High-Quality Speech Synthesis Using Demisyllables and a Variable-Frame-Rate Vocoder

Wolfgang Hess

Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee 47, D-5300 Bonn 1, W. Germany

Abstract. *An experimental high-quality speech synthesis system is described. Demisyllables are used as phonetic units for concatenation; in a first step it is shown that 1665 demisyllables requiring about 0.5 MByte of memory at a data rate of 7.2 kbit/s are sufficient to synthesize a very large German vocabulary. To generate the output speech signal, a special variable-frame-rate vocoder synthesizer is implemented.*

1. Introduction

Text-to-speech synthesis systems principally consist of three major components: 1) an orthographic-to-phonetic transcription (including prosody control); 2) the concatenation block; and 3) a vocoder synthesizer. Usually the output of the orthographic-to-phonetic transcription block is a string of phonetic symbols plus a number of special characters for prosody control. The concatenation component converts this string into a data stream of vocoder parameters which are then transformed into synthetic speech by the vocoder synthesizer.

In the last years work on speech synthesis has concentrated upon higher-level tasks, i.e., orthographic-to-phonetic transcription and prosody control. Nevertheless, there are still a number of unsolved problems in connection with concatenation and even with the vocoder synthesizer; due to these problems, the quality of synthetic speech may still be unsatisfactory even for synthetic utterances with a well-modeled prosody. This paper deals with possibilities of improving the quality of synthetic speech by optimizing the concatenation block (Dettweiler, 1981, 1984) and by designing a vocoder that is particularly well adapted to a speech synthesis system by rule (Heiler, 1982, 1985).

2. Concatenation System for Demisyllable Elements

Concatenation is a central problem in any system for speech synthesis by rule. It provides the link between the phonetic level and the parametric level of the system. In practice concatenation is controlled by a set of rules that act upon a data base of speech data. This data base may contain experimental data, such as tables of formant frequencies; however, it may also consist of (parameterized) natural speech. The design of the concatenation component is determined by a tradeoff between the number and complexity of the concatenation rules on the one hand and the size of the memory required for the data base on the other hand. The crucial question in this respect is that of the phonetic units to be applied.

2.1 The Demisyllable Approach

Besides phonemes and diphonemes, syllabic units supply a viable data base for high-quality synthesis by rule. The influence of coarticulation strongly diminishes when a syllable boundary is crossed (Fujimura, 1981; Öhman, 1966). When syllabic units are used, the number of elements is minimized when the syllables are split up into *demisyllables* (DSs). Demisyllables as units of speech processing were first proposed by Fujimura both for speech recognition (1975) and for synthesis purposes (1976). For German DSs were taken up by Ruske and Schotola (1978) for a speech recognition system; for synthesis by rule they were first used by Dettweiler (1980, 1981).

Usually a syllable is defined to consist of the syllabic *nucleus* (in German this is always a vowel or a diphthong) which is preceded and followed by a number of consonants, the so-called *consonant clusters* (CCs). The consonants preceding the syllabic nucleus form the *initial consonant cluster*, and the consonants following the nucleus represent the *final consonant cluster*. A syllable is subdivided into demisyllables by cutting it within the syllabic nucleus. The initial CC and the beginning of the syllabic nucleus form the *initial demisyllable*, whereas the remainder of the nucleus and the final CC make up the *final demisyllable*.

2.2 The DS Inventory. Synthesizing Monosyllabic Words

A representative DS list for German was compiled by Ruske and Schotola (1978; cf. also Schotola, 1984). The initial CCs contain from 0 to 3 consonants, whereas up to 5 consonants may exist in a final CC. The number of CCs is rather limited due to linguistic constraints: we have to deal with only 51 initial and 159 final CCs (Dettweiler, 1984). Concerning the syllabic nuclei, 23 vowels and 3 diphthongs must be taken into account.

Contrary to speech recognition, where the syllabic nuclei and the CCs can be treated separately (Ruske and Schotola, 1978), the transitions between the syllabic nuclei and the CCs are essential for the quality of the synthesized speech; they cannot be generated by rule and must be available as stored data. For the complete DS inventory the number of elements thus becomes

$$N_C = 26 \cdot 51 \text{ initial DSs} + 26 \cdot 159 \text{ final DSs} = 5460 .$$

Since coarticulation has a strong tendency toward anticipating future articulatory gestures (Delattre, 1968; Fujimura, 1981), it is adequate to establish the DS boundary within the first part of the vowel. Fujimura's proposal (1976) to place the boundary 50 ms after the beginning of a vowel is also applied in our system (Dettweiler, 1981; cf. Fig.1).

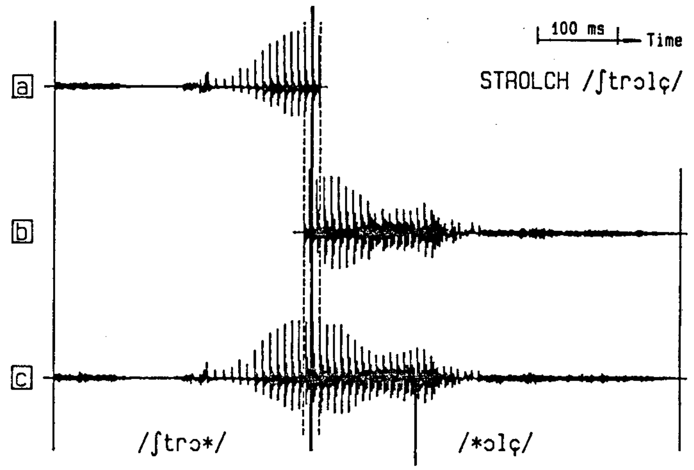

Fig.1a-c. Concatenation within the syllabic nucleus (rule CR1). (a) Initial DS, (b) final DS; (c) complete word after concatenation. The thick vertical line indicates the interconnection point; the smoothing interval is indicated by the dashed lines. The asterisk in the phonetic transcription refers to the position of the syllabic nucleus

2.3 Inventory Reduction

To reduce the number of DSs, two ways seem feasible: 1) vowel substitution, and 2) further splitting of CCs. Both these possibilities have been used in our system; the most important rule being the principle of rudiment and suffix (Dettweiler, 1981; Fig.2).

Certain consonants, when occurring in final position of a DS, may be split off from the DS and form separate units, the so-called affixes (Fujimura et al., 1977). As the experiments suggest, fricatives and stops in final position, like vowels in the syllabic nuclei, represent a natural coarticulation barrier; i.e., sounds following this barrier do not (substantially) affect previous sounds. A splitting scheme which is particularly efficient for German is the principle of *rudiment* and *suffix* (Dettweiler, 1981, cf. Fig.2). A suffix is defined to consist of any (existing) combination of the four consonants /f/, /s/, /ʃ/, and /t/, whereas the remainders of the final DSs form the rudiments. The linguistic constraints of German state that once a suffix consonant, i.e., one of the 4 consonants named above, has occurred in a final CC, the following consonant(s) of that final CC, if existing at all, must be suffix consonants as well.

In practice the rudiment is formed by uttering a DS that contains the remainder of the consonantal cluster (without any suffix consonant) plus a final /t/ and then removing the /t/ together with the pertinent silence before the burst (Fig.2b). Since the rudiment contains all the coarticulatory influences by the following /t/, it is easy to see that the rudiment and the final DS containing an identical consonant cluster without the /t/ are different (cf. Fig.2a,b). Any rudiment and any suffix may be simply concatenated without any smoothing at the interconnection point.

Using all these possibilities of inventory reduction, the total number of elements now decreases to N_R = 1665. Note that these inventory reductions do not degrade the quality of the synthetic speech.

With an average duration of 0.3 s per element, the memory required for this inventory is less than 0.5 MByte if a vocoder at 7.2 kbits/s is used.

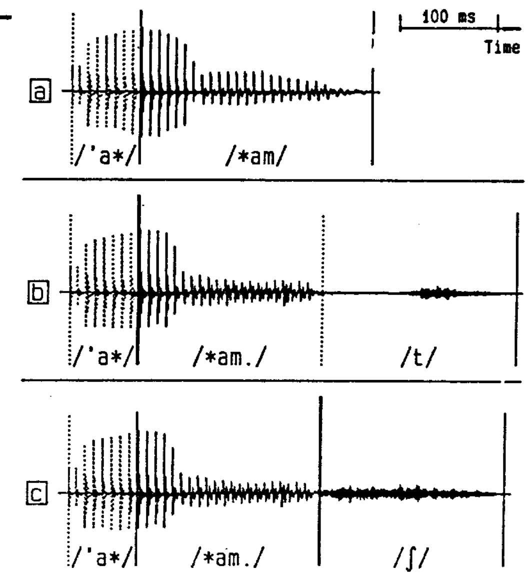

Fig.2a-c. The principle of rudiment and suffix. (a) Ordinary consonant cluster: example /*am/. (b) Rudiment and suffix: the DS /*amt/ is split up into the rudiment /*am./ and the suffix /t/ (the dotted line represents the boundary). (c) Concatenation using rudiment and suffix: /*am./ ¦¦ /ʃ/ —> /*amʃ/. Signals drawn with dotted lines represent DSs that are needed to complete the word, but do not pertain to the DSs involved in rule CR2

2.4 Synthesizing Polysyllabic Words

Polysyllabic words contain *intervocalic consonant clusters* between subsequent syllabic nuclei. This requires additional rules for the concatenation of CCs (Dettweiler, 1984). The procedure is carried out in two steps. First an intervocalic CC is split up into a final CC followed by an initial CC, and the CCs are joined to the respective syllabic nuclei to form DSs. In the second step the DSs are concatenated.

The ICCs are split according to three rules. Firstly, an intervocalic CC must always be split up into a *valid* final CC and a *valid* initial CC. A CC is regarded as *valid* if it is contained in the DS inventory. If this rule does not yield a solution, the DS inventory must be enlarged. On the other hand, if this rule provides several solutions, a second rule states that the one solution is selected where as many consonants as possible are grouped within the initial CC. This "pragmatic" boundary takes into account the anticipatory effect of coarticulation; even when a DS boundary as established by this rule, differed from a given morph boundary. These two rules thus represent an adequate means to split up intervocalic CCs without requiring morphologic knowledge at this level.

When the intervocalic CC only contains one consonant, a third rule switches the system into a diphone mode by assigning this consonant to both the initial and the final DSs.

The way in which intervocalic CCs are concatenated strongly depends on the consonants involved. Due to lack of space, the individual rules cannot be discussed here. A flow diagram is depicted in Fig.3; the labeling of the concatenation rules (CR 3-12) corresponds to that in (Dettweiler and Hess, 1985); for an in-depth discussion, the reader is referred to that publication.

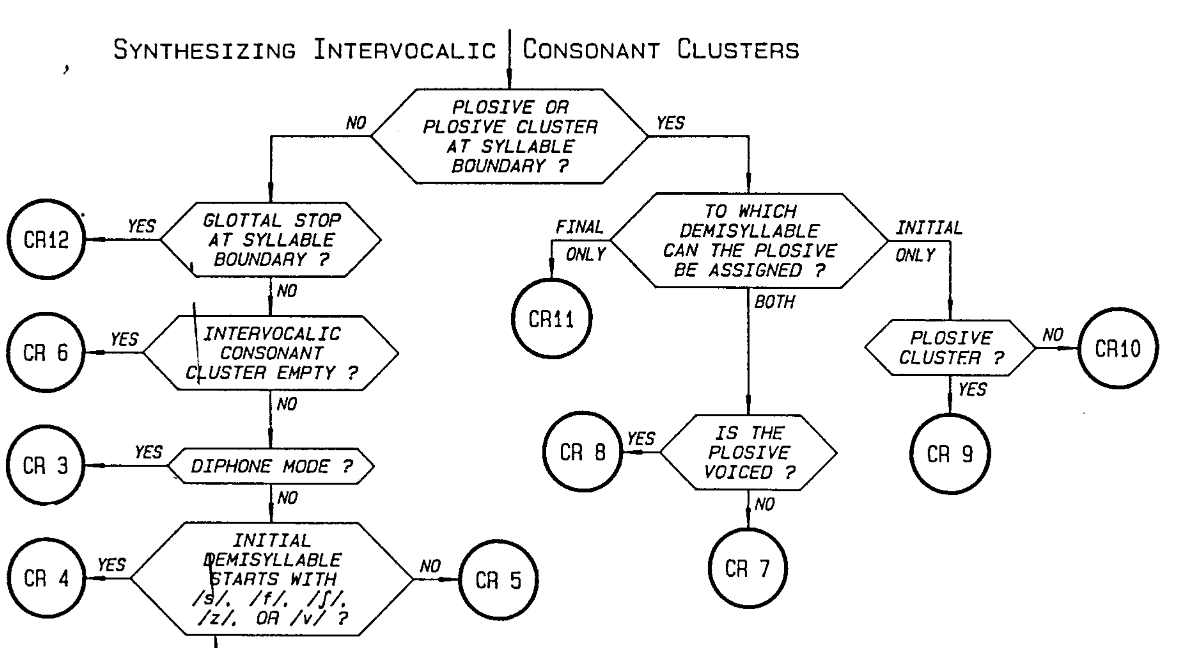

Fig.3. Block diagram of the concatenation step for intervocalic consonant clusters. The labeling of the concatenation rules (CR3 to CR12) corresponds to (Dettweiler and Hess, 1985)

2.5 Realization of an Experimental System

An experimental system was implemented using a 12th-order standard LPC vocoder with a constant frame interval of 10 ms and a signal sampling frequency of 10 kHz. For data acquisition the DSs were embedded in two-syllable meaningless words of the form /<initial DS>tər/ and /gət<final DS>/; the DSs were manually delimited using a display program and an interactive segmentation procedure.

Compared to systems that use phonemes as units for concatenation, the number of concatenation rules in this system is extremely low, and the quality is much better. In an intelligibility test, Dettweiler (1984) showed that the median word intelligibility dropped from 96.6% for vocoded speech to 92.1% for the DS synthesis system (using the same vocoder). The quality of the vocoder speech and the demisyllable synthesis system were judged to be almost the same. However, there were still a number of systematic confusions of fricatives, such as /s/ and /f/; some other errors were due to incorrect segmentation of the demisyllables. Systematic errors due to the concatenation rules were not encountered.

3. High-Quality Variable-Frame-Rate Vocoder

Since Dettweiler's experiments (1984) showed that even a rather high-quality vocoder may be a source of systematic intelligibility errors, it is useful to redesign the vocoder and to adapt it to the special requirements of the synthesis system by rule.

Usually vocoders are designed for the purpose of parametric speech transmission. Important criteria are robustness, good performance even in adverse environment, and real-time operation. If a vocoder, however, is to be optimized with respect to speech quality at a given transmission rate, the principle of variable frame rate (VFR) vocoding best fulfils this requirement (Huggins et al., 1977). In a VFR vocoder the frame rate is adapted to the speed of articulatory movement; i.e., frames are selected and transmitted in rather large time intervals during stationary segments such as vowels, whereas the frame rate during rapid transitions is rather high. This principle is not well suited for transmission purposes since it is rather complex in the analysis part; in addition, it may introduce a substantial processing delay which is intolerable in a dialog. In a speech synthesis system by rule, however, the requirements are substantially different. Analysis of the input data is done off line (even manual interaction may be permitted) and with high-quality data; thus the analysis algorithms may be sensitive and complex. In addition, the question of the processing delay is irrelevant here. Such a system is thus extremely well suited for this kind of application.

Heiler (1985) developed and optimized the so-called "evolution strategy" (Fig.5) for selecting the frames and approximating the parameters to be stored. In this algorithm an utterance (e.g., a DS) is regarded as one unit. Predetermined are 1) the number of frames to be selected, 2) the interpolation procedure for the frames which are not selected (a combination of linear interpolation and simple repeating of the most recent frame proved to give the best results), and 3) the approximation strategy for the parameters at the selected frames. The algorithms starts with the two frames at the beginning and end of the DS that must be selected. The third frame is positioned in such a way that the accumulated approximation error becomes a minimum over the whole utterance. Then another frame is looked for (with the frames kept constant that were already selected) according to the same criterion; this procedure continues until the desired number of frames have been obtained. Due to the successive approximation, however, the selection of frames is not yet optimal. For further optimization one more frame is now added to the selection. To keep the number of frames constant, the algorithm then removes the one frame whose removal contributes least to the global approximation error. If the removed frame is different from the one which was added in that step, this results in a frame shift; if it is the same, the optimization is terminated.

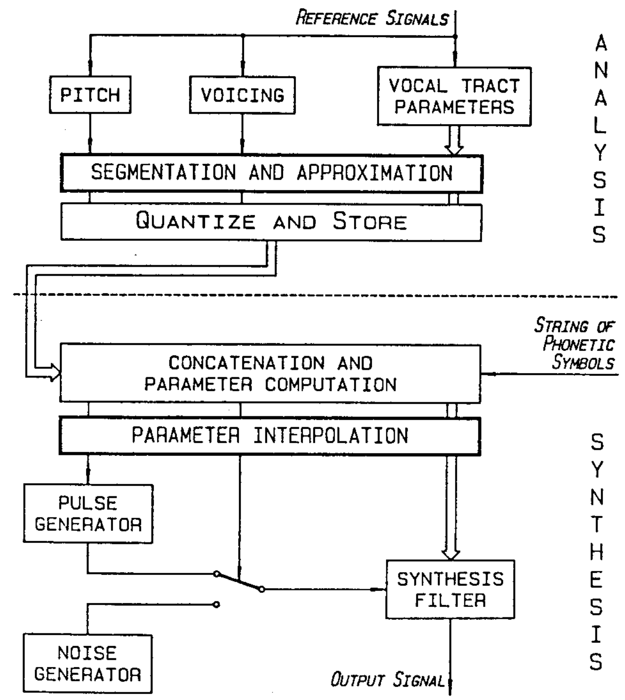

Fig.4. Vocoder configuration for speech synthesis by rule. The analysis (components above the dashed line) is done offline

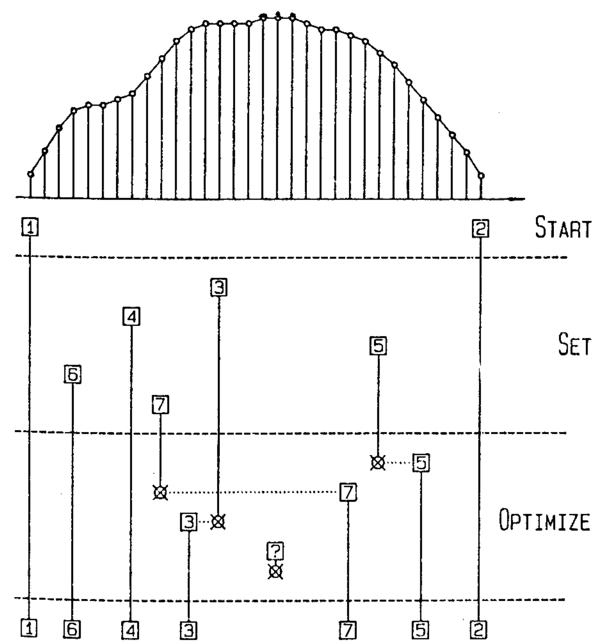

Fig.5. Example for the evolution strategy for a VFR vocoder system. After Heiler (1985)

In subjective listening experiments Heiler (1985) showed that, compared to a vocoder with constant frame rate and no parameter optimization, this VFR principle permits reducing the bit rate by a factor of 3 without a perceptible loss of quality.

4. Discussion and Conclusions

The work descibed in this paper concentrates on quality improvement of text-to speech synthesis by optimizing the front-end steps, i.e., the concatenation block and the vocoder synthesizer.

The use of demisyllables as phonetic units offers the great advantage that about 20 rules and 1650 DSs requiring a data memory of less than 0.5 MByte are sufficient to synthesize (nearly) unrestricted German text. A special variable-frame-rate vocoder synthesizer provides an optimal quality at a given data rate and helps minimizing the required amount of memory.

At the moment the synthesis system by rule and the VFR vocoder still exist as separate units. Efforts are under way to combine the two systems, thus improving the quality of the vocoder in connection with the stored data. A signal bandwidth of 7 kHz requiring a sampling frequency of 16 kHz will eliminate the systematic confusions between the fricatives /f/ and /s/ present in the actual 5-kHz system, and a VFR scheme permitting a minimum frame rate of less than 5 ms without increasing the overall amount of memory will particularly improve the quality of synthetic stop consonants.

Acknowledgement. The major part of this paper was extracted from the Dr.-Ing. dissertations by Dr. H. Dettweiler and Dr. J. Heiler.

References

Delattre P. (1968): "From acoustic cues to distinctive features." Phonetica 18, 198-230 703-706 (VDE-Verlag, Berlin)
Dettweiler H. (1981): "An approach to demisyllable synthesis of German words." Proc. IEEE ICASSP-81, 110-113
Dettweiler H. (1984): Automatic synthesis of German words by means of syllable-oriented segments. Dr.-Ing. dissertation, Technical University of Munich (in German)
Dettweiler H., Hess W. (1985): "Concatenation rules for demisyllable speech synthesis." Acustica 57, 268-283.
Fujimura O. (1975): "Syllable as a unit of speech recognition." IEEE Trans. ASSP-23, 82-87
Fujimura O. (1976): "Syllable as the unit of speech synthesis." Unpublished Bell memorandum (Bell Labs, Murray Hill, NJ)
Fujimura O. (1981): "Temporal organization of articulatory movements as a multidimensional phrasal structure." Phonetica 38, 66-83
Fujimura O., Macchi M.J., Lovins J.B. (1977): "Demisyllables and affixes for speech synthesis." Proc. 9th Int. Congr. on Acoustics, Madrid 1977, paper I107
Heiler J. (1982): "Optimized frame selection for variable frame rate synthesis." Proc. IEEE ICASSP-82, Paris 1982, 586-589
Heiler J. (1985): Minimization of the memory requirements of speech synthesis systems by optimizing the parameter approximation. Dr.-Ing. dissertation, Technical University of Munich (in German)
Huggins A.W.F., Viswanathan R., Makhoul J. (1977): "Speech-quality testing of some variable frame rate (VFR) linear predictive vocoders." J. Acoust. Soc. Am. 62, 430-434
Ruske G., Schotola Th. (1978): "An approach to speech recognition using syllabic decision units." Proc. IEEE ICASSP-78, Tulsa, OK, 722-725
Schotola Th. (1984): "On the use of demisyllables in automatic speech recognition." Speech Commun. 3, 63-87