NEW ALGORITHM FOR SPECTRAL SMOOTHING AND ENVELOPE MODIFICATION FOR LP-PSOLA SYNTHESIS

F.M. Giménez de los Galanes, M.H. Savoji[†] *and J.M. Pardo.*

Dpto. Ingeniería Electrónica. U. Politécnica de Madrid.
E.T.S.I.Telecomunicación. Ciudad Universitaria. 28040 Madrid. SPAIN.
†: Dpto. Electrónica. U. de Cantabria. Avda. de los Castros. 39005 Santander. SPAIN.

ABSTRACT

The final quality of a concatenation synthesis system is directly related to the continuity of the spectrum at the concatenation point. Due to the subjective auditory masking, if we minimize the spectral distortion in the formant frequencies, the quality will increase significantly. In this paper we present, along with results concerning pitch marking, an algorithm capable of modifying the LPC envelope in a flexible way which is the heart of a spectral smoothing module for a diphone-based Linear Prediction Pitch-Synchronous Overlap-Add (LP-PSOLA) concatenation system.

1. INTRODUCTION

A key factor in the final quality of a text-to-speech system based on unit concatenation is the continuity, or smoothness, of the formant trajectories at the concatenation point. In these points we are matching up two different spectra from two units which were obtained from different phonetic context. Even when we use diphone type units (polyphones), which assumes that we are concatenating units using stable segments, every half-allophone will be different somehow (there is a measurable distance between the two spectra on both sides of the concatenation point).

The PSOLA algorithms, specifically Time Domain PSOLA (TD-PSOLA) and Frequency Domain PSOLA (FD-PSOLA), cause an inherent smoothing in the period that is being synthesized from two different units due to the overlap-add process [1]. This smoothing is not enough to avoid "spectral jumps"; formant trajectories have to make the transition in a single period, and that produces an audible distortion of the speech signal.

One possible solution is to minimize the inter-unit distance at the data base collection stage, but considering the number of units and all of the possible combinations, that would be an impossible task. This makes the development of a smoothing algorithm a high priority.

2. THE PSOLA MODIFICATION FRAMEWORK

Under the name of PSOLA techniques, we can find three different approaches, FD-PSOLA, TD-PSOLA and LP-PSOLA [2]. Our work, presented here, can be applied directly to an LP-PSOLA synthesizer, since we already have a modelling of the spectral envelope that is given by the LPC coefficients.

In trying to combine the high performance of the TD-PSOLA modification scheme with the comparatively reduced memory requirements for the database of LPC-based coders, a different scheme was presented in [2]. The idea is to codify the data base using a high quality coder (CELP, MLPC,...). In synthesis time, the unit has to be decoded, and then, the TD-PSOLA step can be performed. The LP-PSOLA method consists of modifying the prosodical characteristics of the Linear Prediction residual by the TD-PSOLA process and then adding the spectral information by inverse filtering. The TD-PSOLA modifications are simple, pitch synchronous, "repetition-elimination" techniques over specially windowed portions of speech. The advantage of using the TD-PSOLA modifications over the residual instead of the signal itself is that the spectral distortions in the formant frequencies are lower (TD-PSOLA over the signal has a bandwidth broadening effect).

This last operation is not linear, so theoretically, the two processes can not be interchanged, but, in practice, the spectrum of the residual is mostly white and continues to be white after the PSOLA modification. The LPC filtering works on a signal with spectral characteristics similar to the original.

The LP-PSOLA synthesis works identically for any LPC-based coding scheme used. A possible alternative is not to codify the residual at all and use this prediction error signal directly as the input to the PSOLA module. This will achieve the highest possible voice quality. Usually, MLPC or CELP schemes are used since they are perfectly suited to this kind of synthesis.

When using a MLPC coder, the process operates by building a "residual" signal by concatenation of the different multipulse frames. For stochastic (CELP) coders, this "residual" is the result of pitch (long term) filter the frames extracted from the codebook.

LPC frames are selected pitch-synchronously, as shown

0-7803-1775-0/94 $3.00 © 1994 IEEE

in figure 1. The operative length of every coefficient set a_k is half the length of a synthesis short-term signal. For every synthesized speech period, its first half is filtered with the coefficients of the first short-term signal, and the second half, with the correspondent to the second overlapping signal.

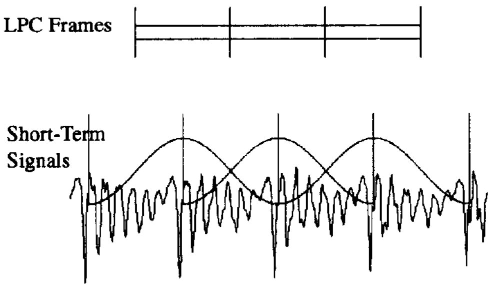

Fig. 1. LPC frames are selected synchronously and centered in the pitch marks used for Short-Term signals windowing.

3. PITCH MARKING

It is said that the exact point of marking the periodicity of the signal has an important effect on the final quality of synthesis [1]. We have made two different systems to generate these marks of synchronism: one, based on the Hilbert envelope, which marks the maximum of the glottal function and the other one, which, with a method derived from the MLPC excitation computation, gives us the sample of causal excitation of an LPC synthesis filter that resembles the original signal [3]. A comparison of the two methods is shown in figure 2.

These two methods were used to pitch-mark a complete database of concatenation units. The results were clear: no noticeable difference was encountered listening to synthesized speech using these two systems [4]. Because the first one is less expensive in terms of computing, we used that database in the following experiments.

4. THE SPECTRAL SMOOTHING ALGORITHM

Since we have an LPC envelope, the smoothing process will be the obtaining of a new set of coefficient a_k. These new coefficients will produce an interpolated spectra from both sides of the concatenation period.

A classic technique is LAR (Log Area Ratios) smoothing. This technique gives good results with low computation cost, but it is a "blind" process; there is no possible control of the formant frequencies and bandwidths.

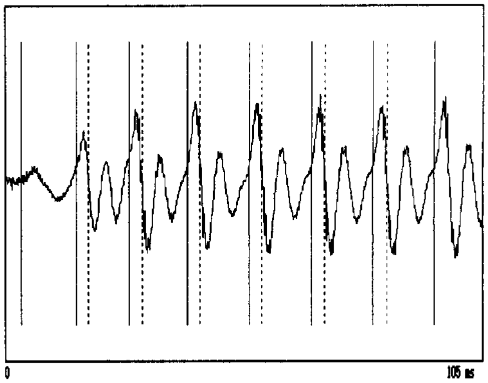

Fig. 2. Performance of the marking system. MLPC readjustment (solid line) is always ahead of the Hilbert marking (dotted line). The first and last marks were generated in a post-processing step.

The algorithm presented here provides us with not only a good smoothing performance [5], but also with a powerful tool for spectral modification, specially suited to joint formant-fundamental frequency modifications (once we have the proper warping function [2,6]).

The method is carried out in three steps: detection of the formant frequencies and amplitudes (we detect amplitudes instead of bandwidths because they are equivalent under controlled modifications), smoothing of these trajectories for at least the first five formants, and synthesis of a new set of LPC parameters a_k.

The first step is implemented using the third derivative of the LPC phase spectrum [7,8]. We detect the first five formants because they are the most important to be smoothed due to their higher energy. This step involves the computation of a FFT (of, for example, 512 points), the module and phase of the resulting spectrum, and the third derivative of the phase. The accuracy of the detection depends on the number of LPC coefficients used. For 16 kHz of sampling frequency, 17 coefficients were employed.

The second step is a linear smoothing of the frequencies and amplitudes. The exact number of reference points to interpolate is 5 points corresponding to formant peaks plus 6 "local minima" points, which are the points positioned in the center of the band determined by two formants or by the first formant and the 0 frequency, or by the last formant and π frequency. They are a good estimate of the valleys in the LPC spectrum.

The final step is a little more complex. We need to warp or modify the spectral envelope and go back to the set of

LPC coefficients associated with this new spectrum. The warping algorithm is derived from the TD-PSOLA modification schemes. We can construct a set of "spectral short-time signals" from the spectral axis of analysis (where we have "formant marks" instead of "pitch-marks") by windowing this amplitude spectrum with windows centered at each "formant mark". These "spectral short-time signals" are moved following the new set of formant frequencies in the synthesis spectral axis. The new envelope is the result of overlap-adding these signals. Once the module of the new LPC envelope is obtained, we compute the inverse FFT of the squared module, which leads to the autocorrelation coefficients of the impulse response of the minimum phase filter corresponding to this envelope [9]. From the autocorrelation coefficients we get the a_k using Durbin recursion.

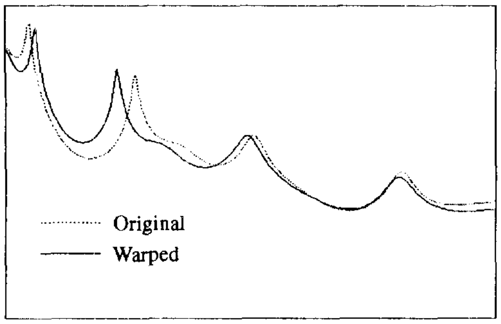

Fig. 3. Spectral envelope of a LPC frame from the vowel [e]. Original spectrum (dotted line); the spectrum warped by the proposed algorithm (solid line).

5. ENERGY EQUALIZATION

In general, the new set of coefficients a_k, and the old LPC filter have different gain. That produces a distortion which is audible and can be seen in a waveform plot. In order to have continuous energy contours, an energy equalization (or gain compensation) is needed.

One way to obtain the compensation factor G is to measure the energy of both filters and apply the formula:

$$G = \frac{E_{old}}{E_{new}}$$

An estimate of the energy (or gain) of a filter is the sum of the squared module of the FFT points representing its spectrum. By Parseval's Relation, this is the autocorrelation function in delay 0:

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi} \int_{2\pi} |X(\Omega)|^2 d\Omega = r_x(0)$$

However, since we know that the function $|X(\Omega)|$ is the inverse of the module of the FFT of a_k, the estimate can be written as follows:

$$\tilde{r}_x(0) = \frac{1}{\sum_{k=0}^{p} a_k^2}$$

and, the final expresion for the gain is:

$$G = \frac{\sum_{k=0}^{p} a(new)_k^2}{\sum_{j=0}^{p} a(old)_j^2}$$

6. RESULTS

This algorithm is capable of smoothing very different formant structures, if they are clear and noise free as they are in the database of concatenation units for a text-to-speech system, as is shown in figure 4. An important parameter to optimize is the length of the smoothing segment. Charpentier [1] recommends 30 mseg segments, which is equivalent to three periods for a standard male voice at 100 Hz. In our system, due to the different lenght of the LPC frames, we have to smooth over 2 to 4 sets of coefficients a_k.

The formant trajectories will be smoother with a longer smoothing segment, but, because of the variable length (in number of frames) of each synthesis unit, it is better to have an adaptive length, depending on the speech production rate.

Acoustically, the waveform obtained has a more continuous sound because the transition noise is minimized.

7. CONCLUSIONS

We propose this spectral modification method as a valid smoothing module for future text-to-speech conversion systems. This method can also be used as the basis for joint source-filter modifications for future research.

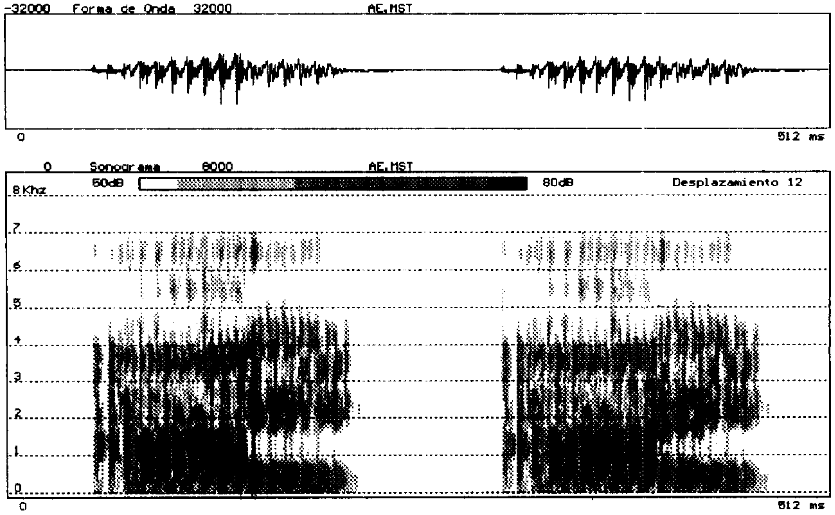

Fig. 4. Waveform and spectrogram of a synthetic speech segment showing the union of two diphones [_a] and [e_]: without smoothing, left; with spectral smoothing over a four-frame window, right.

It can also be a very useful tool for preprocessing databases of synthesis units. There are two possibilities: to create an artificial prototype of an allophone and smooth all the units starting or ending with that allophone to the computed prototype, which may create an unnatural sound, or, to increase the number of units artificially by creating different versions of each polyphone smoothed to the next unit it will be concatenated to, but this, of course, would increase the memory requirements for the synthesizer.

REFERENCES

[1] F. Charpentier: *Traitment de la parole par analyse-synthèse de Fourier: Application à la synthèse par diphones*, Doctoral Thesis, Ecole Nationale Supérieure des Télécommunications, 1988.

[2] Eric Moulines and Francis Charpentier: "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, Vol. 9, pp. 453-467, 1990.

[3] M.H. Savoji: "Automatic readjustment of the estimated excitation points for speech synthesis by waveform concatenation". BTRL Technical Report, June 1990.

[4] F. Giménez de los Galanes, M.H. Savoji and J.M. Pardo. "Marcador automático de excitación glotal". Proc. of URSI'93. Sept. 22-24. Valencia, 1993.

[5] F. Giménez de los Galanes, M.H. Savoji and J.M. Pardo. "Suavizado inter-unidad por modificación espectral para síntesis por concatenación LP-PSOLA". Proc. of URSI'93. Sept. 22-24. Valencia, 1993.

[6] H. Valbret, E. Moulines and J.P. Tubach. "Voice transformation using PSOLA technique", Speech Communication Vol. 11, pp. 175-187, 1992.

[7] B. Yegnanarayana: "Formant extraction from linear-prediction phase spectra", Journal of the Acoustical Society of America, Vol. 63, no. 5, pp. 1638-1640, Mayo 1978.

[8] N. Sridhar Reddy and N.M.S. Swamy: "High-Resolution Formant Extraction from Linear-Prediction Phase Spectra", IEEE Trans. Acous. Speech and SIgnal Proc., Vol. ASSP-32, no. 6, pp. 1136-1144, December 1984.

[9] Vladimir Goncharoff and Suresh Chandran: "Adaptive speech modification by spectral warping", IEEE Proc. of Intern. Conf. on ASSP, pp. 343-346, 1988.