Microsegment Synthesis - Economic principles in a low-cost solution

Ralf Benzmüller & William J. Barry

Institut für Phonetik, FR 8.7, University of the Saarland, 66041 Saarbrücken
e-mail: benzm@coli.uni-sb.de or wbarry@coli.uni-sb.de

ABSTRACT

A low-cost concatenation based speech synthesis system for German is described which combines the advantage of minimal memory requirements with good intelligibility and high segmental and prosodic acceptability. This is achieved by the multiple use of "microsegments", stretches of speech signal varying in length from demi-phone to phone size. All prosodic structuring is carried out in the time domain.

1. INTRODUCTION

Despite the increasing availability of memory at decreasing cost, there is an enduring and increasing need to minimize the memory requirements for speech output systems because, firstly, it operates mainly as an accompanying facility to other primary applications and, secondly, it is becoming an ever more frequently needed facility on personal computer systems. However, economizing on memory should not be at the expense of synthesis quality; synthesis systems have now reached a level where naturalness is becoming an increasingly important criterion.

Against this background, we were challenged[1] to produce a real-time software-based concatenation speech synthesis system of acceptable quality to run under Windows on a 386 PC with 4 MB RAM and a Soundblaster card. This reduces the working memory available for speech-signal data to approx. 1 MB (compare this with the memory requirements in Meyer et al., 1993 and Taylor et al., 1991) and precludes the use of memory-intensive processing methods. In line with developments, the original specifications have now been changed to comply with Microsoft API specifications. It now runs under Windows 95 on a 486 configuration with the currently more common 8MB of working memory. These constraints might appear less severe. However, many commercial programs have also increased demands on memory , maintaining the pressure on memory-resident companion-programs to restrict their needs.

2. MICROSEGMENT SYNTHESIS

UNDERLYING PRINCIPLES

Adopting a concatenation approach for reasons of baseline quality, in particular with respect to the recognizable "human quality" of the voice, and rejecting LPC resynthesis techniques for the combined reasons of computational cost and speech timbre quality loss, we sought a solution exclusively in the time domain (compare Posmyk, 1989). In order to minimize memory requirements we applied speech knowledge to generalise in the selection of acoustic speech-signal segments. The basic rationale is the multiple use of speech-signal segments, both for consonants, which are to a large extent used in a context-free way, and for vowels, which are differentiated within contextual classes defined by place-of-articulation categories but generalised across manner categories, and over different prosodic conditions, i.e., different stress, accent, and boundary conditions.

SEGMENTATION PROCEDURE

Underlying this approach is the extremely careful selection and excision of the microsegments from the original recordings.

The speech-signal segments are excised manually from single-word or short- phrase utterances. This provides basic units with long durations because of their accented and utterance-final position. The recordings were produced on a monotone under low reverberation conditions, and care was taken to maintain the same mouth-microphone position. This means that voiced segments are well matched in phase conditions, and require only minimal F0 adjustment for optimal concatenation. Production on a monotone is the prerequisite for the time domain F0 manipulation and thus for intonational structuring, as described below (section 4).

SEGMENTAL STRUCTURE - THE INVENTORY OF MICROSEGMENTS

With regard to the segmental structure, three main categories of segments are defined which, since they follow different segmentation criteria from diphones and are, in most cases, smaller than speech sounds (phones), we call "microsegments". These categories are:

a) Consonantal segments (phone-sized with some allophonic differentiation), which due to their relatively constant spectral structure during their time course can serve as concatenative units. Some contextual differentiation is required in the case of the sibilant fricatives and plosive release segments. Both these microsegment subcategories are defined according to the following vowel context. In the case of the former, rounded and unrounded vowel contexts are catered for; in the case of plosive release, front

1 The work has been funded by G-Data Software GmbH Siemensstr.16, D 44793 Bochum, Germany. In particular, A. Lüning has been responsible for the programming and software implementation of the principles described here.

non-open rounded and unrounded, open, and non-open back vowel contexts are differentiated.

b) Demi-vowels and -semivowels (sub-phone-sized and diphone-like), using one generalised demi-vowel microsegment for each of the following main places of articulation/articulators:

- (i) labial for bilabial and labiodental ([+labial]);
- (ii) alveolar for dental, alveolar and post-alveolar ([+coronal]);
- (iii) velar for palatal, velar and uvular ([+dorsal]). Because of their unmarked spectral structure (not breathy as in fricative or voiceless stop contexts, nor nasal as in nasal contexts) demi-vowel microsegments are excised from voiced stop contexts. They are defined according to pre- or post-context, depending on whether they represent the first or second half of the speech sound.

c) Quasi-stationary vowel segments (sub-phone-sized and phone-like) taken from the centre of long vowel-realisations. They are not marked for context, and are used

- word-initially,
- with the semi-vowel /j/ and the glottal consonants /h/ and /ʔ/,
- as target elements in diphthongs, and
- as vocalic core elements in vowel-vowel sequences

In the latter two cases they are linked by special transitional vocalic microsegments (cf. category d).

d) Transitional segments. These are vocalic microsegments for linking vowels, including the glottal stop for vowel onsets that are marked juncturally . These have a fixed length of 25ms and are excised from corresponding vowel-to-vowel sequences. With the exception of the glottal stop, they are defined by denoting the left-hand vowel in brackets and the right-hand vowel as context: $(V_1)V_2$ (see section 5).

3.SYMBOL-TO-SIGNAL TRANSFORMATION

The principle of microsegment concatenation requires, in addition to the normal grapheme to phoneme transformation (lexicon-based in the present system), a set of rules for phoneme-to- microsegment transformation. These take the form of a di-phonemic input string in which the segmental identity and, in most cases, either the preceding or the following context are specified. This di-phonemic string determines the microsegment to be selected. Different degrees of stress require segments of different duration. The procedure for rhythmic structuring at this level is described in section 4 below.

The context specification can be summarised as follows:

Pre-context: V in CV sequence; /f, x, l /

Post-context: Plosives, affricates; /v, s, z, ʃ, ʒ, ç, j, h, ʁ, n, m, ŋ /

No context: quasi-stationary vowels, and transitional segments.

An example of the concatenation principles is given in section 5 .

4. PROSODIC STRUCTURING

Prosodic rules are limited at present to duration and F0 modification for rhythmic and melodic structuring, though rules for syllable- and intonation- phrase-based intensity differentiation are being considered, since they present no problems to the time-domain processing principle.

DURATION

Duration differences are used for three phonological length oppositions, namely the /a/ - /a:/ (e.g. Staat - Stadt; state - town), /e: / - / / /e.g. beten - bitten; pray - request), and /o:/ - / / (e.g. Pole - Pulle; Pole - bottle) which do not need to differ in timbre in Standard German. Prosodic duration rules are used at two levels of structuring, a) for the realisation of word stress, and b) for phrasal and utterance phenomena

All durational variants of a microsegment, from completely unaccented to finally lengthened accented syllables, are taken from one signal segment, which has multiple markers; six durations per signal segment are set manually, and, for mnemonic reasons are given the 2 main labels "normal" and "long" x the 3 subsidiary terms "unstressed", "standard" and "stressed". They are set on period-initial positive zero-crossings in periodic signals (this constraint does not apply to non-periodic segments), and progressively shorten the signal-segment from the centre of the original phone towards the segment boundary, i.e. towards the initial or the final transitional part of the phone. In figure 1 below, this means that the longest microsegments will extend from 6 to 0, the shortest from 1 to 0. The six steps are set relative to the duration of the segment, not in absolute terms. The context-free and the transitional vocalic segments are not subject to durational modification in this way.

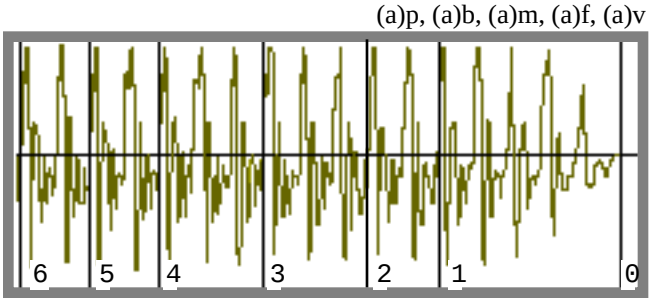

Fig. 1 Duration markers for microsegment used in /a/ - labial context. 0 = end of microsegment, 1 = normal/ unstressed, 2 = normal/ standard, 3 normal/ stressed, 4 = long/ unstressed, 5 = long/standard, 6 = long/ stressed.

Sub-category-inherent durational differences, such as shorter basic durations for higher than for low vowels, have not yet been implemented, since informal auditory assessment indicates that,

while statistically demonstrable under controlled, laboratory conditions, such intrinsic differences are swallowed up by word- and phrase-prosodic variation, and do not significantly affect intelligibility. Other aspects of variation in segmental durations, such as reduced consonant duration in clusters, and syllable-initial vs. syllable- final consonant duration, have been catered for at the level of word prosody (see below).

In the vocalic segments, the shortening of the signal from the centre of the phone has the additional advantage that natural target undershoot in shorter segments is approximated. In addition the psychophysical relationsship between duration and perceived loudness supports the use of originally accented segment-material in unaccented contexts (compare Drullman & Collier, 1991)

WORD PROSODY

Word-stress patterns are defined primarily in the word-stem lexicon. However, specific stress-attracting and stress-repulsing affixes are defined in the affix lexicon for orthographic to symbol-phonetic conversion in cases where the word is not contained in the lexicon. The syllabically defined word stress is transferred to the onset and the nucleus segments of the syllable, but not to the coda, reflecting to some extent the initial-final consonant- strength differences found in natural speech (Campbell, 1993). The above principle of stress transfer from syllable to microsegments is illustrated in the example "Frauenheld" in section 5.

To maintain rhythmic relations between syllables despite different syllable structure, base-durations are reduced for consonants in clusters. This, of course, only applies to homo-syllabic not to hetero-syllabic consonant sequences.

PHRASAL PROSODY

Superimposed on the word-prosodic structure is a second level, phrasal prosody. This caters for intermediate-phrase, and intonation-phrase grouping by introducing phrase-final lengthening and, in the case of the intonation phrases, phrasal pauses. In addition, the accented words in each phrase are determined on the basis of local rules derived from patterns of word-category sequences and key words, identified on the basis of part-of-speech tagging in the lexicon. Accented words thus determined are then given increased duration at microsegment level. Correspondingly, function words are reduced in duration relative to all lexical words.

The lengthening rules are arranged so that they reflect the different syllable-internal lengthening patterns found in accented words vs. phrase-final words (Campbell ,1992, 1993; Berkovits, 1994)

MELODIC STRUCTURE

Intonational structuring is also carried out in the time domain, on all segments defined as "voiced", by modifying the duration of each period during the part of the cycle which can be assumed to correspond to the the open-glottis phase, i.e., where resonant information is at a minimum, and consequently spectral distortion has negligible effects on signal quality. Modification is carried out by using a double-sampled signal which can be partially or totally played out with single, triple, or quadruple samples, depending on the F0 manipulation needs. This approach is only possible because, as already stated, all recordings are monotone, allowing the automatic sequential identification of individual periods. Naturally, careful selection of the speaker is of paramount importance, a) with regard to voice quality, and b) the persons ability to produce the material on a monotone, avoiding undue intensity fluctuations. Any slight fluctuations in either amplitude or F0, which are of course inevitable to a certain extent, have been corrected after microsegment excision, prior to final storage of the segments for use.

5. EXAMPLE UTTERANCE

The word "Frauenheld" contains 9 phonemes /'fåa¥ «nhElt/. The speech signal is shown in figure 2 with segments marked. The microsegment structure is explained in the following:

The first two microsegments ...(f) and (å)a are consonantal segments whose contexts are only defined on one side. They also form a cluster and therefore have reduced individual durations. After the fricative (å)a, there follows a å(a) segment, the first part of the diphthong /a /, which includes a transition from a dorsal consonant towards an /a/. A quasi-constant a(a) microsegment forms the initial target of the /a / diphthong. a (a) is the perceptually important transition subject to durational modification as a function of accentual level, from the first to the second diphthong target ( ). ( )« provides the transition from /a / to /«/. This would be followed by V(V) in other vowel-to-vowel sequences, but it is excluded in vowel-to-schwa and vowel to / / sequences because it makes the unstressed vowel too long and prominent The («)n transition follows immediately, and (n)h is the unilaterally specified nasal segment. Following a consonant, (h)E only has the following context specified; the aspirated onset to an /E/ vowel is the phonetic realisation of /h/ in this context. It is followed by a quasi- stationary E(E) microsegment. (E)l is the transitionary second half of /E/ leading to the apical lateral E(l), which has only the vocalic pre-context specified. The /t/ is formed from a period of silence for the closure t(t) and a release noise (t)..., which is followed by a pause (...). As mentioned above, the syllabic stress definition in the phonemic string ./'fåa¥ «nhElt/ is passed down to the onset and nucleus, i.e. to the microsegments ...(f) (å)a a(a) a (a) ( ). Since there is no coda in this particular syllable, the rule of coda exclusion does not apply.

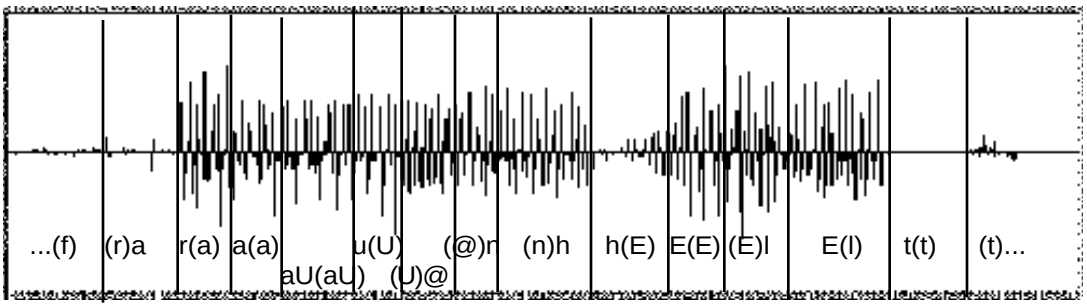

Fig. 2 Segmentation of 'Frauenheld' /'fåa¥ «nhElt/ in Microsegments

6. SUMMARY AND CONCLUSION

The synthesis system described above illustrates the application of speech knowledge to the practical problem of reducing memory requirements and demands on processing time without sacrificing synthesis quality unduly.

Within a general signal concatenation approach, a combined phone- + sub-phone-level inventory of signal segments is employed which respects consonantal place cues and other transitional information contained in vocalic signal portions and exploits the less context-sensitive aspects of consonant realisations.

Rhythmic structuring makes use of psychoacoustic constraints to allow reduced duration signal portions taken from stressed syllables to be used in unstressed contexts. The same durational flexibility is applied to consonants to achieve correct durational patterns for different syllable structures. Melodic structure is obtained in a simple pitch-synchronous procedure which uses the open phase of each period to change its duration.

As a result of the multiple use of each microsegment and the strict time-domain processing, the memory requirements and the processing load are kept to minimum without undermining the real-time performance of the system. The inventory for German comprises 330 microsegments which, at a sampling rate of 22 kHz and 8 bit resolution, require 700 KB of memory, uncompressed.

The microsegment principle has the additional advantage of reducing the cost of extending a system to new voices. The time required for recording and for segmenting the (reduced amount of) speech material is only a small fraction of time needed for producing traditional diphone or demi-syllable material.

7. REFERENCES

Berkovits, Rochele (1994): Durational effects in final lengthening, gapping, and contrastive stress. in: Language and Speech 37, S. 237-250

Campbell, W.N. (1992) Syllyble based segmental duration. in: Bailly, G. & C. Benoit: Talking Machines. Theories, Models, and Designs. Amsterdam, London, New Yoir, Tokyo. pp 221-224

Campbell, W.N. (1993), "Predicting segmental durations for accommodation within a syllable-level timing framework" in: Proc Eurospeech'93, Berlin, 1081-1084.

Drullman, Rob & Collier, René (1991) On the combined use of accented and unaccented diphones in speech synthesis. in JASA 90(4), pp. 1766-1775

Meyer et al. (1993): PHRITTS- A text-to-speech synthesizer for the German language. in: Proc. Eurospeech 93, Berlin, pp. 877-880

Posmyk, Reinhard (1989): Time-domain synthesizer for preserving microprosody. in: Proc. Eurospeech 89, Paris, pp. 191-194

Taylor, P.A. et al. (1991): A real time speech synthesis system. in: Proc. Eurospeech 1991, pp. 341-344

8. SPEECH FILES

1. [SOUND A586S01.WAV]
Paris. Steffi Graf hat das Finale des Damentennisturniers von Paris gewonnen und ist damit wieder Weltranglistenerste. Sie schlug Marie Pierce aus Frankreich mit 6 zu 2 und 6 zu 2.

2. [SOUND A586S02.WAV]
Am Rande eines Hochdruckgebietes über dem östlichen Mitteleuropa bestimmt Warmluft unser Wetter. Die Vorhersage: Heute früh bilden sich - insbesondere in den Flußtälern - Nebelfelder, die sich im Laufe des vormittags auflösen. Dann ist es allgemein sonnig und trocken.

3. [SOUND A586S03.WAV]
Fischers Fritze fischt frische Fische. Frische Fische fischt Fischers Fritz.

Further synthesized utterances can be accessed via www
http://coli.uni-sb.de/phonetik/projects/Sprachsynthese.html