DECOMPOSITION OF SPEECH SIGNALS INTO THEIR MODULATED COMPONENTS TO USE IN VOCODER

Paulo Henrique Carvalho and Allan Kardec Barros
Depto. Eng. Eletrica, Universidade Federal do Maranhão

Abstract: This work presents a simple variation of VOCODER based on two concepts: the formants and the modulated components of the speech signals. The suggested codification method extracts the modulated components (instantaneous amplitudes and frequencies), to be transmitted through a channel. We show the advantages of the proposed technique through comparison to other codification methods.

Keywords: Speech formants, wavelets, instantaneous amplitudes, instantaneous frequencies**,** analytic signal.

1. Introduction

Nowadays there is a great interest in the digital processing of signals for speech transmission through computer and telephone networks. The works in this area usually have a reduction in the bandwidth of the encoded signals. Thus, for an efficient performance of the speech transmission in networks with good quality and low transmission rate of bits per second, researches have been carried out to diminish the transmission rates of the encoders without meaningful loss in the naturalness and intelligibility of the speech.

There are different ways to implement speech encoders [1] and the commonly known can be divided into two basic classes: the wave format encoders, such as PCM (pulse code modulation) and ADPCM (adaptative differential pulse code modulation), and the parametric encoders or VOCODERS. These parametric encoders are based on LPC (linear predictive code). Thus, considering the cited classes, this article presents a simple and original variation of encoders, based on two widely known concepts: the formants and the modulated components of the speech signals.

In the method proposed here, the speech spectrum limited between 0 to 4 KHz is filtered into four narrow bands and the centre of each one is defined by the resonant frequencies or formants [8], which are defined by the power peaks of the spectrum. Then, we extract the AM-FM modulated components of the corresponding filtered analytic signal, which correspond to four instantaneous amplitude signals or envelop (AMP) and to four instantaneous frequency signals (FI). The idea is that the eight obtained signals are quantized, encoded and transmitted instead of the speech signal. Of course, from the modulated components, the original signal is recovered on the reception side.

Although the idea of decomposing the speech signal into AM-FM components is widely known, the particularities of each suggestion are in the way the modulated components are extracted. Maragos, Kaiser and Quartieri [4] [5] [6] [7] developed an algorithm based on the separation of energy in which the detections of the components are carried out with the use of an energy operator. Lu [9] [10] [11] suggested an alternative statistical model, where non-linear filter acts as a bank of band pass filters, and the instantaneous amplitudes and frequencies are determined respectively from the central frequency and the bandwidth of each filtered band. Kumaresan e Rao [12] [13] proposed an algorithm where the speech signal was decomposed into a polynome whose roots are located inside and outside the unitary circle of the complex plain from which the components are extracted. Finally, in this work it is used a model of extraction of instantaneous frequencies already used in electrocardiograms [2], based on the concept of the analytic signal.

2. Method

The method can be divided into three basic parts:

1º Determining the frequencies that correspond to four spectrum formants through an autoregressive model (AR) [14];

2º Filtering four bands around the peak frequencies using a wavelet [2]

3º Finally, determining the instantaneous amplitudes and frequencies of the filtered band from the analytic signal [15] [16] [17].

In this method some important aspects can be observed:

- The encoder can be classified as hybrid, once that the formants are related to the speech parameters, whereas the extracted modulated signals are encoded like the encoders in a wave shape;

- To obtain the instantaneous amplitudes and frequencies it is necessary to use the AM and FM proprieties of the speech spectrum, where the speech signal can be perfectly decomposed into its AM and FM modulated components. The decomposition or demodulation works like a reducer of the speech signals coding rates, once that the bandwidths that are occupied by the modulated signals can be considered much smaller than the original speech band. As it is illustrated in figure 1, each band filtered around the formant is modulated in amplitude by the AMP signal and in frequency by FI, hence when decomposing the original voice signal of each band; the instantaneous signals that correspond to the variations of AMP and FI through time are obtained. The smaller the variation rate of the AMPs and FIs signals, the smaller will be the obtained coding rate for the transmission.

2.1 Estimating the peak frequencies

To estimate the peak frequencies of the formants we use the Yule-Walker's autoregressive model [14]. These frequencies are extracted from where the peaks of the power spectrum occur. In other words, the frequencies can be extracted easily from the poles of the obtained transfer function.

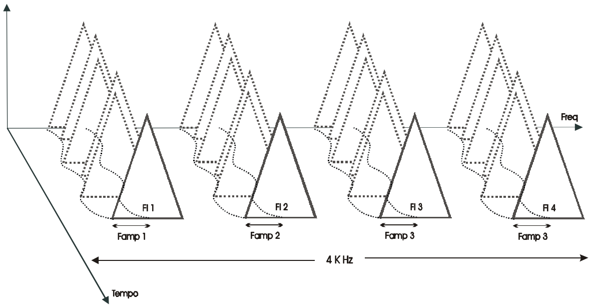
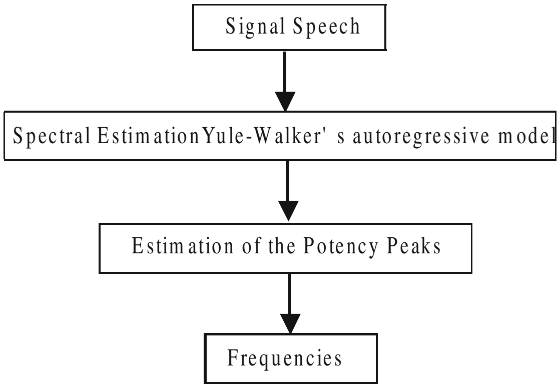


Fig. 1 – Amplitude and Frequency Modulation

The process of estimation of the frequencies can be summarized through the following block diagram:


Signal Speech
Spectral EstimationYule-Walker' s autoregressive model
Estimation of the Potency Peaks
Frequencies

For the spectral estimation, Yule-Walker's autoregressive model uses the autocorrelation method [14], where the coefficients of the autocorrelation function are determined, considering that the signal is stationary signal in that interval. In this work, as it is well-known, we consider an interval of 30ms of speech as being stationary.

After obtaining the coefficients of the correlation function, we determine the roots of the denominator of *(1)* which corresponds to the function's poles. We must observe that c_k in *H(z)* is the autocorrelation coefficients.

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} c_k z^{-1}} \qquad (1)$$

After obtaining the roots of the polynome, we only have to find the angular frequencies that correspond to the function's poles.

We see that the number of frequencies corresponds to the quantity of coefficients, but this method uses only the four first frequencies in order to reduce the quantity of encoded information.

2.2 Band Pass Filtering

The filtering around the frequencies that correspond to the peaks of the window's spectrum are carried out using a basic Wavelet [3] defined in (2).

$$\psi(t) = \frac{1}{2\pi}\frac{d}{dt}\left[\exp\left(-\pi\left\{\frac{\overline{\delta(t)t}}{2}\right\}^2\right)\cos\left(2\pi t\int_{\Omega}\delta(\tau)d\tau\right)\right],$$

$$\overline{\delta(t)} = \frac{1}{\Omega}\sum_{\Omega}\delta(t) \qquad (2)$$

where Ω is a small time interval. The filtered signal in this interval is given by

$$x_{\Omega}(t) = \int_{\Omega}\psi(\tau)\,y_i(t-\tau)\,d\tau \qquad (3)$$

2.3 Determining the Instantaneous Frequencies and Amplitudes

To obtain the instantaneous AMPs and FIs, we must first get the analytic signal, which can be obtained by just the imaginary component of the signal through the Hilbert transform [15] [16] [17] given by

$$\bar{x}_i(t) = \frac{1}{\pi}\int\frac{x(\tau)}{t-\tau}d\tau \qquad (4),$$

where $x(t)$ is the real signal filtered and $\bar{x}_i(t)$ is the obtained imaginary term.
The analytical signal from x(t) is given by

$$s_i(t) = x_i(t) + j\bar{x}_i(t) \qquad (5)$$

From s(t) we obtain AMP according to (6)

$$a_i(t) = \left|H[s_i(t)]\right| \qquad (6)$$

Also, the FI is obtained from $s(t)$ according to (7)

$$\omega_i(t) = \frac{d\phi_i(t)}{dt}, \quad \phi_i(t) = \arctan\left(\frac{\left|H[s_i(t)]\right|}{s_i(t)}\right) \qquad (7)$$

2.4 Recovering the speech signal

Once the components are encoded and transmitted, it will be necessary to recover the reception of the four real bands to form the speech signal. The recovered signal is real part of the analytical signal [17], which means:

$$y_i(t) = a_i(t)\cos 2\pi\int\omega(t)dt \qquad (8)$$

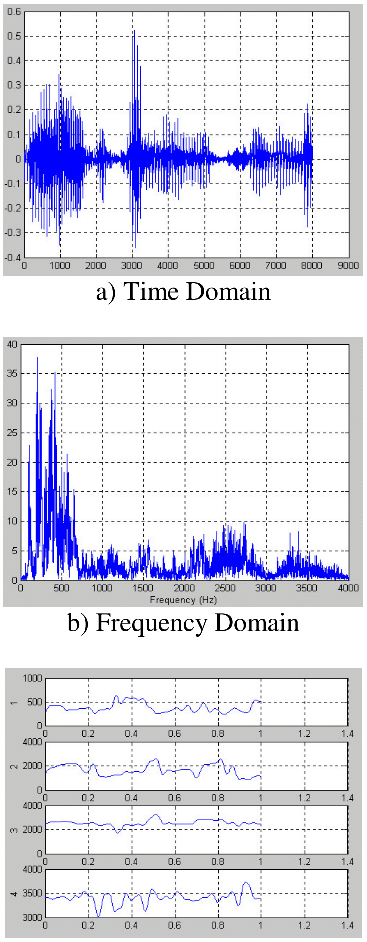
We must point out as well that the recovered signal loses part of the information, since its spectrum is formed only by the resultants of the four bands.

3. Results

We tested the method experimentally, where we used different phrases, in Brazilian Portuguese, with an average duration of three seconds. The phrases were generated with a sample rate of 48Khz and encoded into 16 bits to enhance the quality of the sample. Then they were sub-sampled to 8khz once the main application of the proposed VOCODER are telephone networks.

To illustrate the method, figures 2, 3 and 4 present the practical results obtained from one of the analyzed samples. Figure 2.a represents the first 1000ms of the produced signal and figure 2.b represents the correspondent frequency spectrum.

From the speech sample, the frequencies of the formants were calculated inside the 30ms windows, according to *(1).* Figure 2.c shows the correspondent frequencies of the formants found through time. Then, from the estimated frequencies the wavelets were built according to *(2)* and the bands were filtered, applying *(3)* to 30ms intervals. Figure 3 shows the spectra produced by the four filtered bands.



a) Time Domain

b) Frequency Domain

c) Estimation of the peaks

Fig. 2 – Original Signal

From the four filtered we determed the instantaneous amplitudes and frequencies correspondent (AMPs and FIs) to each band, using the Hilbert transform according to *(4), (5), (6)* e *(7).* Figures 4.a and 4.b show the inicial 100ms of the AMPs and FIs found signals. After obtaining the eight signals we went to the last step, which is the estimation of the codification through the quantization of the respective signals.

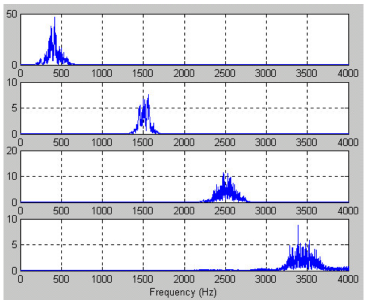
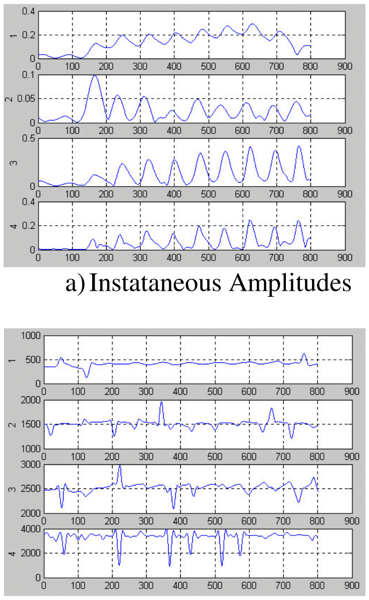


Fig. 3 – Filtered Band



a) Instataneous Amplitudes

b) Instataneous Frequencies

Fig. 4 – Time domain of the six AMPs and FIs signals of the four bands

Although this paper proposes a method for future use in speech codification without considering the transmission rate reached for the speech encoder, it is worth citing the estimation of the obtained band. To do so, it was initially necessary to estimate the width of the spectrum of the AMPs and FIs signals.

In this case, we observed in practice that to obtain an intelligible signal, the AMPs signals would have to have at least 100Hz of length and 400Hz of frequency. Thus, to fix the signal in these lengths, we used low pass filters to with cut frequency of 100Hz and 400Hz.

Another important point about the estimation of the transmission rate is the quantity of codification bits. In this case, we also observed in practice that to obtain an intelligible speech, the AMPs signals would have to have at least 16 quantization levels (4 bits) and FI 32 quantization levels (5 bits).

The final rate estimated was 19200 bps applying (9).

Transmission rate of the Encoded:

$$Tx = [\ (FORMANTS\ \ x\ BITS\_AMP\ x\ 2\ x\ F\_AMP\ ) + (FORMANTS\ \ x\ BITS\_FI\ x\ 2\ x\ F\_FI)\ ]\ bps \qquad (9)$$

were,

FORMANTS: number of formants;
BITS_AMP: quantity of codification bits of the AMP;
BITS_FI: quantity of codification bits of the FI signal;
F_AMP: length of the spectrum of AMP signal;
F_FI: length of the spectrum of the FI signal.

Transmission rate of the Encoded*:*

$$Tx = [\ (4\ \ x\ 4\ x\ 2\ x\ 100\ ) + (4\ x\ 5\ x\ 2\ x\ 400)] = 19200\ bps$$

Finally, to test the degree of intelligibility from the encoded modulated signals in 19200 bps, we applied (8) for the recovering the speech signal, then, we used the subjective opinion method (MOS) for the phrases samples. In this test, we gave five samples to ten listeners, which give a mark for each phrase they listened following the criterion in the MOS. The 50 marks varied between 2 and 3.5 and the average final value was near 3. Figure 5 show the final result of the MOS test of the VOCODER as well as a graphic comparing this to the traditional encoders, considering MOS's mark and the transmission rate in Kbps obtained.

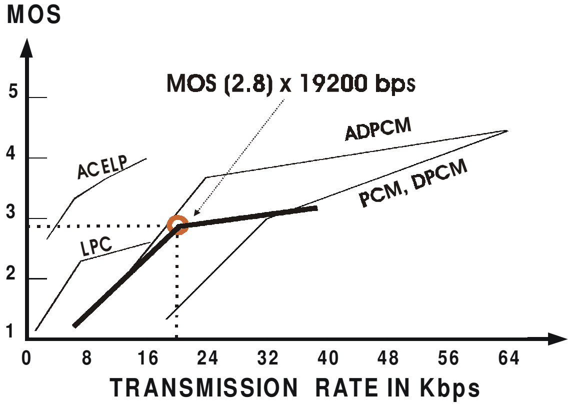


Fig. 5 – MOS x Transmission Rate in Kbps

Also, according to the figure 5, the MOS test was applied to other Kbps estimations of the codified one, where the width of the FIs and AMPs signals according table 1. Applying the 10 combinations to (9), 10 rates that varied from 7200 bps to 38400 bps were obtained. It must be pointed out that variations in the rates of the encoders were made by varying the width of the AMPs and FI signals, keeping the quantization levels constant.

According to results, it was observed that increasing the rate of the codified one from the optimum point (19200 bps), the gain in relation to the MOS is small, whereas the decrease of the rate from the same point makes the MOS drop fast.

F_AMP (Hz)	F_FI (Hz)	BITS_AMP	BITS_FI	Tx (bps)
100	100	4	5	7200
200	100	4	5	10400
100	200	4	5	11200
200	200	4	5	14400
100	**400**	**4**	**5**	**19200**
200	400	4	5	22400
100	600	4	5	27200
200	600	4	5	30400
100	800	4	5	35200
200	800	4	5	38400

Tab. 1 – Transmission Rate in bps

4. Discussions

The explanation for the behavior of the MOS in figure 5 for the different rates in table 1 is the following:

- For rates above 19200 bps, the increase in the bandwidth of the FIs signals does not contribute to the increase in the MOS. The loss of intelligibility of the encoder in this interval (from 19200 to 38400 bps) is directly linked to two factors intrinsic and preponderant to the encoder: the tranking formant LPC technique not very efficient and the loss of information about the speech after the filtration around the formants;

- On the other hand, for rates below 19200 bps, the decrease in the width of the FIs signals meaningfully contributes to the reduction of intelligibility, adding itself to the intrinsic factors mentioned before (filtration around the formants and tranking formant LPC). Hence, the MOS drops fast in the interval that goes from 19200 to 7200 bps;

- For the different rates, the variation in the bandwidth of the AMP signal does not meaningfully contribute to the variation in the MOS.

According to the obtained results, it was verified that the variations of the AMPs and FIs signals were high, according to figure 4. Thus, to implement en efficient encoder that allows the transmission in low rates, for instance, below 8Kbps, without compromising the quality, it was clear that it is necessary to reduce even more the variations of the AMPs and FIs signals. According to analysis, we observed that normally the frequency increases or decreases abruptly, when the amplitude drops to zero.

In addition to this, we observed that the variations of this signals occur because of the inversions of abrupt changes of phases od the speech signal. Thus, future papers should explain the relation between the alterations of the phase of the speech signal and their reflexes in the instantaneous frequencies and amplitudes.

Another issue that deserves deeper studies is the understanding of the inversions of the phase of the speech signal generated by the vocal treat, because once we have this knowledge, we will be able to mask or parameterize the inversions in the transmission and recover them in the reception, facilitating the reduction in the spectra of the modulated signals.

We must also observe that the proposal in this paper is limited by the presentation of the method, therefore the taking formants presented here through the LPC coefficients do not have the desired strength, and thus new studies are necessary for the use of new resources.

5. Conclusion

In this paper, we presented a proposal for the decomposition of speech signals into AM-FM modulated components to be applied to voice encoders.

The method is based on the fact that the transmission of the modelated speech can be substituted by the sending of its AM-FM modulated signals and, to extract these components, the method used the concept of formants and applied Yule-Walker's autoregressive model to obtain the frequency peaks of the speech spectrum. In addition, we used the band pass filtrations based on wavelets to obtain the narrow bands around the peaks, and lastly, we applied the Hilbert transform to get to the modulated components.

References

[1] M. Hasegawa-Johnson and A. Alwan: "Speech coding: fudamentals and Applications".

[2] A.K. Barros, J. Wisbeck and N. Ohnishi: "Heart Instantaneous frequency (HIF): An Alternative Approach to Extract Heart Rate Variability". *IEEE Trans. On Biomedical Engineering,* vol 48, No.8, August 2001, pp. 850-855.

[4] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", in *Proc. IEEE ICASSP-90*, Albuquerque, New Mexico, pp 381-384, April 1990.

[5] P. Maragos, T. F. Quartieri and J. F. Kaiser: " Speech nonlinearities, modulations, and energy operator", in Proc. IEEE ICASSP-91, Toronto, Canada, pp, 421-424, May 1991.

[6] P. Maragos, T. F. Quartieri and J. F. Kaiser: " On separating amplitude from frequency modulations using energy operators", in Proc. IEEE ICASSP-92, pp, II-1-II-4, 1992.

[7] P. Maragos and J. F. Kaiser: " Energy separation in signal modulations with aplication to speech analysis", *IEEE Trans. Signals Processing,* vol. 41, No.10, October 1993, pp.3024-3051.

[8] H. M. Hanson, P. Maragos and A. Potamianos: "Finding speech formants and modulations via energy separation: with aplication to a vocoder", in *Proc. IEEE ICASSP-93*, vol. II, pp 716-719.

[9] S. Lu and P. C. Doerschuk: "Modeling and processing speech with sums of AM-FM formants models*", International Conference on, ICASSP-95*, Acoustics, Speech, and Signal Processing, vol. 1, May 1995, pp 764-767.

[10] S. Lu and P. C. Doerschuk: "Nonlinear modeling and processing of speech based on sums of AM-FM formant models," *IEEE Trans. Signals Processing,* vol. 44, No.4, April 1996, pp.773-782.

[11] S. Lu and P. C. Doerschuk: "Demodulators for AM-FM models of speech signals: a comparison", IEEE International Conference on, Acoustics, Speech, and Signal Processing, vol. 1, May 1996, pp 263-266.

[12] R. Kumaresan and A. Rao: " Algorithm for decomposing an analytic signal into AM and positive FM components", *IEEE International Conference, On* Acoustics, Speech, and Signal Processing vol 3, May 1998, pp.: 1561 –1564.

[13] A. Rao and R. Kumaresan: " On decomposing speech into modulated components," *IEEE Trans. On Speech and Audio Processing,* vol. 8, No.3, May 2000, pp. 240-254.

[14] L. R. Rabiner and R. W. Schafer. "Digital Processing of Speech Signals". Englewood Cliffs, Nj: Prentice Hall signal processing series, 1978.

[15] S. Wardle, "A Hilbert-Transformer Frequency Shifter for Audio" Available online at *http://www.iua.upf.es/dafx98/papers*.

[16] J.O. Smith, "Generalized Complex Sinusoids." Available online at *http://www-ccrma.stanford.edu*.

[17] B.P. Lathi. "Modern Digital and Analog Communication Systems". 3nd Ed.