

A Non-linear Speech Analysis based on Modulation Information

Jean Rouat¹

ABSTRACT The paper proposes an analysis to estimate formant differences and/or pitch by relying on patterns of modulation observed in fibers influenced by a summation of close harmonics. It is shown that such modulation information seems to be robust to noise.

0.1 Introduction

The analysis and the recognition of speech spoken in noisy environment is a difficult and crucial task. Most of the speech recognizers alleviate the difficulties of these task by training on noisy data, assuming that the statistical properties of the noise will remain unchanged between the learning and the recognition phase. Other techniques assume that the speech source and the interference noise sources are geographically different. In summary, most of the effective techniques are useful for specific conditions.

On the other hand, perceptive analysis and auditory models enhance the discriminant information from non-stationary noise and are supposed to yield good performance in adverse conditions. However, their complexity and the difficulty of exploiting the dynamic output information with standard pattern recognition algorithms, restricts their integration in speech recognizers.

According, to Bregman, the spectral integration (or grouping of sounds) has been shown to be partially based on common amplitude modulation characteristics [1].

Furthermore, research work on automatic demodulation of speech can be motivated by the hypothesis that the human brain has neural cells which specialise in Amplitude Modulation (AM) and Frequency Modulation (FM) detection [4][11]. Moreover, simple nonlinear operators can enhance the AM or FM information in a signal [6][9] and can be used to process the output of a cochlea filterbank in order to obtain the AM information characteristic of speech signal [10].

¹Départ. des Sciences Appliquées, Univ. du Québec, CHICOUTIMI, (Québec), CANADA, G7H 2B1

0.2 Modulation in the auditory system

The poor spectral resolution of the cochlea can be an advantage when the speech signal is harmonic, as more than one harmonic of the signal can fall into the same channel producing an amplitude modulated signal with a modulation frequency equal to the fundamental frequency. Therefore, the fundamental frequency can be encoded in the temporal discharge patterns of auditory nerve fibers, the characteristic frequencies of which are very different from the fundamental frequency value.

As the bandwidths of the nerve fibers vary significantly even for fibers tuned to the same frequency and since the bandwidths may be broad, at least some of the auditory nerve fibers should always be able to encode the envelopes relevant to speech signals. Langner [5] shows how such periodicity coding is related to modulation information and analysis the role of the 'ON' and 'Chopper' neurones (in the Cochlear Nucleus) as pre-processors for enhancing the AM information coming from the auditory nerve fibers.

0.3 Application

In recent years, many studies concentrated on the coding of vowels [2][3] in the auditory nerve. For the nerve fibers whose characteristic frequency (CF) is close to a formant frequency, a phase-coupling to the formant frequency or to an adjacent harmonic is observed with little or no envelope modulation as the discharge pattern of the fiber is dominated by a single large harmonic component. Other fibers may show modulations corresponding to harmonic interactions. Therefore, the auditory system is able to track simultaneously formants and pitch by relying on phase-coupling of fibers which CF is close to the formant ("spectral analysis") and by relying on patterns of modulation for fibers influenced by a summation of stimulus harmonics [2][7].

To our knowledge, not much work has been done in order to exploit the pattern of modulation observed in the auditory nerve, for formants interaction estimation and for pitch estimation. Based on the fact that each inner hair cell is connected to 10 to 20 neurones and that each neurone is connected to only one inner hair cell, we assume that some of those neurones have a bandwidth broad enough to encode the envelope of the modulation produced by two interacting formants (F1 and F2 in /a/, F2 and F3 in /i/, F3 and F4 in /i/, etc.). Therefore, we present an analysis which does not require too many cochlear filters (24) and yields an accurate estimation of F2-F1, F3-F2 or F4-F3 and of the missing fundamental by extracting the modulation pattern at the output of cochlear filters with nonlinear operators.

0.4 The Analyse

The filterbank is comprised of a bank of twenty-four filters centred on 330Hz to 4700Hz [8]. The output of each filter is a bandpass signal with a narrow-band spectrum centred around f_i where f_i is the central frequency (CF) of channel i . The output signal $s_i(t)$ from channel i can be considered to have been modulated in amplitude and phase with a carrier frequency of f_i .

$$s_i(t) = A_i(t)\cos[\omega_i t + \phi_i(t)] \quad (1)$$

$A_i(t)$ is the modulating amplitude and $\phi_i(t)$ is the modulating phase.

In the present paper we use two techniques to extract the AM information. The first is a FIR digital Hilbert transformer to approximate the Hilbert transform ($s_i(t)_Q$) of $s_i(t)$, thus

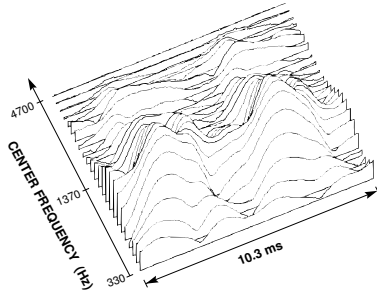


FIGURE 1.

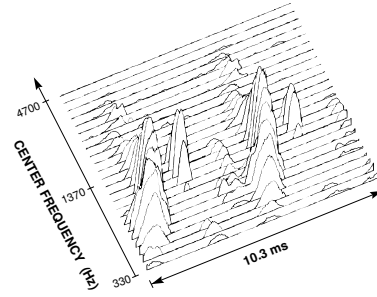


FIGURE 2.

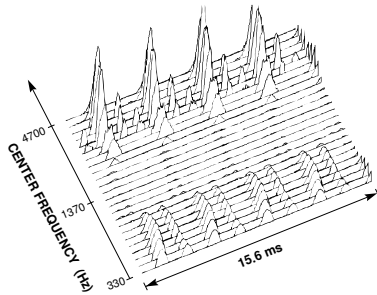


FIGURE 3.

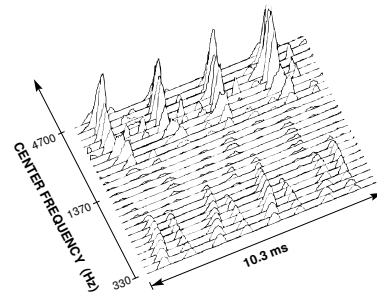


FIGURE 4.

$A_i(t) = \sqrt{[s_i(t)]^2 + [s_i(t)_Q]^2}$. The second uses the Dyn operator [9][10], where $\text{Dyn}[s_i(t)]$ is related to the time derivative of the instantaneous power of $s_i(t)$. The 3D representation is obtained by plotting the envelope $A_i(t)$, or the low-pass-filtered Dyn output, on the vertical axis. The x axis is the time and the y axis is expressed in Hertz according to the ERB scale (Equivalent Rectangular Bandwidth)[8].

0.5 Output of the Analysis

Figure 1 and figure 2 present two pitch periods taken from a french /a/ pronounced in a sentence by a male speaker driving a car at 130 km/h (closed window). Figure 1 represents $A(t)$ and figure 2 is the low-pass-filtered Dyn operator (after hard rectification). Dyn enhances significantly the modulation patterns due to the harmonics (residue pitch) and to the interaction between F1 and F2 (observed in channels 8 to 12). The period of the peaks in channels 8 to 12 is equal to $1/(F2-F1)$ and is independent of the noise power. Figures 3 and 4 compare the output of the low-pass-filtered and rectified Dyn operator for a clean and noisy /i/ segment spoken by a female. The pattern of modulation observed in channels 19 and 20 (F3-F2) and in channel 22 (F4-F3) is still preserved in the noisy speech segment (obtained from the clean segment by adding a white gaussian noise, SNR : 0.3 dB). In both figures, the pitch period is enhanced by the Dyn operator.

0.6 Conclusion

We have proposed a multichannel analysis based on AM extraction for harmonic or formant interactions (beats). We used a bank of cochlea filters whose design is based on psychoacoustic and auditory motivations [8] and we were able to show that patterns of AM modulation due to formant or harmonic interactions exist for such filterbank. It is very important to consider that the modulation patterns depend on the filter bandwidths in relation to the nature of the speech. Therefore, an adaptive filterbank would be made more realistic by taking into account the active processes occurring in the cochlea.

The modulation patterns might be used by the auditory system to perform sound grouping in order to separate speech from noise or to distinguish one talker from others (cocktail party). Depending on the bandwidths of the filters, the AM modulation can be dominated by pitch (residue) or formant interaction information.

In summary, the temporal analysis yields a robust information that characterises some spectral properties of the original signal. It is important to consider that such work is not a model of the auditory system but attempts to exploit some nonlinear and simple transformations performed by the auditory system to improve the contemporary speech analysis methods.

0.7 Acknowledgements

This work has been supported by the NSERC of Canada, by the FCAR of Québec, by the CMC and by the "Fondation" from Université du Québec à Chicoutimi. Many thanks are due to Alcatel R.T. who provided the car speech database. Many thanks to Daniel Morissette for his programming work.

0.8 References

- [1] A. S. Bregman (1985), "Spectral integration based on common amplitude modulation", *Jour. of Perception & Psychophysics*, 37, pp. 483-493.
- [2] B. Delgutte, (1980), "Representation of speech-like sounds in the discharge patterns of auditory nerve fibers", *JASA* 68, pp. 843-857.
- [3] B. Delgutte and N. Y. S. Kiang, (1984) "Speech coding in the auditory nerve : Vowels in background noise", *JASA* 75, pp. 908-918.
- [4] R.B. Gardner and J.P. Wilson (1979), "Evidence for direction- specific channels in the processing of frequency modulation", *JASA* 66, pp. 704-709.
- [5] G. Langner (1992), "Periodicity coding in the auditory system", *Hearing Research*, vol. 60, nb 2, pp. 115-142.
- [6] P. Maragos, T.F. Quatieri, and J.F. Kaiser (1992), "On Separating Amplitude from Frequency Modulations Using Energy Operators", *ICASSP-92*, pp. 2.1-2.4.
- [7] M. I. Miller and M. B. Sachs (1984), "Representation of voice pitch in discharge patterns of auditory nerve fibers", *Hearing Research*, vol. 14, pp. 257-279.
- [8] R.D. Patterson (1976), "Auditory filter shapes derived with noise stimuli.", *JASA* 59, 3, pp.640 - 654.
- [9] J. Rouat (1993), "Nonlinear operators for speech analysis", in "Visual Representations of Speech Analysis", pp. 335-340, edited by M. Cooke, S. Beet and M. Crawford, J. Wiley

and Sons.

[10] J. Rouat, S. Lemieux and A. Migneault (1992), " A spectro- temporal analysis of speech based on nonlinear operators", Int. Conf. on Spoken Language Processing, Banff, October 12 to 16, Vol. 2, pp 1629-1632.

[11] B.W. Tansley and J.B. Suffield (1983), "Time course of adaptation and recovery of channels selectively sensitive to frequency and amplitude modulation", JASA 74, pp. 765-775.