A NEW PARAMETRIC SPEECH ANALYSIS AND SYNTHESIS TECHNIQUE IN THE FREQUENCY DOMAIN

BR. WÉRY, A. LEROUX, H.PH. DELBROUCK, J. LECLERC
University of Liége - Department of Acoustics - Belgium

This work has been supported by belgian governmental funds under contract 85/90-81.

A new method for speech parametric coding at medium output rate is presented. The availability of new electronic components changes the hypotheses assumed in previous studies. A parametric coding scheme with medium output rate now seems appropriate for rule-based synthesis. The weak spot in most parametric synthesizers lies in the poor model used to represent the excitation signal. Our approach consists in coding, in the frequency domain, the residual signal of a linear prediction analysis. The residual signal is analyzed as a function of frequency through a sub-band splitting. Particularly, voicing is analyzed in each band and the decision is progressive from voiced to unvoiced. This parametric, frame based, representation of the residual signal, associated with the linear prediction coefficients, is used for transmission or storage purposes. The synthesizer builds the synthetic residue by generating spectra from these parameters. These spectra are combined through the overlapp-add algorithm to obtain a time domain signal. Our system thus allows complex filtering and phase control and provides a robust representation and a good quality while preserving the flexibility inherent to parametric coding schemes.

1. INTRODUCTION.

Works on fully parametric analysis-synthesis systems are rare since the apparition of the LPC vocoder. The simplicity of the modelling of the residual signal is certainly the major drawback of this classic scheme. Some systems present improvements in the representation of this residual signal but are generally related to waveform coding. They are well suited for transmission purposes but unsuited in text-to-speech synthesis.

Separation of residual signal into voiced and unvoiced frames is a quite difficult task. The harmonic structure of some voiced sounds is not regular, fading in some frequency regions or appearing only at very low frequencies. Thus, the voiced-unvoiced decision mechanism of the LPC vocoder is certainly meaningless in some transition areas. In our approach, we replace this decision by a "voicing function", progressive from fully voiced to fully unvoiced, depending on frequency. This is made through a sub-band decomposition where each band is characterized by a "voicing ratio". Such an approach has also been adopted by D.Griffin and J.S. Lim in [10], [11], [12] and [15].

Creating a fully parametric system, able to generate synthetic speech from parameter vectors transmitted on a frame by frame basis is our objective. Problems appearing at frame connections are resolved in time domain through the use of adaptive filters but the behaviour of such filters is difficult to analyze. Processing in the frequency domain, through the "overlapp-add" algorithm is an elegant way for solving these problems [1].

When no signal processor was available, frequency domain methods were unaffordable because of the important computing overhead it produces. Now, this is changing with the apparition of processors able to process signals in real time through large transforms. Thus, techniques like the one described in this paper will certainly run in real time on today equipments.

To be able to design efficient strategies for connecting phonetic elements in text-to-speech applications, a coding system producing parameters with a clear, easy to understand meaning is needed. This is obtained in our scheme.

2. GENERAL DESCRIPTION

The general flow charts of the analysis and synthesis systems are presented in figure 1. In the following, romanic numbers refer to this figure.

A) ANALYSIS

(i) The speech signal (sampled at 10000Hz, filtered at 4000Hz) is first filtered through a classic linear predictor (SCHUR-LEROUX algorithm). The PARCORS coefficients are transmitted.

(ii) The residual signal is obtained and processed frame by frame. An FFT (1024 samples wide) is applied on each frame (weighted by a 512 samples Hanning window). The spectral representation of the residual signal is thus obtained.

(iii) The cepstrum method is used for pitch detection. This is performed through an algorithm derived from the one described by W. Hess [3, chapter 8, pp 399 - 409].

(iv) The spectrum is then splitted in sub-bands.

(V) Each band is analyzed to obtain an energy value and an estimation of the voicing ratio.

(VI) Parameters (Parcors, pitch, voicing ratio and energy in each band) are coded and transmitted.

B) SYNTHESIS.

First, spectra of fully voiced and fully unvoiced frames must be computed.

(VII) The generation of the unvoiced spectra is quite easy. We use a catalog of spectra of random sequences weighted by Hamming windows. Some care is necessary to ensure a good behaviour of the system when using such spectra through the overlapp-add algorithm.

(VIII) The generation of the voiced spectra is more difficult because of the problem of connection between overlapping frames. Each spectrum must be computed from the pitch period and previous frames and then be multiplied by the spectra of a typical glottic pulse to improve the time domain waveform.

(IX et X) These two kinds of spectra are weighted and summed in each band according to the energy and voicing parameters. The synthetic spectrum of this frame is thus obtained.

(XI) The time domain residual is built by the "overlapp-add" algorithm.

(XII) The synthesized speech is obtained from the synthetic residual signal by LPC inverse filtering.
3. DETAILLED REVIEW OF SOME MECHANISMS.

3.1. SUB-BAND SPLITTING AND VOICING ESTIMATION

The voicing function must be defined as a function of frequency. The simplest way to do this is to split the frequency domain into sub-bands and to estimate an "average voicing ratio" in each band. A more complex solution is to estimate this voicing function for each signal harmonic [12]. This second procedure is pitch dependant and produces a high and variable bit rate.

We use 8 fixed sub-bands. The choice of these bands is dictated by perceptual tests, perception equivalence between bands [6], [7], [14] and stability conditions of the voicing estimator. The bands are 100-400, 400-700, 700-1000, 1000-1300, 1300-1650, 1650-2000, 2000-3000 and 3000-4000 Hz.

The voicing estimation is a new problem. Our method is based on the properties of the autocorrelation function [16]. We suppose the input signal (residual signal) to be the sum of a periodic sound at pitch frequency m and of an uncorrelated (at order m) noise signal. Then the autocorrelation value R(m) is an estimation of the energy of the periodic (voiced) part of the input signal. This assertion is also valid for a band filtered signal if the impulse response of the filter is finite and shorter than M samples.

Knowing that the correlation coefficient R(n) for the current frame can be computed in the frequency domain as (N is the size of the FFT)

\[ R(n) = \sum_{k=0}^{N-1} |S(\omega)|^2 \times \cos\left(\frac{2\pi kn}{N}\right) \]

the ratio ("voiced energy"/energy) \( r_p \) in a band p (with the same assumptions as above) will be given as the ratio of the contributions of the band p to R(m) and R(0)

\[ r_p = \frac{\sum_{k=bs}^{be} |S(\omega)|^2 \times \cos\left(\frac{2\pi kn}{N}\right)}{\sum_{k=bs}^{be} |S(\omega)|^2} \]

where bs and be are the first and last sample indexes of band p.

3.2. QUANTIFICATION

A maximum bound for the data rate of the compressed speech can be computed as follows:

- pitch: 7 bits per frame;
- global energy: 6 bits per frame;
- band energy: 4 bits per band each frame;
- voicing information: 4 bits per band each frame;
- LPC parameter: 59 bits per 2 frames.

This leads to a total of 96 bits per 128 samples (frames) or 512 samples long with 75% overlapping (analysis) or 1024 samples long with 75% overlapping (synthesis). The global data rate is then 7900 bits/s.

3.3. SPECTRUM GENERATION

A synthetic spectrum must be generated for each frame from transmitted parameters. These spectra will be used in the overlap-add algorithm to generate the excitation signal. This process includes three steps:

- generation of a full voiced spectrum;
- generation of a full unvoiced spectrum;
- weighted addition of these spectra.

This paragraph deals with the first two steps.

3.3.1. GENERATION OF THE UNVOICED SPECTRA (FIGURE 2)

The main problem is to generate a sequence of spectra to be used in the "overlap-add" algorithm without producing destructive interferences. Thus spectra must be transforms of 75% overlapped random sequences.

Any dynamic generation of these spectra leads to important overhead and can be efficiently replaced by a catalog mechanism. This catalog is generated through the analysis of a random signal with a 75% overlapping with a weighting Hanning window.

During synthesis, frames are chosen successively. The catalog must be sized so that no cycle appears in a speech element.

3.3.2. GENERATION OF THE VOICED SPECTRUM (FIGURE 3)

This problem is more complex. Interactions between successive spectra are more sensitive and it is important to preserve the continuity of the periodic structure along time. This is important to avoid generated speech to appear like whispered.

Our solution works implicitly in the time domain. We consider the voiced excitation signal as a pulse sequence. The position of each pulse within the current frame is computed. This computation is based on the pitch value and on the position of pulses in the previous frame. Thus, only pulses in the fourth quarter (75% overlapping) of the frame are computed during the frame processing. Others are derived from previous frames. The spectrum of the sequence weighted by a Hanning window is then computed. To avoid a metallic sounding of the speech, these spectra are multiplied by the spectra of a typical glottal pulse.
3.5. RESULTS.

The perceptual evaluation of a coding scheme implies long intelligibility and naturalness tests. Our first results give a syllabic intelligibility score over 80%. In these tests speech is qualified as good, quite natural.

The main qualities of our scheme are:
- a good behaviour towards long consonnants, semi-voiced sounds, nasals and liquids;
- the synthetic speech does not appear metallic;
- a good reproduction in the high frequency range;
- good transitions between different phonetic elements.

The drawbacks are:
- short vowels are attenuated and "melt" into adjacent consonnants;
- the speech sound a bit reverberant, mainly for short vowels;
- a non-stationary noise seems added to speech;
- occlusives are attenuated.

These drawbacks are mainly the consequences of the frame length, which reduces the adaptation capabilities of the system, of lack in our voicing estimator, and of the separation of the LPC synthesis (performed in time domain) and the residual synthesis (performed in frequency domain).

4. CONCLUSIONS.

We have presented a new parametric coding scheme using a representation in the frequency domain of the residual signal of the linear prediction. It uses a progressive voicing estimation defined as a function of frequency through a sub-band splitting. The synthesis is completely performed by the "overlap-add" algorithm.

This system synthesises good quality speech, quite different from the speech produced by other schemes. Our results confirm the interest of this frequency domain approach.

The computing charge is important but the system can be implemented in real time on specialized hardwares.

Studies about the importance of each parameter and about quantification are to be continued. The major problems are the consequences of analysis errors and the development of more robust analysis algorithms is necessary.

REFERENCES.


Authors' Address:
University of Liège
Department of Acoustics
Montefiore Institute B-28
B-4000 Sart-Tilman
Belgium

EUROspeech'89, Paris, France, September 1989 1328