MICROPHONEMIC METHOD OF SPEECH SYNTHESIS

Konrad Lukaszewicz

Institute of Biocybernetics and
Biomedical Eng. IPPT PAN, ul. K.R.N. 55,
00-818 WARSAW, POLAND

Matti Karjalainen

Helsinki University of Tech.,
Acoustics Lab., Otakaari 5 A,
02150 Espoo, FINLAND

ABSTRACT

Waveform concatenation has been the method used in many low-quality speech synthesis experiments. The objective of this study was to find new ways to overcome the inherent difficulties in concatenating speech sample waveforms. Our experiments in Finnish and Polish show that it is feasible to develop simple and inexpensive synthesizers with natural and high-quality human-like characteristics. The implementation can be based on standard microprocessors and D/A-converters without expensive signal processing hardware. This paper describes the results of the experiments and conclusions to the design of speech synthesis that we call the microphonemic method.

INTRODUCTION

The methods of speech signal generation in speech synthesis can be divided into two main classes: *model-based* (source-filter models, i.e. formant and LPC-synthesis) and *waveform-based* (time-domain synthesis) methods. The advantage of model-based synthesis is the flexibility of generating an infinite number of signals according to parametric controls that can be computed by rules, tables etc. This has become the major method especially in speech synthesis by rule.

Time-domain synthesis can be based on a collection of speech signal units like waveform cycles and events, sound segments, phonemes, diphones, syllables etc, taken from real speech. Waveform concatenation of real-speech samples is a simple method that has been used in speech synthesis experiments of low to moderate quality. In principle the sound quality could be very high if it were possible for enough samples of natural speech to be stored and carefully combined. This method needs more memory and faster data rates than model-based synthesis but otherwise it is not as complex and computationally intensive.

A well known example of time domain speech synthesis is the *Mozer* method /1/, where pitch-period-sized prototype units of real speech are manipulated to take as little memory as possible but are still able to be reconstructed in an intelligible form. This moderate quality, low bit rate method is used in some limited vocabulary synthesizers. Our experiments show that the tricks like zero-phasing the signal to lower the bit rate tend to remarkably reduce the quality and speaker identity. The phase properties that are retained are important for very high quality, natural sounding speech in a similar way as in multipulse LPC-coding /2/ .

The term "**microphonemic method**" that is used in this study was adopted from early experiments of similar principles in Poland. *Patryn* /3/ synthesized with phonemic units without transitions and pitch changes. His work was continued by *Kielczewski* in his doctoral thesis (1979). This microphonemic method applied pitch changes for intonation and transitions by mixing parts of neighbouring phoneme prototypes. *Lukaszewicz* et al. have worked on the method at the Institute of Biocybernetics, Warsaw, since 1980. Their synthesizer has found applications in a talking typewriter and a talking calculator.

The quality of speech produced by all of these synthesizers has been low to moderate. The objective of this study was to find methods to overcome the inherent difficulties in concatenating speech sample waveforms. Our experiments showed that it is feasible to develop simple and inexpensive synthesizers with natural and high-quality human-like characteristics.

PROBLEMS ENCOUNTERED IN THE USE OF WAVEFORM CONCATENATION

The main idea of the microphonemic method is based on modelling the time domain signal by using a dictionary of prototypes. These are derived from natural speech utterances and their size can be of different lengths. It is possible to store whole words, syllables, phonemes (allophones) or shorter segments. Using a dictionary of microphonemes and several rules it is possible to generate synthetic speech by concatenating prototypes one after another. Waveform interpolation and concatenation are applied to realize the transitions between consecutive units. There are several problems that need to be solved in order to obtain high quality synthetic voice, e.g.:

CH2396-0/87/0000-1426 $1.00 © 1987 IEEE

- * realizing dynamic and static variations of the units, especially in the generation of smooth and natural transitions between consecutive segments and phonemes,
- * synthesizing consonants, like tremulants (Finnish /r/), etc.,
- * modifing parameters to control intonation, stress and rythm,
- * determining the prototype set which is needed for a good representation of natural speech,
- * extracting these prototypes and their positions in the uttered speech examples,
- * formulating a good strategy when using waveform concatenation for synthesis by rule.

Some of these problems were studied by us at the Helsinki University of Technology, Acoustics Laboratory, by using the following experimental techniques.

WIDE FORMANT TRANSITIONS

One of the main problems to be solved in high-quality waveform concatenation is to realize wide formant transitions e.g. between vowels (/ui/ in Finnish) and in glides. The original idea of the microphonemic method was to apply simple linear interpolation from one pitch prototype to another by amplitude mixing (see Fig. 1). In our experiments we found that this works satisfactorily only if the glide in formant frequencies is less than 2 Barks (critical bands). If the transition is wider, the amplitude-based interpolation is not sufficient to introduce a perceptually acceptable formant movement effect. For highest quality speech even 1 Bark transitions may be needed.

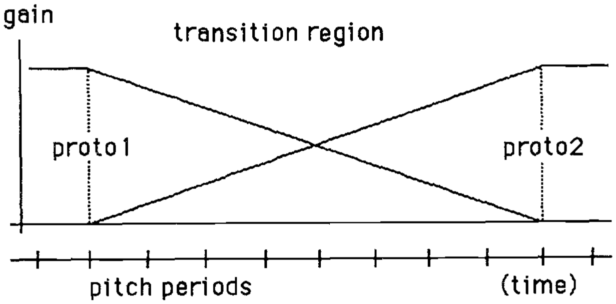

Fig. 1. *Linear amplitude-based interpolation between two pitch-sized prototypes to simulate formant transitions.*

If the formant distances between sound segments larger than 2Barks are needed, some intermediate prototypes should be used to interpolate through (see Fig. 2.). It was possible for all transitions found in Finnish and Polish to be synthesized in this way.

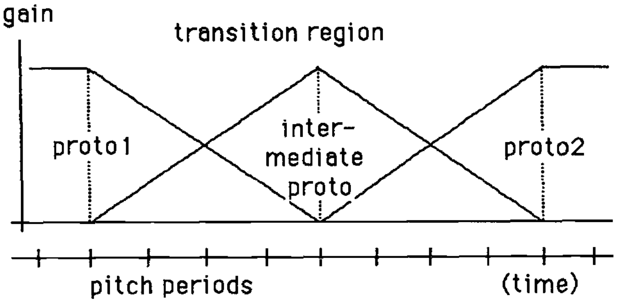

Fig. 2. *Linear amplitude-based interpolation between two pitch-sized prototypes with an intermediate prototype.*

CONSONANT SYNTHESIS

Many consonant classes need special processing. Short non-repetitive units like bursts in stop consonants can be stored as direct waveform segments and as several variants in the context of different vowels or vowel groups. Sometimes the effect of neighbouring consonants must also be analyzed and the context stored for synthesis.

Fricatives need special treatment, too. Prototypes of about 50 ms in total length were found to be suitable and 10 ms units from them were randomly taken for concatenation. The same interpolation rule as in vowels can be applied. Most voiced consonants behave in the same manner as vowels except that the variability according to the context is only higher.

PITCH AND INTONATION CONTROL

Prosodic features reveal some difficulties in concatenation. A simple and fairly successful method to control intonation is the use of minimum-pitch-period-sized prototypes and insertion of zero-signal segments to obtain the desired effective pitch for each moment (Fig. 3). A suitable windowing technique and the overlapping mixing of pitch periods could improve the results still further (Fig. 4). Timing is controlled by counting a proper number of pitch periods.

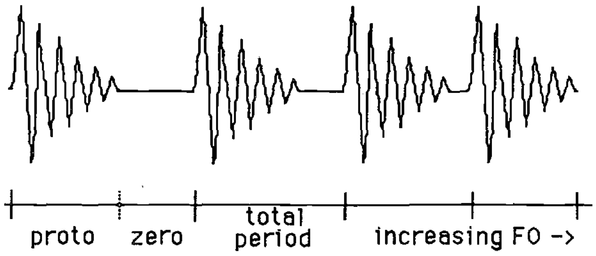

Fig. 3. *Zero signal insertion as a method of controlling pitch in concatenation.*

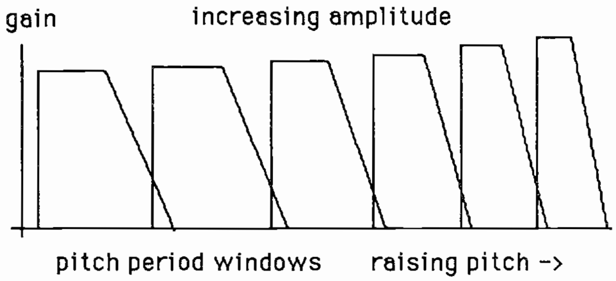

Fig. 4. *Overlapping window summation in pitch control.*

EXPERIMENTAL STUDY

About 70 Finnish and Polish phoneme pairs, concentrating on the synthesis of diphone-like transition segments, were studied experimentally by the microphonemic method. Some other larger units (syllables, words) were also modelled.

A microprocessor-based signal editor (SPS-02) was used to extract the prototype units from real speech. The same editor system was further applied to scale the amplitude, adjust the pitch period and to mix the prototypes for concatenation and synthesis experiments. Another analysis system, ISA /4/, with auditory spectrum and spectrogram display was used to pick up the best positions of the prototypes and to compare the original against the synthetic speech examples. The principle of the auditory model for this analysis is presented in /5/.

Prototypes from the original speech were used to model the phoneme pair transitions with two different principles of prototype selection. The first one was for producing intelligible, moderate quality speech with minimum number of prototypes, which were located in the middle of the quasiperiodic steady-state phonemes and one prototype in the middle of the transition region.

The other method was to produce higher quality speech with a larger number of prototypes. This was accomplished by choosing the prototypes at each point where the formant frequencies started changing. If the change was larger than 2 Barks an extra prototype between the starting and ending points was taken. The maximum difference in any formant transition to be interpolated was always less than 2 Barks. A prototype was selected also at the points where the formants changed their moving direction. For a full synthesis system some of the intermediate prototypes may be selected so that they can be used in several contexts.

As an example, the number of prototypes in the Finnish diphtong /ia/ was three for intelligible and seven for high quality speech. The maximum number of prototypes was never larger than nine for any diphone-like unit. The size of a prototype was usually equal to one pitch period. However, in the case of stop consonants the length of a prototype was two to five times higher and for fricatives also five times higher.

Fig. 5. shows the auditory spectrogram of the original diphtong utterance /ia/ with the related loudness function. Vertical lines with the capital letters A to C mark the positions of the prototypes in the lower-quality experiment. The auditory spectrogram of the synthesized version is shown in Fig. 6. Lines related to digits 1 through 7 in Fig. 5 indicate the places of the prototypes in the case of the highest quality reconstruction. The corresponding auditory spectrogram is in Fig. 7.

The pitch-sized prototypes from real speech retain automatically some speaker-specific features and personality of the voice. The time-domain signal carries the tone quality features related to the detailed amplitude and phase spectrum. Our experiments show that the phase, especially the rapid phase transitions can be very important to the naturalness of some allophones (nasals, liquids etc.) and their combinations. The prototypes may also include inherent pitch and amplitude data of the allophones that will be modified according to the context during the resynthesis.

IMPLEMENTATION ASPECTS

An estimate of memory capacity that is needed for prototypes in a moderate-quality synthesizer (Finnish) is: some 30 "phonemes", in average 8 variants (vowel contexts), and the same amount of intermediate prototypes. This results in a total number of about 500 units, each of 12ms times 14kHz samples (8 bits), which equals to less than 100 kilobytes. At the level of the present ROM-memory technology it is feasible to use up to 256 kbytes of memory for the prototype storage and synthesis rules, thus to achieve a high-quality synthetic speech with personal-sounding voice.

A single microprocessor like Motorola 68000 is capable of doing the synthesis in real time. Serial and/or parallel ports are needed for input and a single D/A-converter (8 to 12 bits) with a reconstruction filter may form the analog output. Another way is to design with multiplying D/A-converters to avoid software multiplications for amplitude scaling in the interpolation. The microphonemic method is also well suited to software-based speech synthesis in microcomputers with special D/A-hardware to support fast analog output.

The software for the microphonemic synthesis by rule can be based on the manipulation of prototypes as allophonic symbols. The dictionary of microphonemes is searched for the best candidate prototypes for the context and these are then interpolated and concatenated to countinuous speech.

The selection of the prototypes during the development of the system is a laborious and critical task, that is difficult to be automated. A semiautomatic segmentation algorithm and pitch period detector could help if the voice of several speakers must be modelled. We are working on the design of two different development systems to continue the studies on the microphonemic method. One will be based on a personal computer,

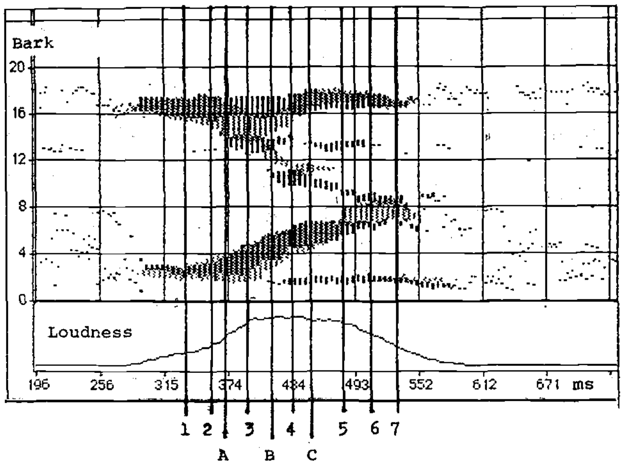

Fig. 5. *Auditory spectrogram of the original speech, (Finnish diphtong /ia/)*

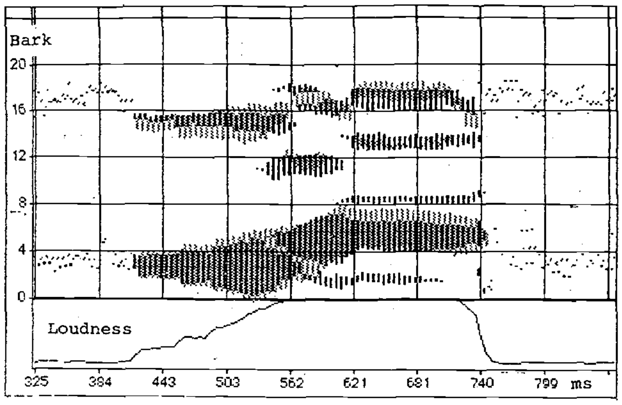

Fig. 6. *Auditory spectrogram of the lower-quality reconstruction by the microphonemic method with three prototypes (A, B, C in Fig. 5.)*

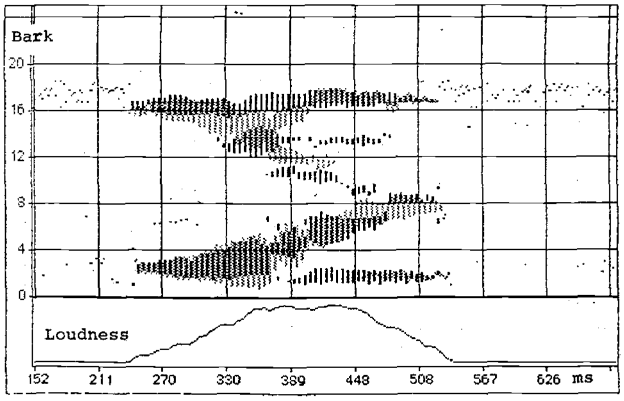

Fig. 7. *Auditory spectrogram of the higher-quality reconstruction by the microphonemic method with 7 prototypes (1 to 7 in Fig. 5.)*

another on Symbolics 3670 Lisp-machine with signal and array processor hardware.

CONCLUSIONS

Our experiments showed clearly that the microphonemic method by waveform interpolation and concatenation has potential to high-quality speech synthesis by rule. Its main technical advantage is that no computationally intensive signal processing is needed. The crucial question to achieve the highest-quality is to extract the optimal collection of prototype segments from real speech and to have a good startegy for the rule-based concatenation. Auditory spectra and spectrograms were found important in the extraction process to find the best segmets that meet the requirements of the human auditory perception.

REFERENCES

/1/ **Costello B.C., Mozer F.S.**, Time-Domain Synthesis Gives Good-Quality Speech at Very Low Data Rates. Speech Technology Sept/Oct. 1982, p. 62-68.

/2/ **Atal B.S., Remde J.R.,** A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates. Proc. of ICASSP-82, Paris, p. 614 - 617.

/3/ **Patryn R.**, Transitionless Synthesis of Speech. Acoustica Vol. 48 (1981) no. 4, p. 275-276.

/4/ ISA, Intelligent Speech Analyser, Instruction Manual, Vocal Systems, Finland, 1987.

/5/ **Karjalainen M.,** A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of ICASSP-85, Tampa 1985, p. 608-611.