GENERATION OF MULTIPLE SYNTHESIS INVENTORIES BY A BOOTSTRAPPING PROCEDURE

Thomas Portele, Karl-Heinz Stöber, Horst Meyer, Wolfgang Hess

Institut für Kommunikationsforschung und Phonetik
Universität Bonn
email: tpo@ikp.uni-bonn.de

ABSTRACT

In concatenative speech synthesis systems the generation of a unit inventory is a tedious task. Hovewer, some applications demand multiple voices.

A semiautomatic method to generate unit inventories is proposed. The units are segmented out of carrier phrases by means of dynamic time warping alignment with a synthesized utterance. This requires at least one existing inventory. The availability of several existing inventories will improve the likelihood of finding one with similar voice characteristics, which will improve the accuracy of results. The method is a bootstrapping procedure. To choose the best segmentation out of a set (e.g. aligned with each voice already implemented) a penalty system was developed that uses timing constraints. The results were compared with manually corrected segmentations and show the validity of this approach.

1. MOTIVATION

As speech synthesis develops from a laboratory tool to applications, the need for multiple voices is growing. Most companies like to have a personal and unique synthetic voice. Furthermore, some applications demand multiple voices.

Concatenative systems yield the best speech quality to date (Kraft & Portele, 1995). The concatenation units are extracted from utterances by a human speaker. As the number of units can easily exceed 2000, manual generation of a synthesis inventory is a tedious task. Therefore, a semi-automatic generation procedure was developed.

2. RECORDING

2.1. Text setup

The inventory for the HADIFIX speech synthesis system (Portele et al., 1994) consists of 2180 units; seven types of units are used. For the recording procedure the units were embedded in carrier phrases that were identical for all units of one type (Table 1). The sentences were set up in such a way that the units were spoken with secondary stress and the articulatory effort was minimized. The phrases were grouped in 110 groups, each group on a single sheet of paper. The carrier phrases were represented orthographically and in a phonetic transcription. A large font (20 pt) was used in the printout.

2.2. Recording

Two inventories were recorded, one spoken by a male speaker, the other one by a female speaker. The carrier phrases were read pagewise by the speakers in an anechoic chamber, recorded on a DAT recorder and simultaneously stored on hard disk. The sampling rate was 32 kHz. The recording program allows the on-line separation into one file for each carrier phrase; incorrectly pronounced or badly recorded phrases were rerecorded immediately. Two people controlled the recording, one was responsible for correct pronunciation, the other for gain control. Four one-hour sessions were required to record one inventory (Figure 1). The recording process resulted in 2180 sample files, each containing a single carrier phrase.

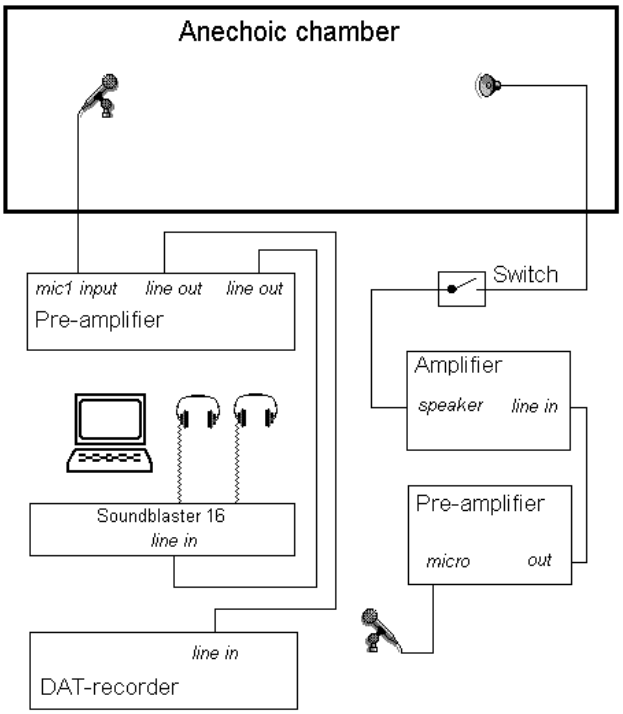

Figure 1: Setup for the recording procedure. The speaker is contacted via the microphone (right side of the figure).

2.3. Processing

The sound files were high-pass filtered and their amplitudes were adjusted. The pitch period marks were automatically determined using

Initial demisyllables:
Man wähne XXXtelei getan. /manvɛ:nəXXXtəlaɪgəta:n/
Final demisyllables:
Das SeegebXXX ist weg. /dasze:gəbXXX?ɪstvɛk/

Table 1: Carrier phrases exemplified for the initial and the final demisyllables. The 'XXX' denote the varied part.

a program developed by Ansgar Rinscheid (Rinscheid, 1993). The whole process of recording and processing required 6 hours.

3. SEGMENTATION

For the segmentation of the units speaker-independent automatic segmentation techniques are often not reliable enough, or they do not deliver all the necessary labels. We worked with one segmentation program from Florian Schiel (Schiel, 1993) that gave good results in general; however, it did not divide plosives into closure and release and used a slightly different vowel categorization etc. These problems are encountered with all speaker-independent solutions.

Speaker-dependent solutions, on the other hand, are trained with material uttered by the speaker. This can lead to good results and the categories can be chosen by the user, but it is necessary to have a certain amount of hand-labelled data.

Both methods are designed to segment every utterance. Using identical or very similar carrier phrases makes the task much easier because a large part of the utterance is the same in in all utterances. Two possible ways to use this fact were explored.

3.1. Unit boundary detection

The first method was the exact determination of the unit boundaries by measuring the difference between the target utterance and twelve reference utterances using a dynamic time warping (DTW) algorithm. The peak of the sum of the difference functions should mark the position of the varied part, i.e. the unit.

Several difference measures were tested (mel-cepstrum, lpc cepstrum, parcor etc.). Their results were similar. Although the maximum of the difference function was always located at the position of the unit, the differentiated function did not show the two prominent peaks that were expected. This is probably due to coarticulatory effects. The results of this procedure could serve as a rejection criterion for subsequent automatic segmentation methods. In this case, however, they were not used because of the similarity to the second method: errors when applying the first method will also appear when applying the second one.

3.2. DTW with an already segmented utterance

The second idea was to apply the DTW algorithm and to align the target utterance with an already segmented reference utterance. The reference utterance was synthesized using one of the voices already present in the system. This bootstrapping procedure implies that in the beginning at least one voice is made by hand labelling. Further voices can then be built using synthetic stimuli from all previous voices.

The alignment was done using six different parameterizations, i.e. LPC, cepstrum, and mel-cepstrum, each of them in an energy-normalized and a non-normalized version. Two inventories had already been constructed; twelve different segmentations were therefore possible. The synthetic utterances had an average F_0 similar to those of the target utterances, and the durational structure of the carrier phrases was adopted.

Timing constraints were used to choose the best versions, taking advantage of the similar structure of the carrier phrases. About 5 carrier phrases for each unit type were segmented by hand to obtain average values for segmental durations in the phrases. These durations served as constraints for a penalty system. A comparison of the results with the hand-labelled versions showed that in every case the best solution was chosen, and that the average error was in the range of 20 ms. Only the four best segmentation methods were kept for the complete procedure. This stage took two hours for each inventory.

The segmentation of all carrier utterances with four different segmentations took 48 hours on a workstation.

An advantage of this procedure is that the labels are exactly in the system format. For instance, concatenation points are also aligned, and they are usually adequate as first guesses.

4. MANUAL CORRECTION

The units were then corrected manually using a mouse-based labelling program written especially for this purpose. About 100 units were checked per hour, resulting in a total amount of 22 hours for one inventory. The complete amount of time for the generation of one synthetic voice is calculated in Table 2.

Recording: 4 hours (with additional 3 hours for breaks)
Processing: 2 hours
Hand segmentation of 35 utterances: 2 hours
Automatic segmentation: 48 hours
Inventory generation: 1 hour
Manual correction: 22 hours
Total time (human): 28 hours
Total processing time (SPARC10): 51 hours
Total time: 79 hours

Table 2: Time to generate one inventory using a SUN SPARC10.

5. EVALUATION

To evaluate the proposed method, a comparison between automatic segmentation results and the manual corrections was performed. Answers were obtained for the following questions:

1. How good is the overall segmentation quality?
2. How good are the different segmentation methods?
3. How good is the penalty system?

The dependence of the results on speaker, unit type (i.e. context) and sound class was also investigated.

5.1. Overall quality

Figure 2 displays a histogram of the absolute difference between automatic and manual segmentations. It can be easily seen that for most labels the difference is less than 10 ms. This means that more than half of the labels remained unchanged. A numerical analysis revealed that the difference is less than 10 ms for 50.9% (no change necessary), less than 40 ms for 74.5% (close to the original position, easy to change), and less than 100 ms for 89.6% (within the same syllable). Only for 10.4% of all labels were changes of more than 100 ms necessary.

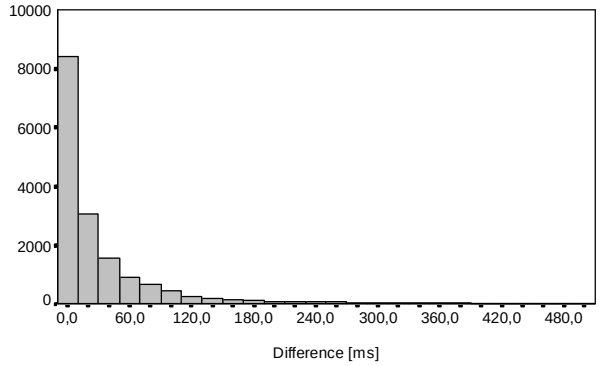

Figure 2: Histogram of the differences in ms between the results of the automatic segmentation process and the manual corrections.

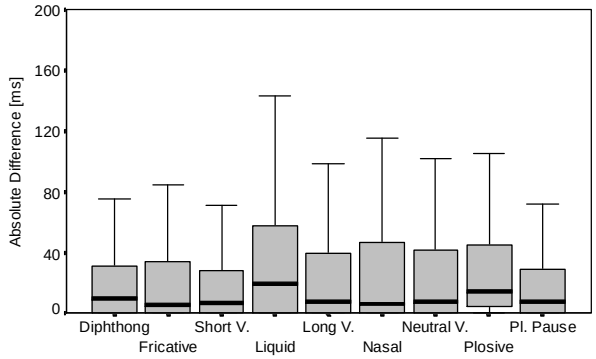

Figure 3: Boxplot of the absolute differences between automatic and manually corrected labels, displayed for each sound class.

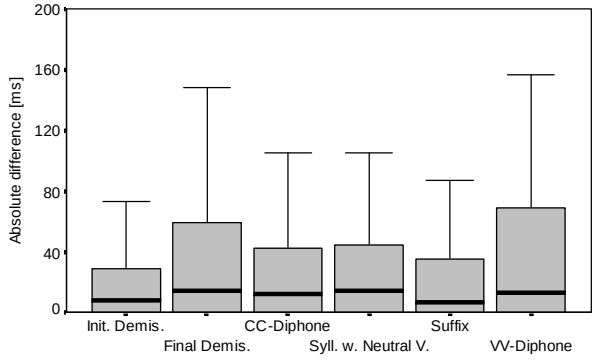

Figure 4: Boxplot of the absolute difference between automatic and manually corrected labels, displayed for each unit type.

The differences between the speakers are negligible. Figure 3 shows the differences between the sound classes. It is not surprising that liquids are badly segmented because of their similarity to vowels (especially in postvocalic position). Plosives were not as exactly placed as necessary due to the frame shift size (10 ms) of the DTW that was too large for this task.

Figure 4 displays the dependency between segmentation accuracy and unit type. The largest errors are found for final demisyllables due to the difficulties in finding the boundary between a vowel and a postvocalic sonorant, and by vowel-vowel diphones, probably for the same reasons (no abrupt changes but slow transitions).

The results show that the problems of the automatic segmentation process are boundaries between vowels and postvocalic sonorants and the exact placement of plosives; this is more or less as expected. Generally, the results are very consistent.

5.2. Distance measurements

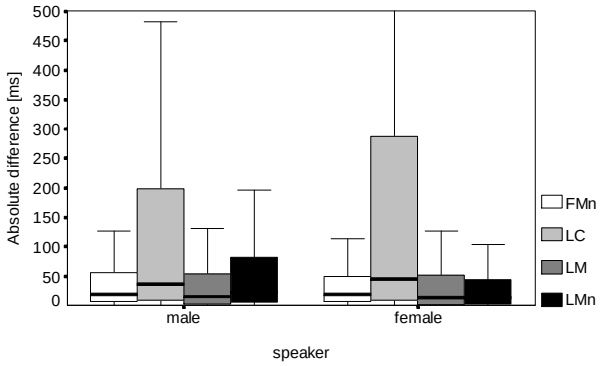

Figure 5: Boxplot of the absolute difference between automatic and manually corrected labels, displayed for speaker and segmentation version (see text).

Four different segmentation methods were used in the final segmentation process: the old female voice with normalized mel-cepstrum coefficients (LMn), the old female voice with non normalized mel-cepstrum coefficients (LM), the old female voice with cepstral coefficients (LC), and the old male voice with normalized mel-cepstrum coefficients (FMn). In every case the euclidean distance was computed. It would speed up the segmentation process if this number could be reduced to two, because every method takes about 12 hours. In order to determine the quality of the four versions they were compared with the manually corrected versions. Figure 5 displays the results. It is obvious that the mel-cepstrum coefficients are the method of choice, and that the method is speaker-independent; the results are only marginally better when the sexes of the natural and the synthetic voices match.

5.3. Penalty system

The penalty system works in two steps. In the first step, all significant differences between expected and segmented durations in the unit get a penalty value of 1. Segmentations with a penalty sum greater than two plus the minimal penalty sum of all segmentations of a certain utterance are excluded.

In the second step the context is analyzed using the result from the test segmentations by hand. Here, the differences between expected and segmented durations are summed up, and the segmentation with the smallest number is chosen for the particular utterance.

The performance of the system depends on careful adjustment of the timing constraint values, and was assessed by computing the differences between the best and the chosen segmentation results and

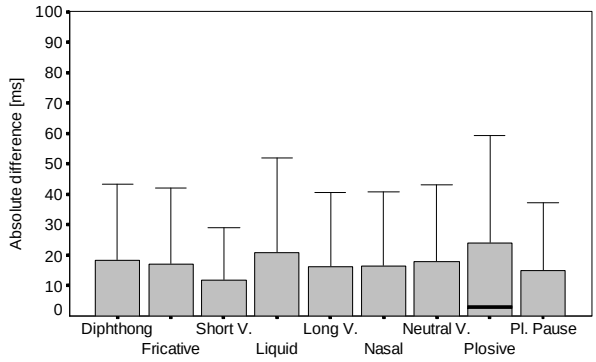

Figure 6: Difference between chosen and best segmentations, displayed for each sound class.

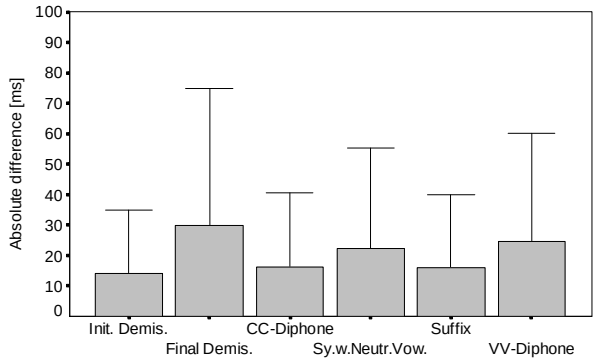

Figure 7: Difference between chosen and best segmentations, displayed for each unit type.

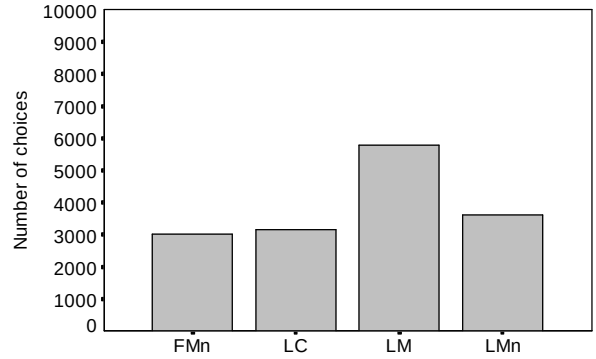

Figure 8: Number of choices of a segmentation method by the penalty system.

by counting the number of correct decisions. The correct segmentation was chosen in 67.6% of cases; in 75.9% the chosen and the best labels differ by less than 20 ms, in 87.1% by less than 50 ms.

The penalty system performed equally well for both speakers. Figure 6 displays the difference between chosen and best version for each sound class; the results are very consistent. Figure 7 shows that final demisyllables and vowel-vowel-diphones were most difficult; this might be due to a less accurate modelling of the durational structure. The penalty system proved to be an effective way of choosing the best segmentations. Figure 8 indicates that such a system is necessary because all segmentation versions have their share in the global result.

6. CONCLUSIONS

Multiple voices are necessary for many applications. The method described here allows the generation of new voices by using existing ones in a bootstrapping procedure. Its successful application to the construction of two new inventories demonstrates its potential. The more voices a system features, the better the results of the bootstrapping procedure are likely to be. This approach is rather machine-time-consuming, but machine time is cheaper than human time. Improvements in computer hardware will directly increase the speed of the inventory generation process proposed here. Many possible extensions exist; especially the refinement of the penalty system using additional information (such as voice-voiceless distinction).

The greatest improvement will probably come from a different set of carrier phrases. The current phrases were chosen solely for neutrality of context and ease of articulation; they are difficult to segment (for instance, the context before an initial demisyllable is completely voiced: /manvɛ:nə/). Carefully chosen phrases that are easy to articulate but switch between voiced an unvoiced segments as anchor points will further increase the rate of correctly labelled units.

Acknowledgements: We thank Monika [SOUND A331S02.WAV] and Jürgen [SOUND A331S01.WAV] for their voices, Monika Rauth, Barbara Heuft and Gerit Sonntag for helping with the manual corrections, and Simon King for many helpful comments.

7. REFERENCES

Kraft, V.; Portele, T. (1995) "Quality assessment of five German speech synthesis systems", *Acta Acustica* **3**, 351-366, 1995

Portele, T.; Höfer, F.; Hess, W. (1994a) “Structure and Representation of an Inventory for German Speech Synthesis”, *Proc. ICSLP’94*, Yokohama, 1759-1762, 1994

Rinscheid, A. (1993) "Automatische Bestimmung von Perioden-marken mit dem emark-Algorithmus." *Fortschritte der Akustik - DAGA '93*, Frankfurt, 1048-1051, 1993

Wesenick, M.B.; Schiel, F. (1994) “Applying Speech Verification to a Large Data Base of German to Obtain a Statistical Survey about Rules of Pronunciation”, *Proc. ICSLP’94*, Yokohama, 279-282, 1994