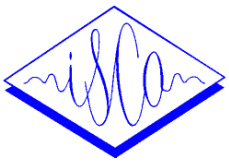


4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

IMPROVED TRANSIENT REPRESENTATION AND QUANTIZATION FOR SINUSOIDAL SPEECH CODERS

M.S. Torres-Guijarro*, F.J. Casajús-Quirós
e-mail: marisol@gaps.ssr.upm.es
ETSI Telecomunicación-UPM
Ciudad Universitaria
28040 Madrid, SPAIN
* ETSI Telecomunicación-Universidad de Valladolid

ABSTRACT

In this contribution we propose an improvement of the multiband excitation vocoder on the basis of preserving temporal and spectral features of transients. Proper representation of unvoiced-voiced transitions is achieved by means of reliable detection of the point where harmonics are born and application of specific analysis and synthesis procedures for the frames placed before and after this point. Description of temporal features requires the definition of new parameters, their extraction and quantization having been studied. Additional methods for parameter reduction and quantization will also be addressed.

1. INTRODUCTION

Proper representation of transitions is an important drawback of speech coders that, as the sinusoidal coder, try to closely represent frequency features of voice signals. The characterization of such frequency features is usually performed on a basis of short-term transformation, which gives no information of the temporal evolution of the speech signal. This information is commonly substituted by linear interpolation between frames of the spectral parameters, such as harmonic amplitudes.

This linear approximation introduces small distortion in stationary regions, but is clearly inappropriate in transients. Resulting effects are specially serious during the perceptually crucial unvoiced to voiced transitions, due to two major causes. Firstly, spectral characteristics before and after transition change dramatically. A traditional sliding window analysis with constant shift will place at least one window half on the unvoiced region, half on the voiced one. Therefore the estimated parameters will mix both informations, and the corresponding synthetic segment will neither have proper unvoiced features, nor voiced ones.

Secondly, it seems important to preserve not only spectral information, but the temporal envelope of the original signal also. This applies to the growing envelope of the voiced onset and to the previous noise signal, specially in certain transitions such as plosives, otherwise lost. As a solution to these problems, we propose an MBE based scheme with specific analysis and synthesis procedures for the case of unvoiced-voiced transition.

To begin with, onset frames need to be detected. As detailed in section 2, this detection is carried out on a basis of a phonetic classification similar to the proposed in [1], taking into account voicing and energy information of each frame. Once the onset frame has been pointed out, an analysis of the segmental energy leads to the determination of the *transition point*, there where the harmonics are born.

Work supported by National Project TIC92-0329

At this moment, some flexibility can be given to the analysis window positioning: a first analysis window is located ending at the transition point. Spectral characteristics and energy profile from the noisy zone previous to transition will be extracted from it. A second analysis window, starting at the transition point, will give reliable information about energy and phases of emergent harmonics. Details of the analysis procedure will be given in section 3.

Associated to the transition analysis procedure is an specific synthesis algorithm (see section 4) which makes use of the specific parameters extracted before. In addition, a nonlinear profile is applied to the amplitudes of the harmonics in order to approximate their natural evolution.

With the proposed method, a faithful synthesis of onsets can be obtained, as shown in figure 1. It represents the syllable "ta" from the Spanish word "respuesta" (response) uttered by a woman, synthesized with the IMBE [2] coder and with the proposed algorithm. It can be noticed the improvement of the temporal characteristics of the synthetic speech due to three reasons: first, the energy profile applied before the transition reconstructs the plosive pulse; second, nonlinear interpolation of spectral amplitudes preserves the temporal envelope of the onset; and third, the original initial phases of the emerging harmonics lead to a more reliable reconstruction of the pitch pulse. As a consequence, paired comparison with standard IMBE coder shows a clear advantage of the proposed scheme.

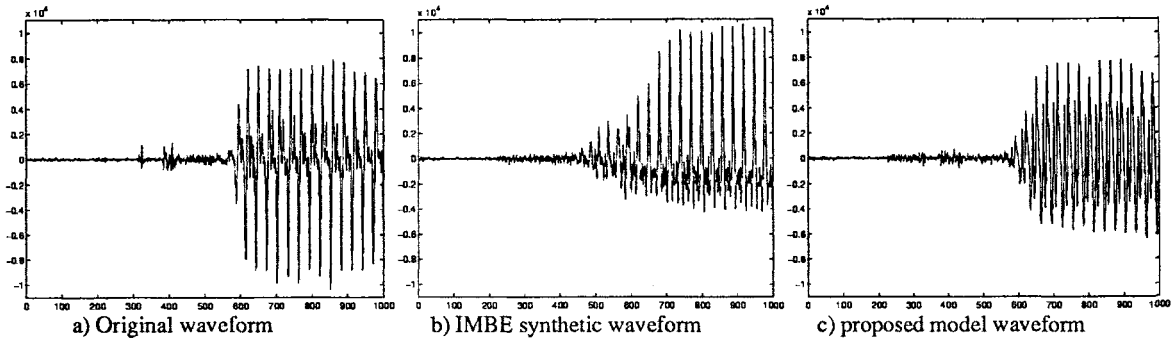
The way we handle transitions calls for additional information to be transmitted to the decoder: place of birth of harmonics, energy profile of the frame previous to transition and initial phases of emerging harmonics. The way this parameters can be coded without increasing the global bit rate of the coder is addressed in section 5. Additional methods for parameter reduction have also been studied and summarized in section 6.

2. ONSET DETECTION

The first step in the analysis of unvoiced-voiced transitions is their detection. This is accomplished in two stages: detection of the window containing the onset in a traditional constant-shift sliding window analysis; and determination of the precise point where harmonics are born inside that window, namely the *onset point*.

2.1 Detection of transition window

An unvoiced-voiced transition is usually characterized by a significant increment of energy and sonority, therefore the detection of a window containing a voiced onset should involve measures of these parameters over itself and perhaps neighboring windows. Several possible measures of energy and sonority have been tested, the next having been selected for their robustness in the detection:



a) Original waveform b) IMBE synthetic waveform c) proposed model waveform

Figure 1: Syllable "ta".

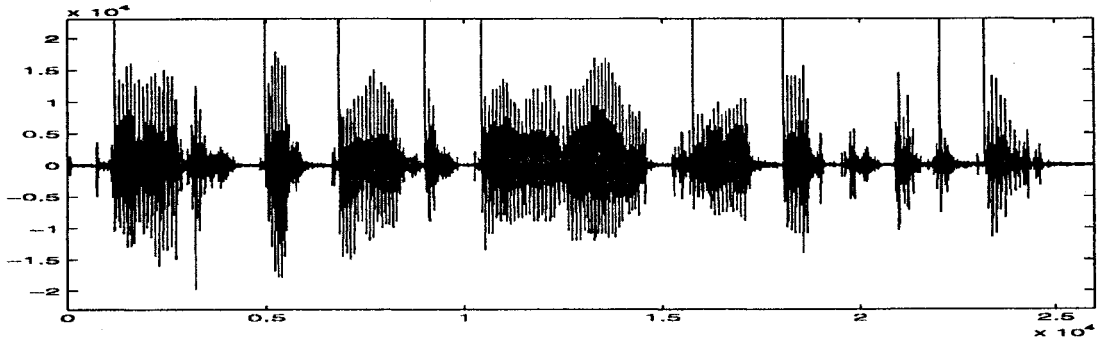


Figure 2: Example of onset detection

Sonority measure: the gross-error function of pitch period [6] given by:

$$E = \frac{\sum_{n=-\infty}^{\infty} w^2(n)s^2(n) - \Psi(P)}{\left(1 - P\sum_{n=-\infty}^{\infty} w^4(n)\right)\int_{-\pi}^{\pi} |S_w(\omega)|^2 \, d\omega}$$

where $w(n)$ is the analysis window, $s(n)$ is the speech signal, P is the integer pitch period, $S_w(\omega)$ is the spectrum of the windowed signal and

$$\Psi(P) = P\sum_{k=-\infty}^{\infty} \phi(kP)$$

with $\phi(m)$ the autocorrelation function of $w^2(n)s(n)$.

Energy measure: the segmental energy of the windowed signal, calculated over 8 non-overlapping segments.

A window is decided to contain an onset if both the following conditions are verified:

1.- the increment of the sonority measure between the preceding and following windows exceeds certain empirical threshold

2.- the maximum relative increment of energy between consecutive segments of the window exceeds a second empirical threshold.

The use of the gross-error function as a sonority measure has the following advantage: as it is already computed by the pitch refinement algorithm for two windows in advance, the first rule does not mean an increment of the processing delay.

2.2 Determination of the onset point

Finally, the birth of the harmonics is located in the frontier between the segments with maximum relative increment of energy. This gives a precision in the measure of 1/8 of the window length, which has proved to be enough for our purposes.

The performance of this method does not depend on the speaker and exceeds the 80% of detection, while few false detections have been found in the processed speech. Figure 2 shows an example of performance in a sentence uttered by a male speaker. Vertical lines indicate detected onsets.

3. TRANSIENT ANALYSIS

Once located the onset point, unvoiced and voiced segments can be analyzed separately by means of repositioning the analysis windows. The window finishing closer to the onset point, namely *pre-onset window*, is moved in order to finish exactly there. Hence, the MBE analysis applied to it will provide pure unvoiced spectral information. It can also be used to extract the energy envelope of the noise previous to the onset, where a plosive pulse could be found.

Similarly, the window starting closer to the onset, namely *post-onset window*, is moved to start at this point. Its analysis gives reliable information about the pitch period, the amplitude of the newborn harmonics and their sonority. If their initial phases are also extracted, they allow a closer reconstruction of the pitch pulse. Figure 3 shows an example of the new placement of windows. As it can be seen, the window lying between the *pre-onset window* and the *post-onset window* is discarded, as it produces distorted information mixing voiced and unvoiced characteristics.

4. TRANSIENT SYNTHESIS

The synthesis of the unvoiced component of the speech is essentially identical to the basic IMBE vocoder. Despite the displacement on the analysis windows around the onset point, the overlap-add synthesis procedure is applied as if the analysis window were moved with constant shift, thus the amount of overlapping can be kept constant. This simplification does not introduce significant distortion.

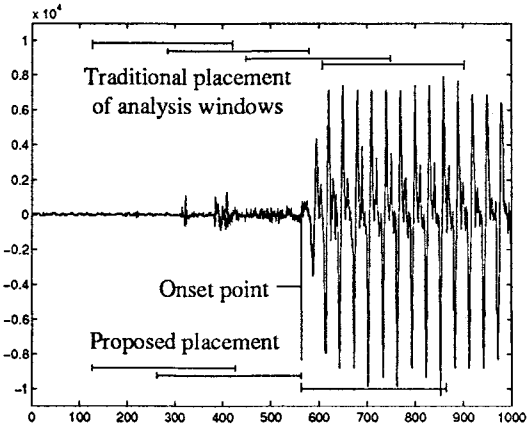


Figure 3: Proposed placement of analysis windows

The spectral information of the window lying between the *pre-onset window* and the *post-onset window* is substituted by linear interpolation of the other two, as it can introduce pre-echoes.

In relation to the voiced synthesis, the following modifications have been introduced to the basic IMBE algorithm of [2]: all the harmonics declared voiced in the *pre-onset window* die before the onset point; and all the harmonics declared voiced in the *post-onset window* are born in the onset point. The latest start with original phase information instead of the usual linear component, in order to better reproduce the pitch pulse of the voiced segment. The temporal envelope of the onset is better reconstructed if the growing amplitudes follow a raised-cosine law, instead of linear.

Finally, the pre-onset energy profile is applied to the sum of the voiced and unvoiced components of the synthetic speech.

5. ADDITIONAL PARAMETERS CODING

Proper representation of onsets requires additional information to be transmitted to the decoder: onset point, energy profile of the *pre-onset window* and initial phases of emerging harmonics. This does not mean a raise in the bit rate, as this information can be placed instead of the information corresponding to the window which was discarded. 48 bits are thus available with a transmission rate of 2.4 Kbps and a frame rate of 20 ms, and can be located as shown below

Phases	9
Onset point	3
Energy profile	36
Total	48

Three bits are enough to code the onset point, computed from the energy of the window divided in eight segments. The energy profile previous to the onset point can be vector quantized with far less than 36 bits.

Concerning the phase information, and according with our results, only three original onset phases need to be sent, those corresponding to the largest voiced harmonics, and can be coded with three bits each. The remaining phase spectrum is recoverable from the amplitude spectrum on the basis of a minimum-phase model, as detailed below. The basic idea for reducing the phase information is to approximate the phonetic system to a minimum-phase system, followed by a temporal delay. This assumption allows the substitution of most part of the original phase information for the minimum-phase spectra associated to the system without great degradation of the resulting waveform.

The minimum phase of the system is computed from samples of the amplitude response through the discrete cepstrum coefficients, defined as [7]

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} \log A(k)\cdot e^{j2\pi kn/N} \quad ,0 \le n \le N-1$$

This expression corresponds to the inverse Fourier transform of the logarithmic spectral envelope of the speech signal, where $A(k)$ represents the samples of the envelope and N the length of the FFT. The values $A(k)$ are computed through linear interpolation of the amplitudes of the harmonics.

From the discrete cepstrum coefficients, samples of the spectral phase are obtained as

$$\phi_{\min}(k) = -2\sum_{n=1}^{N/2-1} c_n sen(2\pi kn/N)$$

which linearly interpolated lead to the minimum phases of the harmonics.

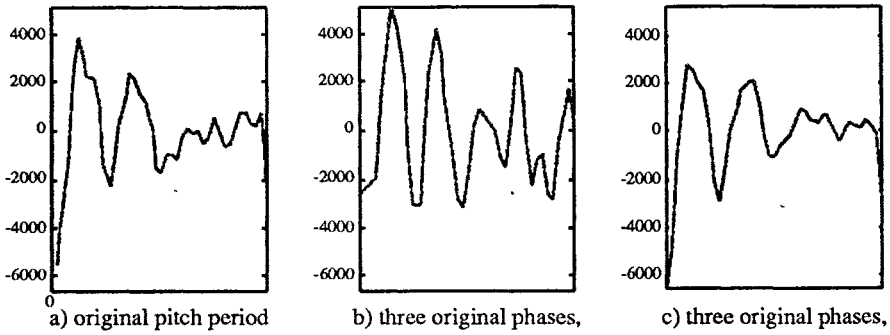
As the original phases of the voiced harmonics are replaced with the respective minimum phases, the voice signal waveform gradually degenerates, but never as much as if random phase information were used instead of minimum phase. Well known, the coexistence of original and minimum phases forces us to articulate some kind of temporal correlation between harmonics, so that new phases suffer the same initial delay as original ones. The criteria followed to choose the phases to be substituted is to preserve those of the most audible harmonics, that is to say, the larger ones. Figure 4 shows a female pitch period, reconstructed with three original phases and random or minimum remaining ones.

During the non transient operation, no original phase information is used. Initial phases of harmonics are computed as in [2], with a linear component that preserves continuity of the phase function over frame frontiers and a second component that provides the necessary randomness of the phases of the lower harmonics. When an onset is detected, all this phase model is substituted by the three original phases plus the minimum phase spectrum. It should be noted that linear continuity is no longer needed as there is no continuity of harmonics over the onset point.

6. PAREMETER REDUCTION

In order to reach a bit rate which is low enough, the parameters resulting from the analysis must be quantized. In this process our main concern is the quantization of the spectral envelope as represented by the amplitudes A_m. Central to this problem is the dimensionality normalization process, motivated by the fact that the number of samples of the spectral envelope (A_m) depends on the pitch period in the frame under consideration. Two approaches are commonly used. One of them [3] uses interpolation in order to estimate the spectral envelope on a normalized spectral grid (40 to 50 grid points are typical). In this way a high-dimension vector - of fixed length - is obtained and must be quantized. The procedure leads to a vector quantization problem of high dimensionality. Another possibility, which copes with the high-dimensionality problem, is to model the spectral envelope by means of a small, finite and fixed number of parameters, quantize them and recalculate the A_m from the quantized parameters.

Following this idea, in [4] the autocorrelation of the frame is calculated from the A_m, it is then used in order to obtain the LPC parameters of the frame, which model the spectral envelope.



a) original pitch period b) three original phases, rest random c) three original phases, rest minimum

Figure 4: Reconstructon of the pitch pulse with different phase information

The performance of this method is poor for high-pitched voices, since the harmonics are few and the sampling of the envelope is inadequate.

The quantization scheme we propose tries to overcome both problems [5]. It intends to obtain a low-dimensionality representation of the spectral envelope, as independent of pitch period as possible. It uses also perceptual properties of hearing in order to improve results. Therefore consider first the approximation not of the A_m but of their logarithm by an all-pole discrete power spectrum:

$$\log A_m \approx \log\left(1\Big/\sum_{k=0}^{p} a_k \cos k\omega_m\right) \qquad (6.1)$$

Where p is the order of the model, and ω_m the normalized frequency of the m-th harmonic. Since A_m $(m=1,...,N)$ is a finite duration sequence, the approximation can be exact for $p=N$. Now p is the order of the model and must be fixed, and N is the number of harmonics, so it changes from frame to frame. This implies that for high pitched voices (N low) the approximation can be very good. In the general case we can calculate the a_k that minimize the following error criterion:

$$E=\sum_{m=1}^{N}\left[\log\left(1\Big/\sum_{k=0}^{p} a_k \cos k\omega_m\right)-\log A_m\right]^2 =$$

$$=\sum_{m=1}^{N}\left[-\log A_m \sum_{k=0}^{p} a_k \cos k\omega_m\right]^2 \approx \sum_{m=1}^{N}\left[A_m \sum_{k=0}^{p} a_k \cos k\omega_m - 1\right]^2$$

which means that:

$$\frac{\partial E}{\partial a_n}=2\sum_{m=1}^{N}\left(A_m \sum_{k=0}^{p} a_k \cos k\omega_m - 1\right)A_m \cos n\omega_m = 0, \quad n=0,...,p$$

that is to say:

$$\sum_{k=0}^{p}\left(\sum_{m=1}^{N} A_m^2 \cos n\omega_m \cos k\omega_m\right)a_k = \sum_{m=1}^{N} A_m \cos n\omega_m, \quad n=0,...,p$$

These last equations can be put into matrix form as $\mathbf{C.a}=\mathbf{b}$, where

$C_{nk}=\sum_{m=1}^{N} A_m^2 \cos n\omega_m \cos k\omega_m$; a_n is the n-th parameter, and

$b_n=\sum_{m=1}^{N} A_m \cos n\omega_m$

Therefore, since $\mathbf{a}=\mathbf{C}^{-1}b$, matrix inversion of $\mathbf{C}$ allows us to calculate the parameters of the model. Expanding the cosine products of $\mathbf{C}$ in sum of cosines, $\mathbf{C}$ can be recognized as the sum of a Töplitz matrix and a Hänkel one. Hence $\mathbf{C}^{-1}$ can be efficiently calculated with complexity proportional to p^2.

After quantization and transmission of the a_k, the A_m can be estimated by use of expression 6.1. We can take of perceptual factors by transforming the frequency axis into the place axis according to:

$$x_m=\pi \sinh^{-1}\left(\frac{f_s \omega_m}{1300\pi}\right)\Big/\sinh^{-1}\left(\frac{f_s}{1300}\right)$$

We then use x_m instead of ω_m in all the preceding expressions. In this way the estimated values of A_m are more precise at the low frequency range, where human ear has better resolution.

As a preliminary evaluation, several segments of speech were analyzed and synthesized with no quantization by means of: the original A_m, the envelope resulting of LPC modeling and the spectral envelope as obtained through the all-pole method we have explained. For low-pitched voices both methods show some degradation when compared with direct use of A_m, although the degradation is somewhat smaller for the all-pole method. For high pitched voices and LPC the degradation clearly increases whereas the all-pole method noticeably improves, producing results perceptually identical to those yielded by the unmanipulated A_m. The all-pole model is sometimes ill-conditioned when N is large (low pitch). Perceptual weighting alleviates the problem, but it does not disappear. In practice we use LPC for low pitch and all-pole models for medium to high pitch.

References:

[1] A. Das, A. Gersho, "Enhanced multiband excitation coding of speech at 2.4 kb/s with phonetic classification and variable dimension VQ", Proc. of EUSIPCO 94, pp 943-946.

[2] Digital Voice Systems, "Inmarsat-M Voice Codec, Version 2", *Inmarsat-M specification*, Inmarsat, London, February 1991.

[3] M. Nishiguchi et al., "Vector Quantized MBE with Simplified V/UV Division at 3.0 kbps", Proc. of ICASSP 93, pp II.151-II.154.

[4] D. Rowe and P. Secker, "A robust 2400 bit/s MBE-LPC speech coder incorporating joint source and channel coding", Proc. of ICASSP 92, pp II.141-II.144.

[5] Torres-Guijarro, M.S. and Casajús-Quirós, F.J., "Improved Analysis/Synthesis Methods for the Multiband Excitation Coder", Proc. of MELECON 94., pp 57-60

[6] D.W. Griffin and J.S. Lim, "Multiband Excitation Coder", IEEE Trans. on ASSP, vol. 36, no. 8, pp 1223-1235, August 1988.

[7] R. J. McAulay and T. F. Quatieri, "Low-Rate Speech Coding Based on the Sinusoidal Model", in "Advances in Speech Signal Processing". ed. Marcel Dekker, 1992.