A DIPHONE SYNTHESIS SYSTEM BASED ON TIME-DOMAIN PROSODIC MODIFICATIONS OF SPEECH

Christian HAMON, Eric MOULINES, Francis CHARPENTIER

Centre National d'Etudes des Télécommunications
22301 LANNION FRANCE

ABSTRACT

This paper presents a new time-domain algorithm for text-to-speech synthesis using diphone concatenation. The algorithm is based on the pitch-synchronous overlap-add (*PSOLA*) approach, and is capable of good quality prosodic modifications of natural speech. The algorithm can be seen as a simplification of a previous algorithm combining the *PSOLA* approach and frequency-domain transformations. On the other hand, it appears as a generalization of previous time-domain methods that perform pitch-synchronous cut-and-splice operations on the speech waveform.

This algorithm is used in the CNET diphone synthesis multilingual system, actually supporting three langages: French, Italian and German. The resulting speech has been tested on French and is judged of much better quality than for an LPC-based synthesizer.

INTRODUCTION

Recent advances in diphone synthesis have brought about substantial improvements in the voice quality of speech synthesized from text. These advances were due to the the pitch-synchronous overlap-add *(PSOLA)* approach used for the prosodic modifications of concatenated waveforms [1]. For time-scaling of speech, the *PSOLA* approach boils down to a simple time-domain algorithm, analogous to the previous *SOLA* algorithm [2]. In this paper, we show that the time domain *PSOLA* approach is also applicable to the problem of pitch-scaling, and therefore provides a very efficient algorithm for the prosodic modifications of speech.

In the general *PSOLA* framework, the original speech signal is transformed into a sequence of overlapping short-term signals (ST-signals), obtained by pitch-synchronous windowing. This stream of ST-signals is then modified to produce a stream of synthesis ST-signals. The final synthetic speech is obtained by use of an overlap-add synthesis procedure [3,4].

In the *PSOLA/FFT* algorithm, that was proposed earlier for pitch-scaling [1,5,6], the synthesis ST-signals were obtained by frequency domain modifications of the analysis ST-signals. These modifications were performed under narrow-band conditions, in order to adjust the pitch harmonics so that the periodicity inherent in the synthesis ST-signal would be consistent with the synthesized pitch value.

In the *PSOLA/MPLPC* method proposed more recently [7], the prosodic modifications operations are applied to the multipulse excitation using a time-domain *PSOLA* processing scheme. From a computational point of view, this method is considerably simpler than the *PSOLA/FFT* method, although it still involves an explicit separation of the source and filter components.

In the *PSOLA/KDG* method presented in this paper (*KDG* means "time-domain overlap-add" in the breton language), the time-domain operations are applied directly to the speech waveform [8,9]. This algorithm is capable of pitch modifications with good quality. The computational complexity of the algorithm is very low since it does not need an explicit extraction of the short-term spectral envelope. Real-time applications on standard processors are therefore possible, provided a fast access to the digitized diphone waveforms.

In this new time-domain method, the frequency-domain operations of the *PSOLA/FFT* method are simply by-passed. In fact, the algorithm only involves two types of operations:
(1) elimination or duplication of the ST-signals;
(2) adjustment of the delays between the ST-signals to comply with the specified time-scale and pitch-scale modifications.

The algorithm can work for different values of the window length, ranging from narrow to wide band synthesis conditions. Under narrow-band conditions, the periodicity inherent in the ST-signal is not adjusted to the synthesis pitch value and this imposes a certain kind of distortion to the synthetic signal, similar to a reverberation distortion. On the other hand, wide band conditions entail a different kind of distortion, that can be interpreted as a spectral leakage that increases the formant bandwidths. The optimal synthesis conditions result from a trade-off between these to kinds of distortions. In fact, for a practical range of window length values, intermediate between narrow and wide band conditions, the distortions, although slightly perceptible, do not appear as a speech quality degradation, so that the method is well-suited for speech synthesis applications.

The *PSOLA/KDG* algorithm appears also as a refinement of previous time-domain techniques of speech synthesis, in the context of time-scaling [10] and pitch-scaling [11]. These methods are based on automatic pitch-synchronous editing of the speech signal. Individual pitch periods are excised from the original speech signal, and spliced together at a different rate to produce the synthetic signal. When raising the pitch, the pitch periods are added at a faster rate so that a certain overlap occurs between successive periods. When lowering the pitch, zeroes are inserted to fill up the gaps between periods. Such pitch period splicing is generally performed by use of short trapezoidal windows. The method presented here differs by the choice of a smoother window (e.g. Hanning window) and of a higher overlap-factor, at least of 50%.

Finally, this method has been applied to our text-to-speech system using diphones. In comparison with a *LPC*-based diphone synthesis system, our time-domain *PSOLA* system requires a much larger size diphone dictionary, but it produces a much more natural sounding quality of the synthesized speech.

1. PROSODIC MODIFICATIONS ALGORITHMS

(a) PSOLA analysis-synthesis

The digitized speech waveform $s(n)$ is decomposed into a sequence of short-term overlapping signals $s_m(n)$. These ST-signals are obtained by multiplying the signal by a sequence of analysis windows $h_m(n)$:

$$s_m(n) = h_m(t_m - n)\,s(n) \qquad (1)$$

CH2673-2/89/0000-0238 $1.00 © 1989 IEEE

In this equation, $h_m(n)$ represents a Hanning window, centered around the time origin $n=0$. The successive instants t_m, called pitch-marks, are set at a pitch-synchronous rate on the voiced portions of the signal. The stream of analysis ST-signals is then processed to produce a stream of modified synthesis ST-signals $\tilde{s}_q(n)$ synchronized on a new set of pitch-marks $\tilde{t}_q$. The correspondence between the analysis and synthesis ST-signals is specified by a time-warping function $\tilde{t}_q \rightarrow t_m$, relating the synthesis pitch-marks to the analysis ones.

The synthetic speech $\tilde{s}(n)$ is obtained by overlap-adding the stream of synthesis ST-signals, by means for instance of the least-squares overlap-add (OLA) synthesis scheme [3]:

$$\tilde{s}(n)=\frac{\sum_q \alpha_q \tilde{s}_q(n)\tilde{h}_q(\tilde{t}_q-n)}{\sum_q \tilde{h}_q^2(\tilde{t}_q-n)} \qquad (2a)$$

where $\tilde{h}_q$ denotes the synthesis windows. The use of this synthesis scheme is equivalent to minimizing the quadratic error between the spectra of the analysis ST-signals and the corresponding short-time spectra of the synthetic speech.

In fact, the *PSOLA/KDG* algorithm does not modify individually any ST-signal, so that it is reasonable to use the same analysis and synthesis windows: $h_m(n)=\tilde{h}_q(n)$, where $\tilde{t}_q$ and t_m are related by the time warping function mentioned above. Consequently, the synthesis formula can be seen as the simple overlap-add procedure [4], with the synthesis windows $\tilde{h}_q(n)$ being replaced by their squared versions $\tilde{h}_q^2(n)$. The denominator plays the role of a time variable normalization factor: it compensates for the energy modifications due to the variable overlap between the successive windows. Under narrow bandwidth conditions, this factor is nearly constant. Under wide bandwidth conditions, it can also be kept constant, provide an appropriate choice of synthesis windows. The additional normalization factor α_q is introduced to compensate for the energy modifications due to the pitch modification procedure.

Whenever the normalizing denominator can be considered constant, it is advantageous to use the simplified overlap-add scheme:

$$\tilde{s}(n)=\sum_q \alpha_q \tilde{s}_q(n) \qquad (2b)$$

(b) The PSOLA/KDG prosodic modification algorithm

The pitch-scale and time-scale modifications require an appropriate determination of the synthesis pitch-marks $\tilde{t}_q$, and of the time-warping function $\tilde{t}_q \rightarrow t_m$, mapping the stream of synthesis pitch-marks onto the analysis ones. In fact, this mapping relates every given synthesis ST-signal $\tilde{s}_q(n)$ to the analysis ST-signal $s_m(n)$ that needs to be copied in its place, and the values of the $\tilde{t}_q$ determine the delays to be used between successive synthesis ST-signals. This is summarized in the following equation:

$$\tilde{s}_q(n)=s_m(n-t_m+\tilde{t}_q) \qquad (3)$$

If the signal is to be simultaneously time and pitch-scaled by the same factor β, there will be one-to-one mapping between the analysis and synthesis ST-signals, as shown on Fig.1. The algorithm therefore simply copies each analysis ST-signal on to the synthesis time-axis, only modifying the delays by the factor β. In the general case where independent time and pitch-scaling factors must be applied, the mapping is no longer one-to-one, but either duplicates or eliminates individual analysis ST-signals.

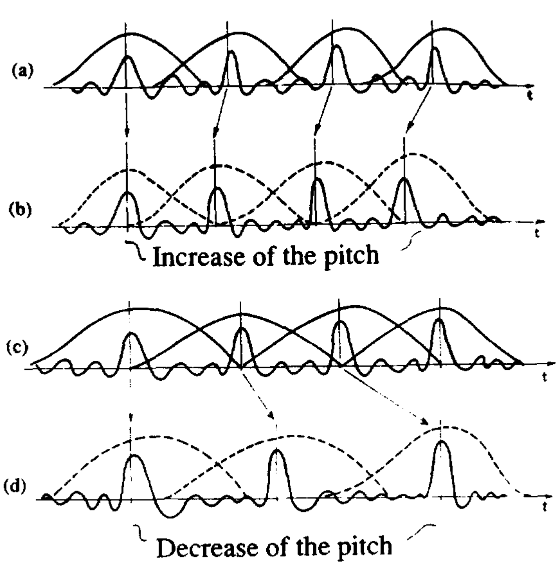


Figure 1: *The modification of the pitch of a voiced speech segment.* An analysis windowing is performed at the original pitch synchronous rate (a,c); then short-time analysis signals are assembled at the synthesis pitch synchronous rate, using the overlap-add procedure (b,d).

A consistent choice of the window length is to set it proportional to the local pitch period:

$$\tilde{h}_q(n)=h\left(\frac{n}{\mu P}\right) \qquad (4)$$

where $h(t)$ denotes the window with length normalized to unity, P indicates the local pitch period, and μ is a proportionality factor denoting the number of pitch periods spanned by the window. Such a window proportionality rule has two consequences: the spectral resolution implicit to the processing remains constant, and the normalization denominator fluctuates around a constant value related to the average overlapping factor.

In fact, two different rules are possible for the choice of P: it can either be equal to the analysis pitch period P_m or to the synthesis one $\tilde{P}_q$. It is interesting to combine the synthesis period proportionality rule ($P=\tilde{P}_q$), in which case the normalization factor α_q can be omitted, and the choice of a window proportionality factor $\mu=2$, for which the normalizing denominator is approximately equal to one.

(c) Interpretation

To some extent, it is possible to interpret the spectral effect of the *PSOLA/KDG* algorithm within the least-squares *OLA* framework corresponding to Eqn.2a. The algorithm implicitly minimizes the differences between the spectra of the analysis ST-signal and the ST-spectra of the final synthetic speech. In relatively large bandwidth conditions (μ < 2), there will be a reasonably good fit between the short-term spectra of both the original and synthetic signal. Unfortunately, this interpretation is not valid in the case of narrow bandwidth conditions (4 < μ), since the spectra that are being fitted have a different harmonic structure.

However, another interpretation is available in the framework of the simplified overlap-add method corresponding to Eqn.2b [13]. Assume the original speech is periodic with period $P_m = P_0$ and the analysis-synthesis windows are identical $h_m(n) = h_0(n)$. The analysis ST-signals are all equal to a single prototype ST-signal:

$$s_m(n) = s_0(n) = h_0(t_0 - n)s(n) \quad (5)$$

The synthetic signal is obtained by the periodization of $s_0(n)$ at the new synthesis period $\tilde{P}_q = \beta P_0$. The effect of this operation is well-known, and it is equivalent to modifying the spectral properties of the signal by sampling the spectrum $X_0(f)$ of the prototype signal at a new set of harmonic frequencies $\tilde{f}_k$. If $w(n)$ denotes an analysis window, the ST-spectrum of the synthetic signal will be given by the convolution of the spectrum $X_0(f)$ sampled at the new harmonic frequencies by the response of the window $W(f)$:

$$\tilde{X}(f) = \sum_{f_k} W(f - \tilde{f}_k)X_0(\tilde{f}_k) \quad (6)$$

In other terms, the spectral envelope of the synthetic signal is identical to the short-term spectrum of the original signal, with a spectral resolution determined by the analysis-synthesis window $h_0(n)$. Therefore, the amplitude of each harmonic of the synthetic signal is distorted with respect to the ideal spectral envelope. In the case a short synthesis window, this distortion appears as a smearing of the formant resonances (Fig.2a). In the case of a long window (Fig.2b), the spectral envelope has a harmonic structure related to the original pitch so that the pitch harmonics of the synthetic signal are all the more attenuated as they depart form an original pitch harmonic. This effect increases with the window length, eventually cancelling out certain harmonics. This distortion appears as degradations of the waveform energy contour, and the sound quality becomes reverberant.

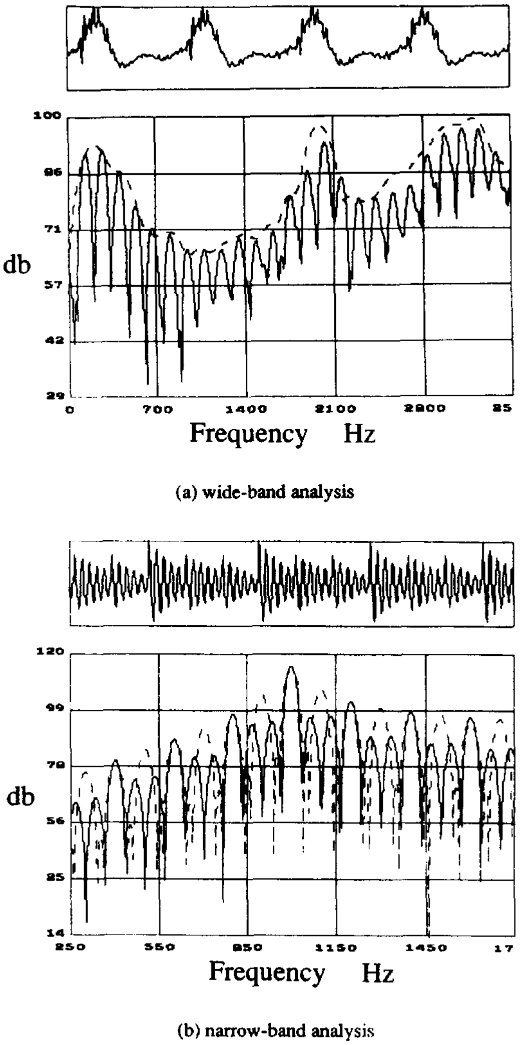
In practice, it is possible with the simple overlap-add scheme and a window factor μ = 2 to retain good speech quality. Due to the window squaring property, the least-squares procedure can be used with a longer window length (μ = 3) without introducing a reverberation effect.

(d) Implementation

Fig. 1 illustrates a specific setting of the *PSOLA/KDG* algorithm using the simplified overlap-add procedure and a window proportionality factor μ = 2. In addition, it is advantageous to apply the analysis period proportionaly rule ($P = P_m$) when decreasing the pitch, and the synthesis period proportionality rule ($P = \tilde{P}_q$) in the other case. By tabulating the synthesis window, it is possible to reduce the computational load to only 2 multiply-adds per synthesized sample. This computational effort is low enough to meet the real time constraint on a personal computer with a 386 processor.

2. DIPHONE CONCATENATION ALGORITHMS

The *PSOLA/KDG* diphone synthesis system works presently for three european languages, French, German, and Italian. It is implemented on a general purpose computer (microVAX II) and on a personal computer. The waveform dictionaries consist of about 1200 diphones for French (male and female voice), of 1200 diphones for Italian (male voice) and 1800 for German (male voice). Prosodic modules for each language specify a duration and a piecewise linear pitch pattern for every phoneme of an input phonetic string. The pitch and duration of the diphones are modified by the *PSOLA/KDG* algorithm in order to comply with the specified prosodic patterns. The energy mismatches between diphones can be corrected by a simple linear rescaling of the waveforms. The spectral mismatches occurring at the diphone boundaries are corrected by a simple time-domain smoothing technique [5].



(a) wide-band analysis

(b) narrow-band analysis

Figure 2: *The spectrum of the synthetic signal.* The short-time spectrum of the original signal (dashed line) provides the envelope of the synthetic signal (solid line). The synthesized signals are diplayed above the spectral representations. (a) A short window produces a wide-band spectrum, close to the envelope spectrum of the original speech signal (natural vowel). (b) A long analysis window provides an harmonic envelope for the synthetic signal: the modification of the pitch (here by a factor of 1.5) introduces a selective altering of the amplitude of the pitch harmonics (synthetic vowel).

The diphone waveforms are sampled at a relatively high frequency (16 kHz). The size of the dictionaries is large, about 5 Mbytes for French. The dictionaries are also available at any lower sampling frequencies.

The decimation of the diphone waveforms can be performed by use of the *PSOLA/FFT* algorithm. The advantage of this method is the possibility to use an arbitrary decimation factor. The method is reminiscent of a former method proposed in the framework of Fourier analysis and synthesis [14], but it works in a pitch-synchronous fashion. The principle of the method is to resample the spectrum of the ST-signals by the inverse of the decimation factor and to discard the spectrum above the target Nyquist frequency. This resampling of the frequency axis optionally includes a spectral envelope rescaling factor, which results into a systematic shift of all the formants and therefore into a modification of the voice quality. Such preprocessing of the diphone dictionary may be used to enhance to voice quality of the synthetic speech. For instance, we observed that a slight lowering of the formant frequencies by 5% made the French female voice sound closer to the actual speaker that recorded the diphones. Inversely, it turned out that our French male voice was judged somewhat too grave by naive listeners. A slight increase in the formant level was performed to produce a more preferable voice.

The text-to-speech system runs for a 16 kHz sampling rate in 3 times real-time on a 386-based personal computer, and approaches real-time for telephone bandwidth quality. The diphone dictionary is stored in 5 Mbytes of memory extension to ensure a fast access to the diphones.

A formal test was conducted for the French male voice, using 16 naive subjects, comparing the four systems, an *LPC*-based system, and the three *PSOLA* systems (*FFT,MPLPC* and *KDG)*. All three were judged of much better quality than the *LPC*-based system, and they were judged almost equivalent among themselves.

CONCLUSION

We have presented in this paper a new time-domain algorithm for speech synthesis with low computational complexity. The algorithm is based on pitch-synchronous processing of the speech waveform and uses an overlap-add synthesis scheme. It is capable of prosodic modifications of natural speech with a good sound quality. It is particularly efficient in the context of text-to-speech synthesis using diphones, since the pitch can be predetermined in that case and stored with the diphones waveforms. The memory size required to store the speech database is in the order of several Mbytes. However, the algorithm can be advantageously combined with speech compression techniques, such as standard ADPCM, or 64 kb/s PCM, or other speech coding techniques.

REFERENCES

[1] F. Charpentier, M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation", *Proc. Int. Conf. ASSP, 2015-2018, Tokyo, 1986*

[2] S. Roucos, A. Wilgus, "High quality time-scale modification of speech", *Proc. Int. Conf. ASSP, Tampa, 493-496, 1985*

[3] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Trans. ASSP, 32(2), 236-243, 1984*

[4] J.B. Allen, L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", *Proc. IEEE, 65(11), 1558-1564*, 1977

[5] F. Charpentier, E. Moulines, "Text-to-speech algorithms based on FFT synthesis", *Proc. Int. Conf. ASSP, New York, 667-670, 1988*

[6] F. Charpentier, "Traitement de la parole par analyse-synthèse de Fourier: application à la synthèse par diphones", *Doctoral thesis, Ecole Nationale Supérieure des Télécommunications, 1988.*

[7] E. Moulines, F. Charpentier, "Diphone synthesis using *multipulse linear prediction", Proc. FASE Int. Conf., Edinburgh, 1988*

[8] C. Hamon, "Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'ondes", *patent No. 8811517, 1988.*

[9] C. Hamon, "Synthèse de la parole par concaténation de formes d'ondes", *17eme JEP, 1988*, 239-243.

[10] E.P. Neuburg, "Simple pitch-dependent algorithm for high-quality speech rate changing", *J. Acoust. Soc. Am., 63(2), 624-625, 1978.*

[11] K. Lukaszewick, M. Karjalainen, "Microphonemic method of speech synthesis", *Proc. Int. Conf. ASSP, Dallas, 1426-1429, 1987.*

[12] M.R. Portnoff, "Short-time Fourier Analysis of sampled speech", *IEEE Trans. ASSP, 29(3), 364-373, 1981.*

[13] E. Moulines, *Doctoral thesis, in preparation.*

[14] J.B. Allen, "Applications of the short-time Fourier transform to speech processing and spectral analysis", *IEEE Int. Conf. ASSP, Paris, 1012-1015.*