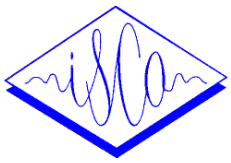

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

AUTOMATIC SPEECH SEGMENTATION FOR CONCATENATIVE INVENTORY SELECTION

Andrej Ljolje *Julia Hirschberg* *Jan P.H. van Santen*

AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974, USA

Abstract

Development of multiple synthesis systems requires multiple transcribed speech databases. Here we explore an automatic technique for speech segmentation into phonetic segments applied to an Italian single speaker database. The output segmentation is compared to manual segmentations by two human transcribers. The performance is very good on voiced stop to vowel boundaries and unvoiced fricative to vowel boundaries, while vowel to vowel and voiced fricative to vowel boundaries are estimated less accurately.

I. Introduction

Construction of a concatenative text-to-speech synthesis system usually requires the segmentation and labeling of speech recorded by a single speaker, and this segmentation and labeling must be done only once. However, every time a new language or voice is required, the process of segmentation and labeling has to be completely re-done. An automatic procedure for transcribing speech would alleviate much of the time-consuming effort that goes into building a TTS system. It could take advantage of the fact that training and testing involve a single speaker [1], which presents a much easier problem than the processing of speaker independent speech databases [2]. In addition to enhanced efficiency, an automatic procedure insures consistency in the placement of phone boundaries within the constraints of its knowledge of the speech signal, as specified by the speech model. However, due to the limited amount of speech used to train the algorithm, plus the inherent limits in parametrization of the speech signal and the speech model structure, the accuracy of the transcription is inferior to that achieved by human transcribers.

II. Automatic Transcription Algorithm

The transcription system used in this work is based on a single ergodic continuously variable duration hidden Markov model (CVDHMM) where each state is a different phone. Each state of the CVDHMM is modeled as a three-state left-to-right conventional hidden Markov model (HMM) using a separate continuous probability density function (pdf) for each of the three states. Each pdf consists of a parameter space rotation and a weighted mixture of Gaussian pdfs. It acts as a phone recognition system which is constrained by a phonotactic model. Since the phonotactic model allows only the legal phone sequence for the sentence which is being transcribed, it in fact performs phonetic segmentation. In the following experiments only one phone sequence per utterance was allowed in both training of the acoustic models and segmentation. The system can also accept a network of possible realizations if they are available and if accurate manual transcriptions are not available. The input speech is parametrized into cepstra and energy with their first and second time derivatives, using a window of 20ms and a window shift of 2.5ms. Given the speech parameters and a phone sequence, the system returns the location of phone boundaries. The model is initialized using uniform segmentation of the training utterances. It is then used to segment those same

utterances. We then repeat this procedure iteratively until there is virtually no change in the boundaries and the acoustic models. Since all of the utterances were embedded in carrier phrases, we use the whole utterance for a few iterations, but then the carrier phrase is ignored and the models are based only on the target phrases for the final few iterations.

III. Segmentation Experiments

The segmentation experiments were conducted on a database of Italian utterances, all of which were embedded in carrier phrases. We used a subset of 100 utterances for testing, with four different transcriptions. One was derived from concatenation of dictionary preferred pronunciations, the same as all of the training data. Two other transcriptions were done separately by two different human transcribers, and the fourth represented the consensus of the two human transcribers. The phonetic segments were clustered into seven categories, as shown in Table 1.

SYMBOL	Phonetic category
V	vowel
P	unvoiced stop
B	voiced stop
S	unvoiced fricative
Z	voiced fricative
L	liquid, glide
N	nasal

Table 1. This table shows the phonetic categories used to evaluate the accuracy of the automatic segmentation algorithm.

Table 2 explains the various labels which are used to describe the results in the experiments.

SYMBOL	Description
%D>10	percentage of cases where the absolute difference exceeded 10 ms
%D>20	percentage of cases where the absolute difference exceeded 20 ms, etc.
MEAN	average difference in the boundary placement (indicates biases)
MED ABS	median absolute difference
SD	standard deviation of differences
N	number of such boundaries in the test data

Table 2. The segmentation results are presented in terms of the parameters shown in this table.

We usually use the consensus transcription as the reference transcription. However, the test phone sequences can differ from the reference phone sequences, since those are obtained in several different ways. Only the boundaries that appear in all of the different transcriptions are used in the performance evaluation.

The first experiment compared the automatic segmentation based on the automatically obtained transcription from the dictionary to the consensus boundaries. The results can be seen in Table 3.

BND	TYPE	MEAN	MED ABS	SD	%D>10	%D>20	%D>30	%D>40	%D>50
P-V	-1.6	7.5	19.4	38.6	22.8	14.0	8.8	1.8	57
V-V	-4.5	12.8	35.7	59.2	38.8	28.6	14.3	12.2	49
V-N	-4.8	10.0	22.8	50.0	20.6	8.8	4.4	2.9	68
V-B	-13.9	12.5	20.3	60.0	30.0	13.3	10.0	10.0	30
V-L	-23.2	20.5	19.6	75.5	51.0	32.7	20.4	6.1	49
V-P	2.2	5.9	11.6	20.0	7.5	2.5	2.5	0.0	40
V-Z	-15.8	19.5	14.1	76.9	48.7	17.9	0.0	0.0	39
L-V	11.1	13.2	21.1	63.9	29.5	14.8	4.9	3.3	61
Z-V	15.4	21.3	24.8	78.6	52.4	40.5	16.7	7.1	42
S-V	2.7	5.2	20.8	18.8	9.4	9.4	3.1	3.1	32

Table 3. The segmentation results for the automatic segmentation algorithm when compared to the consensus boundaries

Table 4 shows the results after the removal of the bias.

BND	MEAN	MED ABS	SD	%D>10	%D>20	%D>30	%D>40	%D>50	N
P-V	0	7.1	19.4	36.8	22.8	14.0	8.8	3.5	57
V-V	0	12.3	35.7	65.3	36.7	26.5	18.4	10.2	49
V-N	0	6.8	22.8	29.4	14.7	5.9	2.9	2.9	68
V-B	0	6.4	20.3	33.3	13.3	10.0	10.0	6.7	30
V-L	0	13.5	19.6	57.1	30.6	10.2	6.1	2.0	49
V-P	0	5.8	11.6	27.5	7.5	2.5	2.5	0.0	40
V-Z	0	9.0	14.1	41.0	10.3	5.1	2.6	0.0	39
N-V	0	16.2	23.6	77.8	29.6	16.7	9.3	3.7	54
B-V	0	3.7	8.8	20.5	2.3	2.3	0.0	0.0	44
L-V	0	11.4	21.1	52.5	27.9	6.6	4.9	3.3	61
Z-V	0	18.8	24.8	81.0	47.6	16.7	7.1	4.8	42
S-V	0	7.8	20.8	25.0	9.4	9.4	3.1	3.1	32

Table 4. The segmentation results for the automatic segmentation algorithm when compared to the consensus boundaries with the bias removed

When the extra effort is made to provide the manually obtained transcription, the automatic segmentation performance is not changed, as can be seen by comparing the results in Table 5 to the results in Table 4. The differences between the automatic segmentation output and the manual segmentations can be very large. The differences between two different human transcribers remain much smaller, as can be seen in Table 6, where their segmentation is compared, with the bias removal.

IV. Results

After the mean bias was removed the automatic segmentation achieved very good results for transitions between voiced stops and vowels (97.7%% of the boundaries were within 20ms of the boundary produced by the human transcribers consensus), unvoiced fricatives and vowels (90.6% are within 20ms of the consensus boundary) and vowels and unvoiced stops (92.2% are within 20ms of the consensus boundary). For these types of boundaries the human

BND	MEAN	MED ABS	SD	%D>10	%D>20	%D>30	%D>40	%D>50	N
P-V	0	7.1	19.4	36.8	22.8	14.0	8.8	3.5	57
V-V	0	12.3	35.7	65.3	36.7	26.5	18.4	10.2	49
V-N	0	6.8	22.8	29.4	14.7	5.9	2.9	2.9	68
V-B	0	6.4	20.3	33.3	13.3	10.0	10.0	6.7	30
V-L	0	13.5	19.6	57.1	30.6	10.2	6.1	2.0	49
V-P	0	5.8	11.6	27.5	7.5	2.5	2.5	0.0	40
V-Z	0	9.0	14.1	41.0	10.3	5.1	2.6	0.0	39
N-V	0	16.2	23.6	77.8	29.6	16.7	9.3	3.7	54
B-V	0	3.7	8.8	20.5	2.3	2.3	0.0	0.0	44
L-V	0	11.4	21.1	52.5	27.9	6.6	4.9	3.3	61
Z-V	0	18.8	24.8	81.0	47.6	16.7	7.1	4.8	42
S-V	0	7.8	20.8	25.0	9.4	9.4	3.1	3.1	32

Table 5. The segmentation results for the automatic segmentation algorithm applied to the manually obtained transcriptions when compared to the consensus boundaries with the bias removed

BND	MEAN	MED ABS	SD	%D>10	%D>20	%D>30	%D>40	%D>50	N
P-V	0	1.3	2.5	0.0	0.0	0.0	0.0	0.0	57
V-V	0	14.1	22.6	65.3	26.5	14.3	10.2	4.1	49
V-N	0	3.6	13.1	7.2	2.9	1.4	1.4	1.4	69
V-B	0	1.9	4.4	6.7	0.0	0.0	0.0	0.0	30
V-L	0	6.6	14.8	32.7	8.2	4.1	2.0	2.0	49
V-P	0	8.0	9.6	25.0	0.0	0.0	0.0	0.0	40
V-Z	0	5.9	8.4	23.1	0.0	0.0	0.0	0.0	39
N-V	0	3.8	12.1	18.5	9.3	3.7	1.9	1.9	54
B-V	0	1.0	3.5	0.0	0.0	0.0	0.0	0.0	44
L-V	0	8.8	18.9	45.9	26.2	9.8	3.3	3.3	61
Z-V	0	1.4	4.5	2.4	0.0	0.0	0.0	0.0	42
S-V	0	0.7	1.4	0.0	0.0	0.0	0.0	0.0	32

Table 6. The segmentation differences between the two human transcribers with the bias removed

transcribers never disagreed by more than 20ms.

Other types of boundaries were determined less accurately. Vowel to vowel transitions, for example, were within 30ms of the consensus boundary 73.5% of the time and voiced fricatives to vowel boundaries, the same as nasal to vowel boundaries, were within 30ms of the reference boundary in 83.3% of the cases. Human transcribers also had difficulties with the vowel to vowel boundaries, and had a comparable performance for the liquid to vowel boundaries as the automatic technique.

References

[1] Ljolje, A., and Riley, M.D., "Automatic Segmentation of Speech for TTS," 3rd European Conference on Speech Communication and Technology, EUROSPEECH, Berlin, pp. 1445-1448, Sept. 1993.

[2] Ljolje, A., and Riley, M.D., "Automatic Segmentation and Labeling of Speech," Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc., Toronto, pp. 473-476, 1991.