic speech is also dependent on many other factors such as phone sets, languages, speakers of the corpora, corpus sizes, and segment selection algorithms.

Therefore, in this paper the authors evaluate segmentation precision in the context of our TTS developed for Japanese. This paper is organized as follows. In section 2, the consistency of phone boundaries between human labelers is analyzed in order to clarify the upper bound of the performance of automatic phone segmentation. In section 3, optimal settings of HMM training are investigated. In section 4, perception tests are conducted in which naturalness is pair-compared between synthetic speeches generated from manually labeled data and from automatically labeled data. Section 5 summarizes the paper.

2. CONSISTENCY BETWEEN HUMAN LABELERS

When human labelers time-align phones of an identical speech data, they do not always agree with each other on phone onset times. Therefore, we conducted an experiment to quantitatively clarify the degree of inconsistency among human labelers.

The speech data is a set of 50 Japanese sentences uttered by a male professional narrator in an anechoic room. The speech was digitized at a sampling frequency of 48 kHz with 20-bit precision. The total number of phonemes is 3377, and the total duration is 237 seconds.

The speech data was first automatically segmented into phones according to phonemic transcription, and then they were simultaneously given to four human labelers. They corrected the phonetic transcription and phone onset times based on visual and audio inspection using an in-house graphical tool. They were requested not to discuss specific cases during the segmentation work.

Table 1 shows absolute errors of segment onset times for phone classes. The error here is defined as the time difference between the farthest result and the mean of results from different labelers for the same phone segment. ***Table 1*** shows that disagreement is small for plosives and flaps, while it is large for silences (i.e. closures and pauses), voiced fricatives, and semivowels. Although the overall disagreement for vowels looks rather small, error lengths differ largely depending on the preceding phones: they are small for unvoiced plosives, unvoiced fricatives, and silences, while they are large for nasals, semivowels, and vowels. Short and long pauses have the largest maximal errors. This is because it is not easy in many cases to determine the end of voicing from waveforms, which is also true for other voiced phones.

Table 1: Errors of manual segmentation

Phonetic Class	Num.	Mean Abs. Err. (ms)		Max (ms)
		Best 90%	Worst 10%	
Unvoiced plosives	345	0.7	6.8	37.5
Flaps	125	2.1	11.8	22.2
Unvoiced affricates	70	2.3	17.8	32.0
Voiced plosives	181	2.9	15.1	27.2
Devoiced vowels	2	3.2	12.5	12.5
Vowels	1370	3.2	17.6	35.0
Unvoiced fricatives	200	3.6	20.9	39.2
Nasals	328	4.2	19.7	38.2
Short closures	353	4.9	12.8	19.5
Long closures	29	5.8	16.3	18.2
Voiced fricatives	48	7.4	17.6	21.5
Vowel tails	3	9.6	19.8	19.8
Short pauses	109	9.5	32.3	58.0
Long pauses	50	10.9	37.1	55.2
Semivowels	78	11.9	28.5	32.2
All phonemes	3291	3.5	19.5	58.0

Table 2: Errors of automatic phone segmentation

Phone Class	Num.	Mean Abs. Err. (ms)		Max (ms)
		Best 90%	Worst 10%	
Unvoiced plosives	2815	2.4	10.9	117.3
Flaps	1226	3.5	17.2	114.8
Vowel tails	768	3.9	16.2	50.0
Unvoiced affricates	514	4.3	16.6	28.4
Short closures	3146	4.4	17.2	104.4
Devoiced vowels	78	4.7	18.8	29.4
Vowels	12440	4.7	27.5	305.1
Long closures	381	4.9	16.8	35.4
Nasals	3020	5.0	28.4	145.7
Voiced plosives	1715	5.1	21.7	123.7
Unvoiced fricatives	1890	5.3	28.4	122.1
Short pauses	923	5.6	26.9	68.8
Voiced fricatives	563	6.3	26.5	55.0
Semivowels	865	8.1	34.9	90.7
Long pauses	501	14.2	55.0	126.2
All phonemes	30845	4.6	25.9	305.1

3. AUTOMATIC PHONE SEGMENTATION

3.1. Speech data

The training data consists of 2259 sentences uttered by a male professional narrator who uttered the speech data used in the previous section. The content of the sentences is newspaper articles and travel conversations. The total number of phones and the total duration are 188674 and 3.69 hours, respectively.

The test data consists of 501 phonetically balanced sentences uttered by the same narrator. The total number of phonemes and the total duration are 30344 and 0.61 hours, respectively.

The training and test data were hand-segmented by the human labelers who participated in the previous experiment. Although each utterance was segmented once by a labeler, it was revised once or twice for precision and consistency.

3.2. Acoustic analysis

The speech data were first down-sampled to 16 kHz, pre-emphasized with a coefficient of 0.97, and windowed with an 11 ms-long Hamming window at every 6 ms. The windowed signals were then parameterized into a 26 component vector consisting of 12th order MFCC (Mel-Frequency Cepstrum Coefficients), Δ MFCC, log energy, and Δ log energy. ΔΔ MFCC and ΔΔ power were not used because they were shown to be ineffective for temporal precision of phone segmentation by a preliminary experiment. Acoustic parameterization and HMM training were conducted using the HTK [4], which is a standard toolkit for automatic speech recognition.

3.3. Precision of automatic phone segmentation

Monophone HMMs were trained based on the standard Baum-Welch maximum likelihood estimation using the training data described above. Embedded training was not conducted, that is, only initial training was conducted where phone boundaries were fixed to manually determined positions. HMMs consist of Gaussian distributions with diagonal covariances. The total number of states and Gaussian distributions are 225 and 954, respectively. The number of states and mixture components per HMM differ depending on phones: the numbers were experimentally optimized as described in the next section. The states are connected in a left-to-right network without skip transitions. An explicit duration modeling is not incorporated.

The HMMs were used to segment the test data. The phone sequences were fixed to manually determined ones. ***Table 2*** summarizes absolute errors of segmentation with reference to manually segmented reference data. The order of phone classes in this table is similar to that of ***Table 1***: unvoiced plosives and flaps have relatively small errors while silences and voiced fricatives have large errors. The reason why vowels have the large maximum is that their amplitudes decrease sometimes very gradually at utterance ends. The dependency of errors on the preceding phone was also similar to the consistency between human labelers.

The grand mean for the best 90% in ***Table 2*** is 31% larger than that in ***Table 1***. Although manual segmentation is surely more precise than automatic segmentation, it is also true that the both precisions are almost comparable. This implies that further improvement of precision cannot be verified easily, even if it exists, since the reference is not sufficiently reliable. The maximum errors, on the other hand, are substantially larger than those in ***Table 1***.

3.4. Optimization of training conditions

Optimizations were conducted on several training conditions to improve segmentation precision. The baseline HMMs are a set of 3-state monophones with 5 mixture components each. Silence HMMs have one state. The total number of states and Gaussian distributions are 192 and 995, respectively. The segmentation error for the baseline is shown in ***Table 3*** as condition (1).

3.4.1. Embedded training

Embedded training was conducted using the baseline HMMs as seed models. As shown in ***Table 3***(2), the segmentation precision deteriorated as iteration proceeded. This is because the

Table 3: **Optimization of training conditions**

Conditions			Abs. Err. Mean (ms) Best 90%	Worst 10%
(1)	**Baseline**		5.7	32.1
(2)	**Embedded training**	iteration=1	6.3	33.7
		iteration=2	6.8	34.9
		iteration=3	7.1	35.6
		iteration=4	7.4	36.6
		iteration=5	7.6	37.1
		iteration=10	8.1	39.6
(3)	**Biphone**		5.8	32.6
(4)	**Triphone**		5.9	33.5
(5)	**Tied-state triphone**		7.2	37.8
(6)	**Mixture number optimization**		5.4	31.1
(7)	**State number optimization**		**4.6**	**25.9**
(8)	**Training data reduction**	1/2	5.7	32.2
		1/4	5.7	32.0
		1/8	5.6	32.1

training in this case is based on the ML (Maximum Likelihood) criterion, while the reference data is determined based on the temporal precision criteria: as iteration proceeds, the effect of the ML criterion increases.

3.4.2. Context-dependent modeling

First, biphone HMMs were trained using the baseline HMMs as seed models. The phonetic contexts considered were left phones for vowels and right phones for consonants. A total of 388 biphone HMMs with 10 or more training observations were used for a segmentation experiment, resulting in use of biphone HMMs for 85.3% of phones in the test data. As shown in ***Table 3***(3), no improvement was obtained. Likewise, no improvement was obtained by triphone modeling as shown in ***Table 3***(4).

Second, following the standard training procedure, tied-state triphone HMMs were trained whose total number of Gaussian distributions was 785. As shown in ***Table 3***(5), the precision degraded substantially compared to the baseline. This is because embedded training is involved in the procedure of the training of tied-state HMMs by the design of the HTK.

3.4.3. Number of mixture components and states

First, the number of mixture components per state was optimized for each phone HMM by the hill-climbing algorithm. An optimal number of mixture components was searched in a linear manner based on the segmentation errors for a part of the training data. This procedure was run through by turns for the entire set of the phone HMMs until no improvement is obtained. The mixture number was set equal for all of the states in a phone HMM. As shown in ***Table 3***(6), the precision slightly improved.

Second, the number of states per HMM was optimized likewise in addition to the mixture number optimization. As shown in ***Table 3***(7), the precision improved substantially in this case. This is probably because the state number works as a restriction on phone duration in place of duration models, which were not incorporated in our experiments.

3.4.4. Amount of training data

Monophone HMMs were trained using reduced training data. The precisions are shown in ***Table 3***(8) for several reduction rates. Degradation of precision is not found even for 1/8 reduction.

3.5. Disambiguation of pronunciation variation

Conversion from phonemic transcription into a phone sequence is subject to ambiguities due to pronunciation variations. In the case of Japanese, important pronunciation variations include vowel devoicing, vowel elongation, and short pause insertion. It is desirable that such ambiguities be resolved automatically.

Therefore, we conducted a segmentation experiment in which ambiguities in a phone sequence are represented as several paths in a phone HMM network. The best phone sequence is selected on the ML basis. As a result, accuracies of 66.2%, 94.9%, and 96.7% were obtained for vowel devoicing, vowel elongation, and short pause insertion, respectively, where accuracy is defined as 100×("total number of samples" – "number of errors")/ "total number of samples". The accuracy for vowel devoicing is quite low: some new acoustic information other than MFCC and power, such as F_0, is necessary for improvement.

4. PREFERENCE TEST FOR HAND-SEGMENTED VS. AUTO-SEGMENTED SPEECH CORPORA

4.1. Speech stimuli and procedure

Three perception experiments were conducted to directly evaluate the effect of segmentation errors on the naturalness of synthetic speech. The speech stimulus was a pair of two synthetic speech components of the same sentence generated from a hand-segmented or from an auto-segmented speech corpus. The inventory of waveform segments is the training data in Section 3, except for the second experiment. The HMMs used for auto-segmentation is ***Table 3***(7) except for the third experiment.

The two components in a pair were sorted in random order. For each perception experiment, 53 pairs of different sentences were prepared. A set of stimuli was presented to listeners twice through headphones in a soundproof room. In the second presentation, the order of the speech samples in each pair was reversed. Nine listeners participated in the experiments. They were requested to answer which one in a pair seemed more natural than the other.

4.2. Results

In the first experiment, hand-segmentation and auto-segmentation were compared. The result is shown in **Fig. 1**(1). Apparently, the hand-segmentation is superior.

In the second experiment, the size of the auto-segmented corpus was doubled. The result (**Fig. 1**(2)) implies that doubling the size of a corpus is equivalent to conducting hand-segmentation in terms of naturalness.

In the third experiment, the effect of embedded training was assessed. The HMMs for the "Auto-embed" condition were made by conducting embedded training for 10 iterations using the HMMs in ***Table 3***(7) as seed models. The result (**Fig. 1**(3)) shows that embedded training neither harms nor enhances the naturalness, although it certainly increases segmentation errors. In other words, the difference in segmentation errors between the two conditions has no effect on the naturalness.

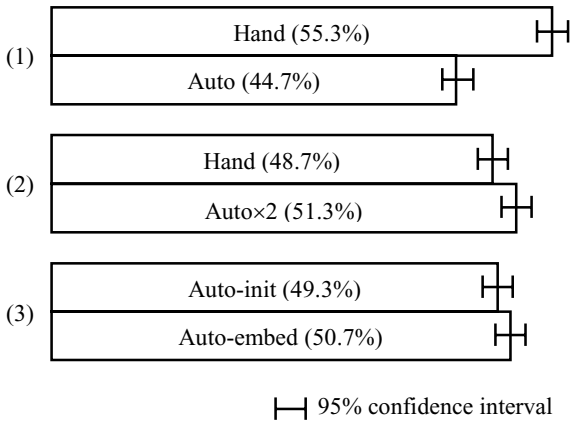

Fig. 1: Results of preference tests for speech samples that were synthesized from manually vs. automatically segmented speech corpora.

4.3. Analysis

The results of the preference test shows that manual segmentation contributes to improving the naturalness of synthetic speech. In this sub-section let us analyze the relationship between the length of segmentation errors and the degree of naturalness degradation.

Table 4 summarizes mean absolute errors and costs of segment selection for different methods of phone segmentation shown in **Fig. 1**. The errors were calculated for segments comprising the stimuli. The segments were extracted from the waveform inventory that was also used for the training of the HMMs: mean absolute errors in this table are for a part of the training data of the HMMs (closed condition). The grand means of absolute errors for the entire training data were 4.2 ms (best 90%) and 26.1 (worst 10%) for "Auto-init" condition and 5.1 ms (best 90%) and 29.3 ms (worst 10%) for "Auto-embed" condition, respectively. The number of segments for "Auto×2" condition is almost half of the other conditions since half of the segments were selected from the augmented part of the waveform inventory for which reference segmentation is not available. The cost, which guides the selection of an optimal sequence of waveform segments, is calculated from acoustical or perceptual distances between targets of synthesis and candidates of segments [5].

The mean absolute errors show in ***Table 4*** are greater or equal to those for the entire waveform inventories, which indicates that our segment selection algorithm does not have an ability to filter out segments with a large segmentation error. In other words, the cost function does not respond to the difference of the segmentation errors between "Hand" and "Auto-init": costs for these conditions are almost the same.

The perceptual experiment showed that "Auto-init" and "Auto-embed" are equivalent in terms of naturalness. This is probably because their mean absolute errors and costs are almost the same. It should be also noted that the mean absolute error of "Auto-embed" under the closed condition is smaller than that under the open condition shown in ***Table 3***(2).

Comparison between "Hand" and "Auto×2" conditions suggests that naturalness degradation caused by segmentation errors was recovered by a substantial decrease in costs that is a result of improvements in several factors affecting segment selection such as prosodic parameters and compatibility of phonetic contexts.

Table 4: **Segmentation errors for different segmentation methods.**

Segmentation condition	Number of seg.	Mean Abs. Err. (ms)		Cost ratio
		Best 90%	Worst 10%	
Hand	811	0.0	0.0	1.00
Auto-init	846	4.9	31.2	1.05
Auto×2	395	4.9	27.6	0.84
Auto-embed	816	5.3	31.8	1.08

5. SUMMARY

This paper studied the performance of automatic phone segmentation from two viewpoints: (1) temporal precision, (2) effects on the naturalness of synthetic speech. First, consistency between human labelers was measured. The absolute error of the phone onset time for the best 90% and worst 10% were 3.5 ms and 19.5 ms depending on phone classes.

Second, precision of automatic phone segmentation was evaluated. The absolute error of the phone onset time for the best 90% and worst 10% were 4.6 ms and 25.9 ms, respectively, and these values are comparable to discrepancies among human labelers. By optimizing training conditions, it was also shown that embedded training and context-dependent modeling degrade the segmentation precisions.

As the result of a perception test in which naturalness was pair-compared between synthetic speech generated from hand-segmented data and from auto-segmented data, it was found that there is a statistical difference. This difference can be compensated by doubling the corpus size.

ACKNOWLEDGEMENTS

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus".

REFERENCES

[1] J. K. Kominek, C. Bennett, and A. W. Black, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," Proc. EUROSPEECH-2003, pp. 313-316, 2003.

[2] A. Ljolje and M. D. Riley, "Automatic Segmentation of Speech for TTS," Proc. EUROSPEECH-1993, pp. 1445-1448, 1993.

[3] M. J. Makashay, C. W. Wightman, A. K. Syrdal, and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," Proc. ICLSP2000, pp. 431-434, 2000.

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Entropic Ltd. 1999.

[5] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Perceptual evaluation of cost for segment selection I concatenative speech synthesis," IEEE Workshop on speech synthesis, 2002.