DIPHONE SYNTHESIS USING AN OVERLAP-ADD TECHNIQUE FOR SPEECH WAVEFORMS CONCATENATION

F.J. CHARPENTIER and M.G. STELLA

Centre National d'Etudes des Telecommunications
22301 LANNION FRANCE

ABSTRACT

A new method is presented for text-to-speech synthesis using diphones. The diphone database consists of the diphone waveforms labeled with pitch-marks indicating the pitch-periods. At synthesis time, the diphone waveforms are processed through a new analysis-synthesis system, providing an independent control of all prosodic parameters, while retaining a good degree of naturalness. This system is based on a representation of the speech signal by its short-time Fourier transform (STFT) at a pitch-synchronous sampling rate. The synthesis part of the system works by overlap-adding the modified short-term signals and it ensures a smooth concatenation of the diphone waveforms. The synthetic speech obtained by this method sounds more natural than with the conventional LPC method.

INTRODUCTION

The most common technique used for diphone or demi-syllable synthesis is linear predictive coding of the speech elements. This scheme is very convenient because it provides a flexible control of the prosodic parameters through the excitation signal of the synthesis filter. Unfortunately, LPC coded speech suffers from a lack of naturalness, especially for sounds like voiced fricatives. Many attempts to enhance the quality of LPC speech were made through improving the nature of the excitation signal. Most recently, researchers have used the multipulse model of the excitation signal [1,2], but this model is not well adapted to the manipulation of prosodic parameters. In a similar experiment [3], we obtained a good degree of naturalness by a raw concatenation of multipulse coded diphones. The prosody of the synthetic speech was then corrected by use of a high quality system to modify the pitch and the duration of natural speech, based on short-term Fourier synthesis [4]. Unfortunately, the combination of this system and of the LPC synthesis scheme introduced an undesirable hoarseness of the speech quality.

In this paper, we propose to build a text-to-speech system entirely on a short-term Fourier representation of the speech signal. Consequently, the LPC synthesis filter is replaced by a short-term Fourier synthesis scheme. A lot of systems based on modifications of the STFT have been proposed to manipulate some of the speech parameters while retaining a high degree of naturalness. The earlier systems used the filter bank summation (FBS) approach, but this usually led to excessively heavy or intricate algorithms [4,5]. The overlap and add (OLA) approach was preferred in our case because it yields faster and simpler algorithms [6,7]. The OLA methods generally consist of splitting the speech signal into overlapping short-term signals (ST-signals) at a certain analysis frame rate, modifying them appropriately and then overlap-adding them at a different frame rate. A non-uniform frame rate at the synthesis stage, as recently proposed by the synchronized OLA method (SOLA), can produce high-quality speech with a modified time-scale [8]. This non-uniform frame rate corresponds to a sort of resynchronization of the modified ST-signals with the pitch of the output signal.

The pitch-synchronous OLA method (PSOLA) presented in this paper is an extension of the SOLA method, in which the pitch-synchronization of the ST-signals is introduced at the analysis stage. Every ST-signal is made to correspond to one pitch period of the speech signal, both at the analysis and synthesis stages. This implies a preliminary segmentation of the speech signal into individual pitch periods. In this respect, the PSOLA method is similar to a previous method for the time-scaling of speech, the time-domain harmonic scaling (TDHS) algorithm [9]. However, the analysis rate is not exactly pitch-synchronous in the TDHS algorithm, but it is rather proportional to a true pitch synchronous rate, and it depends on the time-scaling factor in complicated way.

A major novelty of the PSOLA method is the integration in the overlap-add scheme of an algorithm for modifying the pitch and the global spectral shape. This is obtained by relatively simple operations in the frequency domain, that do not require any phase computation. Finally, the PSOLA analysis-synthesis system provides a flexible tool for an independent and time varying control of all prosodic parameters, and it is therefore a suitable candidate for a text-to-speech system.

CH2243-4/86/0000-2015 $1.00 © 1986 IEEE

THE ANALYSIS-SYNTHESIS SYSTEM

In the PSOLA analysis-synthesis synthesis system, the speech signal is analyzed into a succession of pitch-synchronous ST-signals. These analysis ST-signals are then modified, either in the time or in the spectral domain, in order obtain a sequence of synthetic ST-signals, synchronized with a modified pitch contour. Finally, the synthetic speech is obtained by overlap-adding the synthetic ST-signals. We now describe in more detail the different components of the system.

Pitch-synchronous overlap analysis

A reasonable length of the analysis window is chosen, in order to cover at least three times the longest pitch period. In our experiments, a 512 point Hamming window, corresponding to a usual 30 ms duration, has proved to be suitable in processing speech at a 16 kHz sampling rate.

The initial representation of the speech signal is the speech digital waveform with successive "pitch-marks" distributed along the time-scale. These pitch-marks (P-marks) are set synchronously with the pitch periods over voiced portions, but their positions are arbitrary over unvoiced portions, provided that two analysis windows centered on two successive P-marks should overlap by a sufficient amount. The analysis ST-signals are obtained by multiplying the signal by the analysis window centered on the corresponding P-marks. Each ST-signal is labeled with a voiced/unvoiced flag associated with its P-mark.

In this manner, the signal is expanded into a sequence of overlapping, pitch-synchronous, short-term signals. The unvoiced ST-signals are not converted to the frequency domain since they will only need time-scaling. But the representation has to be further refined for the voiced portions: a short-term spectrum is computed by DFT with the time origin set to coincide with the P-mark; it is then split into a global spectral envelope, and a "source component", which is the short-term spectrum divided by the spectral envelope.

Frequency-domain modifications

The pitch and envelope modifications algorithm is described in Fig.1. They are obtained through a frequency-domain processing of the ST-signals. The source or the envelope components of the spectrum are interpolated separately in order to obtain two separate rescalings of the frequency axis. In the case of the source component, a linear interpolation is performed on both the real and imaginary parts of the spectrum. If necessary, high frequencies are regenerated by copying the lower part of the spectrum to the upper part. Such a rescaling of the source component results into a modification of the pitch. In order to avoid corrections of the phase spectrum, the spacing between the P-marks is simultaneously modified by the pitch modification factor, and this results into a modification of the time-scale. The envelope component is also open to modifications that can be useful to modify the voice quality. For instance, a female voice can be converted to a male or a child sounding voice by lowering or raising the formants by an average 15% and by suitably altering the average pitch level.

Time-scale modifications

Time-scale modifications are performed entirely in the time domain. The time-scaling procedure performs directly on the unvoiced ST-signals but on the voiced ones only after processing by the pitch modification procedure. In the latter case, the algorithm must compensate for the time-scale modification implicit to pitch modification.

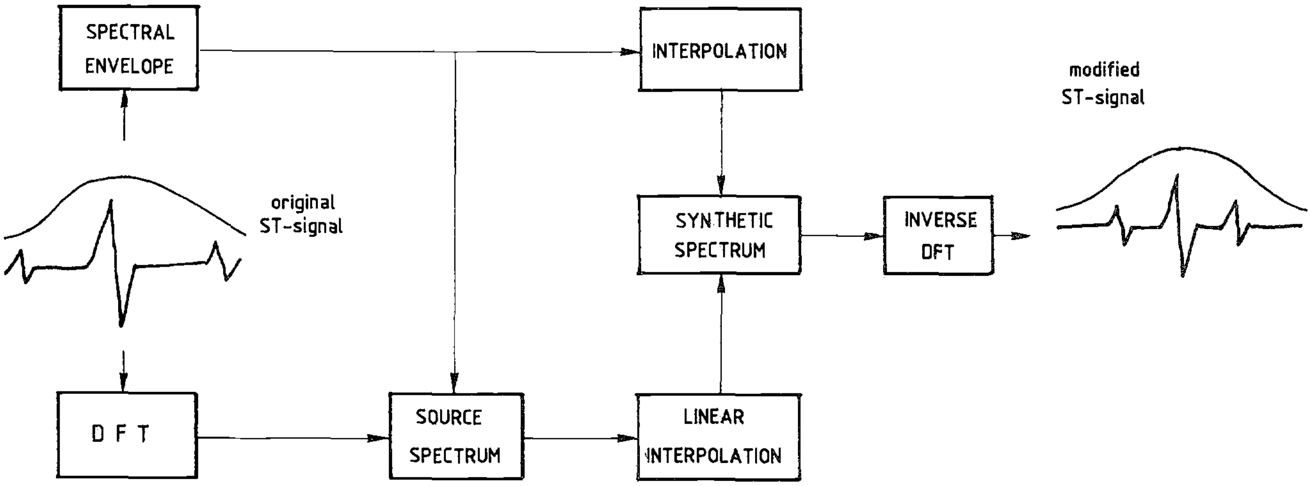

Fig.1 Block diagram of the frequency-domain modifications procedure.

The algorithm is illustrated in Fig.2. The desired time-scale modification can be defined as a time-warping function mapping the analysis time-scale onto the synthesis one. From the distribution of analysis P-marks, the algorithm generates a new set of synthesis P-marks in such a way as to preserve the pitch contour through the time-warping function. Generally, a synthesis P-mark will not correspond exactly to an analysis P-mark. Therefore an approximation scheme is needed to assign a reasonable ST-signal to every given synthesis P-mark. The simplest solution is to assign the analysis ST-signal corresponding to the nearest analysis P-mark through the time-warping function. This is equivalent to either cancelling or duplicating some of the pitch periods. A slightly more complicated scheme has been used here, consisting of time averaging the two nearest analysis ST-signals.

Pitch-synchronous overlap synthesis

The synthesis of non-unformly sampled short-term spectra is possible by use of the least-squares OLA synthesis procedure [7]. The synthesis procedure is described in Fig.3. It consists of two overlap and add buffers that are rotated simultaneously by a variable number of samples P, equal to the current synthesis pitch period. The synthesis ST-signals are windowed and overlap-added in the first OLA buffer and the corresponding synthesis window is squared and similarly overlap-added in the second OLA buffer. After each overlap-add operation, the OLA buffers each produce P samples, and the P signal samples output by the first OLA buffer are rescaled by the P weighting coefficients of the second OLA buffer, to produce the final P samples of synthetic speech. The two OLA buffers are eventually right-shifted by P samples and the missing samples are replaced with zeroes.

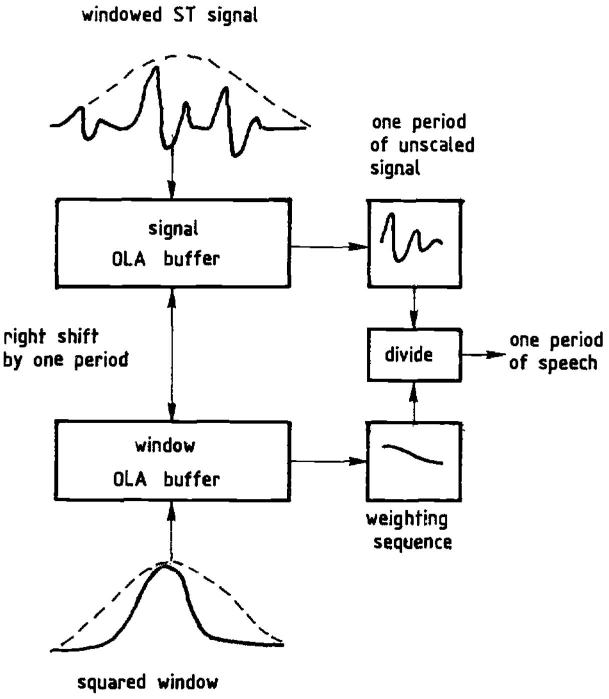

Fig.3 Least-squares OLA synthesis procedure.

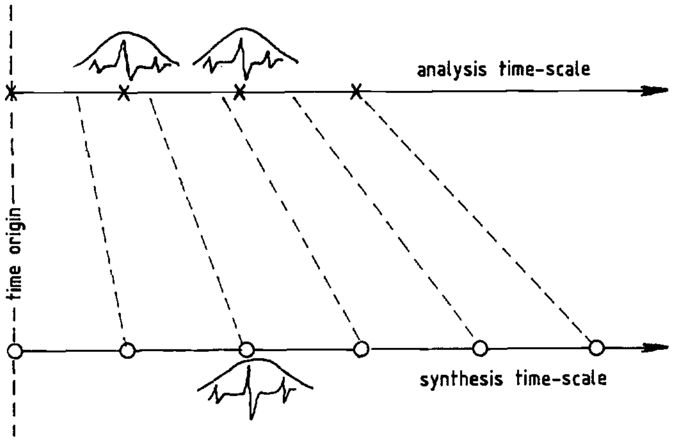

Fig.2 Time-scale modification scheme.
Analysis and synthesis P-marks are indicated respectively by crosses and circles; the dashed lines represent the time-warping function.

THE DIPHONE SYNTHESIS SYSTEM

This analysis-synthesis system has been applied to diphone synthesis of French. The block diagram of the text-to-speech system is shown in Fig.4. The diphone database consists of two parallel sources of information: the diphone waveforms on one hand, the corresponding P-marks on the other. The labelling of the pitch periods was obtained for the 1200 diphones dictionary by use of a semi-automatic procedure. The P-marks were set in a consistent manner so as to coincide with the greatest peak of each period, corresponding roughly to the instant of glottal closure. Voiceless portions were segmented into a number of nominal length windows.

At synthesis time, the diphone waveforms are concatenated as follows:

(a) each diphone is processed by the PSOLA analysis-synthesis system in order to correct the timing and to fit a desired pitch contour; the prosodic parameters are obtained by the prosodic module of our conventional LPC diphone system [10]; modification parameters for the pitch and the time-scale are obtained by comparing the desired prosody to the intrinsic prosody of the diphone;

(b) at the boundary between two successive diphones, the ST-signals of the following diphone are simply concatenated to those of the previous diphone, and the smoothing is ensured by the overlap-add synthesis scheme; the transition can be lengthened by creating a variable number of additional periods obtained by time averaging the two boundary ST-signals; an energy correction is made on the following diphone to eliminate a possible energy mismatch.

RESULTS AND DISCUSSION

Two diphone dictionaries, corresponding to a female and to a male speaker, were prepared in order to test our text-to-speech method. The synthetic speech obtained with both voices sounds more natural with our new method than whith our conventional LPC method. It does not suffer from the usual buzzyness of LPC speech. It also posesses an improved quality in comparison with the synthetic speech previously obtained from multipulse coded diphones [3]. There is still a slight roughness and it is probably due to the mismatch of the formants. Further improvements are expected since the system provides a flexible control of all prosodic parameters and of the spectral envelope.

The PSOLA algorithm runs on an array processor in 20-times real-time for natural speech sampled at a 16 kHz rate. With the diphone synthesis procedure, a second of synthetic is generated in approximately a minute. The memory space required by the diphone dictionaries is quite large (7 Mbytes). By using appropriate waveform compression or coding schemes, it is likely that the storage for the speech database could be reduced to a size comparable to that of the lexical database, usually needed to perform high quality text-to-speech conversion.

CONCLUDING REMARKS

A pitch-synchronous OLA algorithm has been presented in this paper for high quality prosodic modifications of natural speech. It requires a preliminary labelling of the pitch periods using a semi-automatic procedure, but this task is tractable for diphone synthesis because the speech database is relatively limited. Consequently, a new text-to-speech system has been designed and it produces synthetic speech of improved naturalness compared to the LPC technique. With the PSOLA method, the pitch and energy mismatches between successive diphones can be completely eliminated. Still, more knowledge has to be gained in reducing the mismatch of the formant parameters. However, the main advantage of this technique is that it is capable of retaining fine spectral details of the original diphone elements without the smoothing or distorting effects inherent in the formant or LPC synthesis techniques.

REFERENCES

[1] B.E. Caspers, B.S. Atal, "Changing pitch and duration in LPC synthesized speech using multipulse excitation", JASA, 73, S5, Spring 83

[2] J.P. Van Hemert, "Multipulse Excitation: the possibilities and restrictions of a new speech synthesizer", IPO annual progress report No.19, 20-24, 1984

[3] M.G. Stella, F.J.Charpentier, "Diphone synthesis using multipulse coding and a phase vocoder", Proc. ICASSP, 740-743, Mar.85

[4] S.S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction ", IEEE Trans. ASSP, 30(4), 566-578, Aug.82

[5] J.L. Flanagan, R. Golden, "Phase vocoder", Bell Syst. Tech. J., 45, 1494-1509, Nov.66

[6] J.B. Allen, L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", Proceedings IEEE, 65(11), 1558-1564, Nov.77

[7] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform", IEEE Trans. ASSP, 32(2), 236-243, Apr.84

[8] S. Roucos, A. Wilgus, "High quality time-scale modification for speech", Proc. ICASSP, 493-496, Mar.85

[9] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals", IEEE Trans. ASSP, 27(2), 121-133, Apr.79

[10] J.L. Courbon, F. Emerard, "SPARTE: a text-to-speech machine using synthesis by diphones", Proc. ICASSP, 1597-1600, May 82

AKNOWLEDGEMENT

The authors wish to thank F. Emerard for segmenting the speech databases into diphones units.

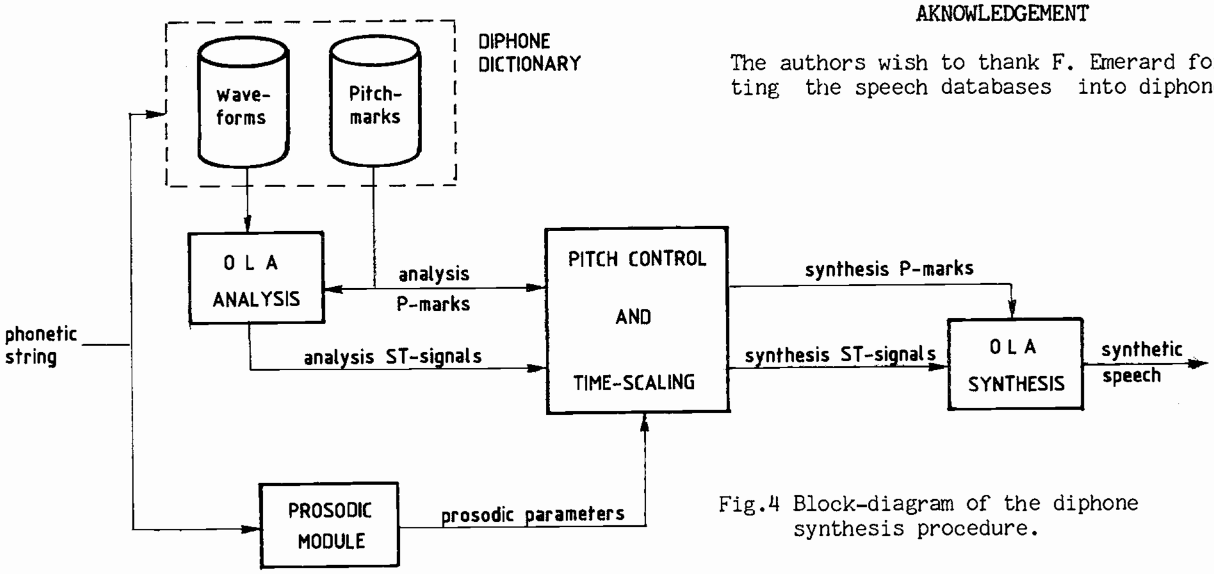

Fig.4 Block-diagram of the diphone synthesis procedure.