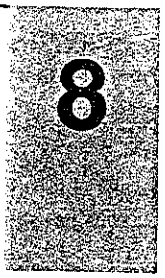
Auditory-based wavelet representation



Christophe d'Alessandro

1 INTRODUCTION

This paper presents briefly the auditory-based wavelet representation (AWR). The adaptation of an auditory frequency scale to the wavelet representation was proposed in d'Alessandro & Beautemps [3] and discussed for speech in d'Alessandro & Beautemps [2]; a speech representation model, closely related, to AWR was proposed in d'Alessandro [1]. The reader is referred to these papers for more information and references on these methods.

The wavelet representation is a linear nonparametric representation method, closely related to linear filtering. It provides a local time-frequency description of the signal: at the analysis stage, wavelet coefficients are obtained by correlation between the wavelets and the signal, and at the synthesis stage the signal is reconstructed as a discrete weighted sum of wavelets. These two points (speech analysis using a spectrographic format, and speech synthesis using wavelets) will be discussed in relation to the Sheffield data. Displaying the wavelet coefficients provides auditory-based wavelet spectrograms (AWS). Different types of AWS are discussed.

Compared to models of the auditory periphery, AWS can be considered as a somewhat simplified functional representation of the first stage of analysis (i.e. cochlear filtering). The aim is not to provide a refined auditory model, but to propose an auditorily-justified tool in acoustic-phonetics.

Another application of AWR is speech synthesis. AWR gives a complete resynthesis scheme. For speech synthesis or modification, it is possible to reduce this redundant representation to its most important components. Resynthesis from AWS and reduced AWS indicates those of the acoustic speech parameters that seem more perceptually relevant. Section 2 presents a brief description of the methods. Section 3 discusses of these methods in relation to the Sheffield data.

2 DESCRIPTION OF THE METHODS

2.1 Overview of AWR

AWR may be interpreted in terms of linear filtering, both at the analysis and at the synthesis stages. These wavelets are defined on a set of points in the time-frequency domain, and are weighted by a set of coefficients which are dependent on the analysed signal.

To interpret the representation, the wavelets must be localised functions both in time and frequency. In other words, they are chosen with a main spectro-temporal maximum, and with negligible values outside a (small enough) time-frequency domain. One can therefore interpret the local behaviour of the signal by comparison with the analysing wavelets, and one can consider AWR as a decomposition of a signal on a discrete set of time-frequency points.

The AWS examples presented here were obtained using a critical-band (Bark scale) finite impulse response (FIR) filterbank. The prototype wavelet was a Hamming window. The filterbank used constant 1 Bark, 6 dB bandwidth filters. In each band, analysing wavelets (re-

Visual Representations of Speech Signals: Martin Cooke, Steve Beet and Malcolm Crawford (eds.)

lated to the impulses responses of the filters) were obtained by contraction/dilation and modulation of the Hamming window.

2.2 Amplitude, phase and filtered instantaneous frequency AWS

As wavelet analysis is equivalent to filtering, one can display multi-band filtered signals using amplitudes and phases of the wavelet coefficients. This display is fairly different from a classical spectrogram because of the linear (instead of bilinear) nature of the analysis, and because of the frequency scale. We prefer here a spectrographic format where only the coefficients with positive phases are plotted (i.e. half-wave filtered signals). This spectrographic format provides several representations of the important features of speech signal (dominant frequencies and periodicity). As complex coefficients are computed at the analysis stage, in each band, computation of the time-derivative of phase provides instantaneous frequencies of the filtered signal. Dominant frequencies are enhanced on these spectrograms.

2.3 Reduced AWS

The ability of the representation to retain relevant acoustic parameters may be checked using resynthesis. The first type of resynthesis is direct resynthesis: the sum of all the weighted wavelets. The quality obtained with resynthesised signals is perfect, excepted for a little bandwidth reduction, as the chosen wavelets are not ideal bandpass filters. A reduced resynthesis scheme, using only the wavelets present at dominant time-frequency points was defined. Local energy concentrations in time and frequency were detected using linear predictive analysis and short-time Fourier analysis, and peak picking. Fourier wavelets (i.e. with constant resolution) were selected on these frequency tracks, and a signal was resynthesised using only this reduced wavelet representation. These reduced resynthesised signals are perceptually identical to resynthesised signals.

One step forward in AWS reduction is to synthesise the signal directly from the acoustic parameters extracted from the AWS. This reduction was performed using the following parameterisation: sinusoidal representation of the F0-F1 area, formantic representation (using dominant frequencies and temporal envelope modulations) above. The quality obtained is excellent, although the reduced synthetic signal sounds a bit different from the direct resynthesised signal. The reduced signal is not identical with the original signal because of the non-linear relationship between the acoustic parameters used for synthesis (temporal envelopes, dominant frequencies) and speech production parameters (formants, F0). Therefore, the parameters used for synthesis are only approximations of the true speech parameters. This approximation is good enough to preserve the general quality, but not to give identical signals.

3 APPLICATION TO THE SHEFFIELD DATA

3.1 Sheffield data

Owing to the limited space available for diagrams, we limited analyses to sound examples referred to as 'timit.syl', 'clean.syl' and 'dirty.syl'.

'Timit' : this example is a well-recorded sentence, uttered by a male American speaker, and sampled at 16 kHz. Formants are clearly visible, both on wideband spectrograms and AWS. This example was judged easy to read by an expert in (American English) spectrogram reading. Fig. 1 shows 'timit.syl' in AWS format; an instantaneous frequency AWS corresponding to fig. 1 is shown in fig. 2, and a reduced AWS in fig. 3.

'Clean': the quality of this example is rather poor, due to the recording conditions. The signal is low-pass filtered at 2.8 kHz. Without any prior knowledge, this sentence appeared difficult to understand both for French and American listeners in our laboratory. The spectrogram reading experiment was also not completely successful, because it appeared difficult to find the formant frequencies and motions, and because of the unusual phonetic realisations of

some vowels. Figure 4 shows 'clean.syl' in AWS format, and an instantaneous frequency AWS corresponding to fig.4 is given in fig. 6.

'Dirty': Figure5 shows 'dirty.syl' in AWS format, and an instantaneous frequency AWS corresponding to fig.5 is given in fig. 7.

3.2 AWS reading

The acoustic features that are apparent on AWS are dependant both on the acoustic signal produced by the vocal apparatus, and on the time-frequency resolution of the analysis. Analyses of quasi-periodic signals, like voiced speech, indicate the following:

- Below a frequency threshold, individual harmonics are resolved. This threshold is dependant both on F0 and critical bandwidth. In each band a time-varying sinusoid is obtained, whose amplitude and phase are set according to vocal tract and source amplitudes and phases, and whose frequency is F0 times the harmonic number.
- Above this threshold, two or more harmonics merge into a single filter, but formants are resolved (i.e. a single band contains at most one formant). The signal is therefore an amplitude and frequency modulated sinusoid, resulting from beats between the harmonic components. Unfortunately, the relationship between amplitude and frequency modulation of the waveform in a band, and the underlying speech production parameters (F0, formants, etc.) is nonlinear, and generally not analytically tractable. Nevertheless, one can prove analytically, at least for two components beating in a band, that the average period of amplitude modulation equals 1/F0, and that the mean of frequency excursion during a fundamental period is a local maximum of the spectral envelope.
- Above another frequency threshold, two or more formants merge into a single filter. The beats pattern is very complex, resulting in a spectral mass on AWS, with an amplitude modulation frequency greater than F0, and a rather wide frequency modulation.

For speech parameters, generally speaking, the first point above corresponds to the F1 region, the second point to F2 and F3, and the final point to the region above F3.

4 CONCLUSIONS

In this paper an auditory-based wavelet representation was introduced for speech analysis and synthesis.

The auditory spectrograms obtained are somewhat different to classical spectrograms; it is not clear that they will prove better for phoneme identification. Generally speaking, AWS show less contrast than spectrograms, because it is a signal decomposition and not an energy distribution. Other auditory-like processing might be applied after this first stage of analysis, to enhance contrast. For instance lateral inhibition may enhance spectral contrast and short-term adaptation may enhance temporal contrast.

On the other hand, both reading and resynthesis indicates that the acoustic parameters which are visible on AWS could give a more complete description of the speech signal than those visible on spectrograms: F0 and parameters related to voice quality are visible. It is well known that voice quality (naturalness, speaker individuality, etc.) is difficult to assess precisely using spectrograms. AWS could work better here. For AWR reduced synthesis, the accurate estimation of speech parameters from the wavelet coefficients is a difficult problem, as soon as several harmonics merge in a single band.

This representation takes advantage of the interplay between perception and production in speech analysis. It might provide another tool for studying several open questions such as:

how is F1 perceived in relation to F0, what is the influence of higher formants on vowel quality, and what is the perceptual relevance of (the amplitudes and phases of) lower harmonics?

Acknowledgements: I would like to thank Denis Beautemps for programming the analysis-synthesis system, Lori Lamel for help in spectrogram readings and phonetic transcriptions, and both of them for fruitful discussions.

REFERENCES

[1] C. d'Alessandro (1990), 'Time-frequency speech transformation based on an elementary waveform representation', *Speech Comm.*, **9**, 419-431.

[2] C. d'Alessandro & D. Beautemps (1991), 'Justification perceptive du spectrographe auditif', *Proc. XIIth Int. Cong. Phonetic Sciences*, vol. 5, 86-89.

[3] C. d'Alessandro & D. Beautemps (1991), 'Transformation en ondelettes sur une échelle fréquentielle auditive', *Proc. XIIIth GRETSI Symp.*, 745-748.

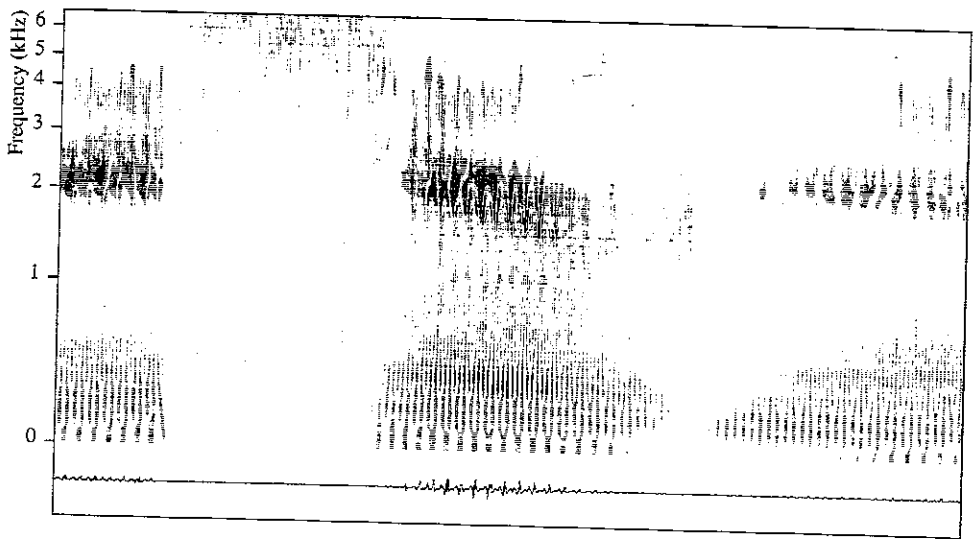


Fig. 1 AWS of 'timit.syl'.

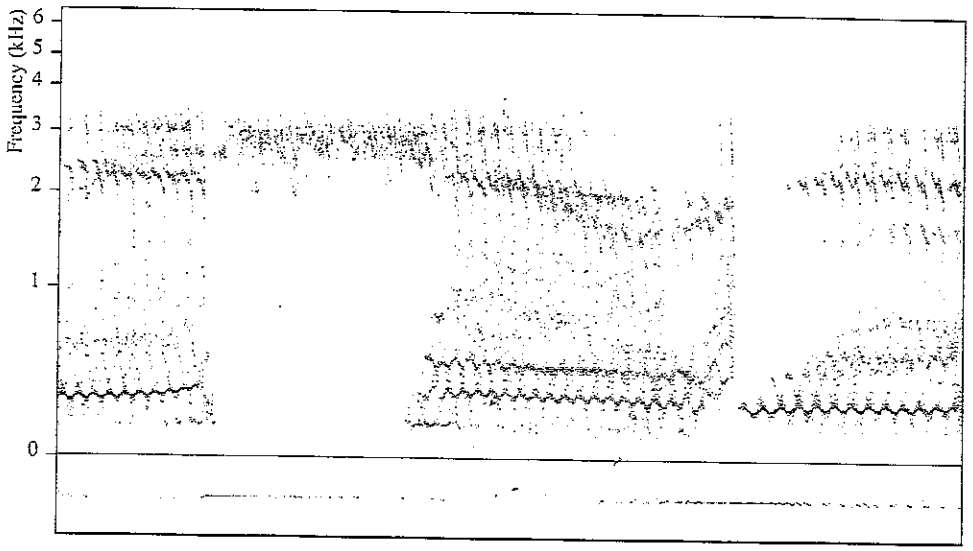


Fig. 2 Instantaneous frequency AWS of 'timit.syl'.

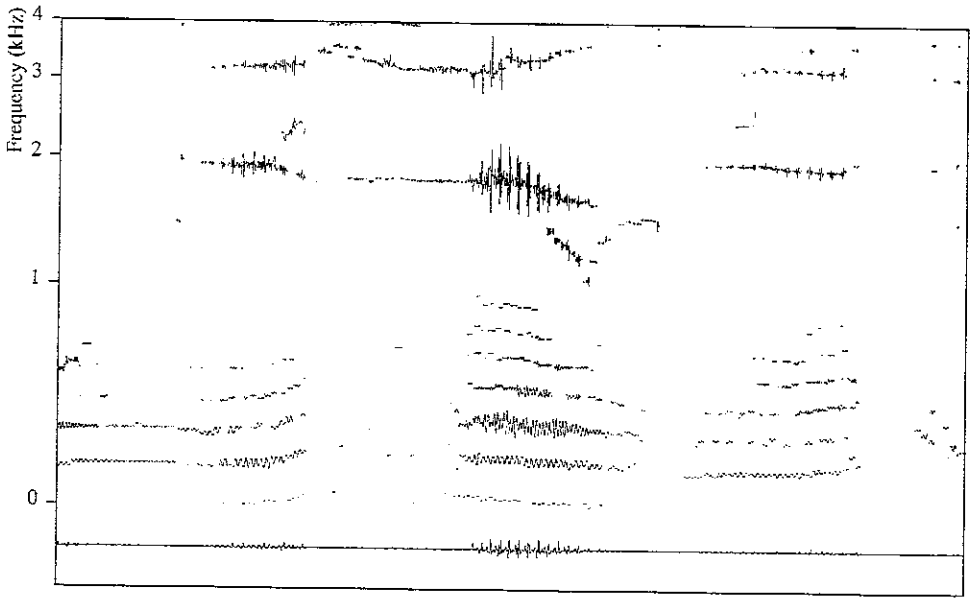


Fig. 3 Reduced AWS of 'timit.syl'.

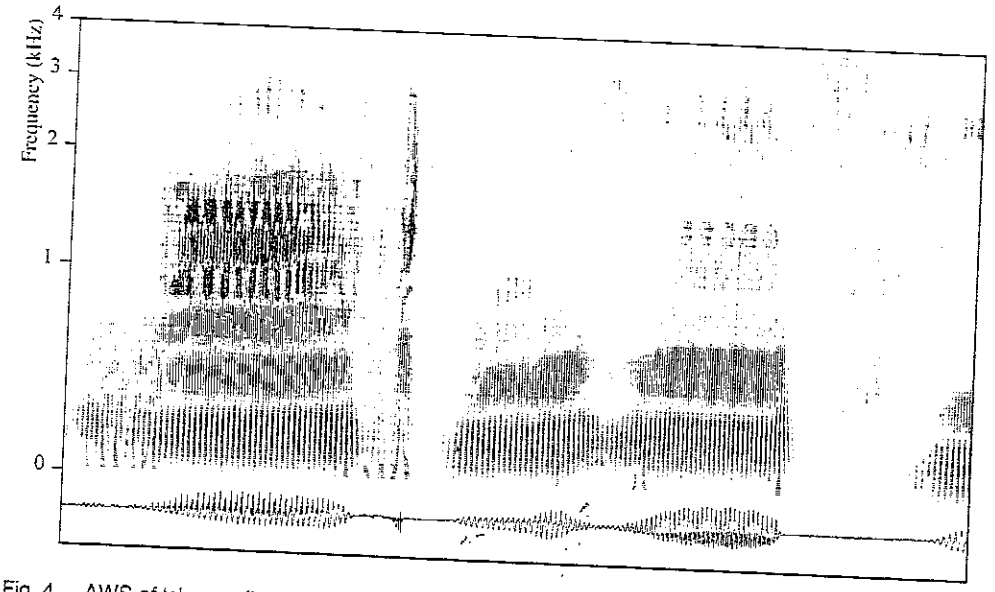


Fig. 4 AWS of 'clean.syl'.

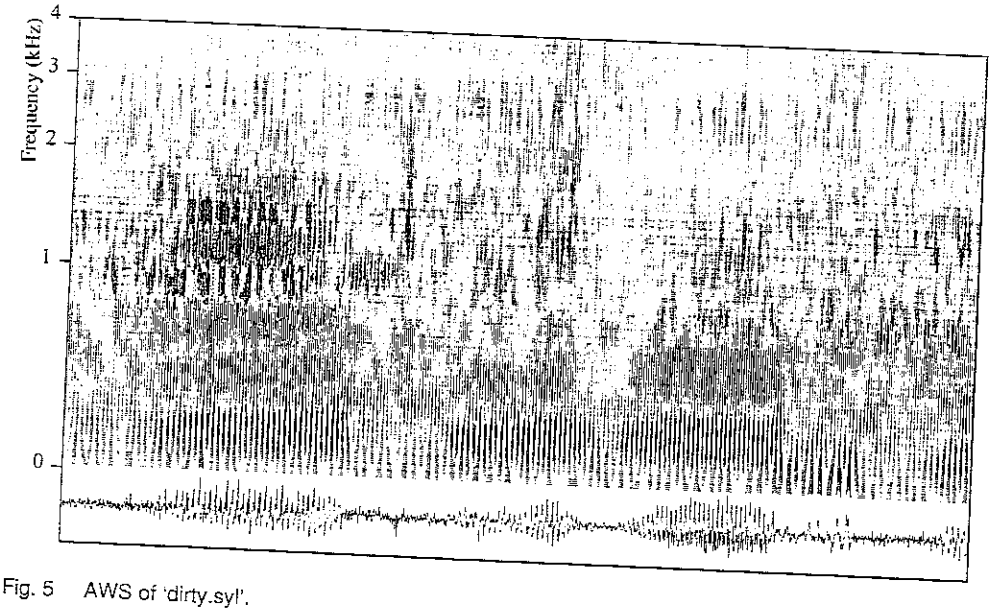


Fig. 5 AWS of 'dirty.syl'.

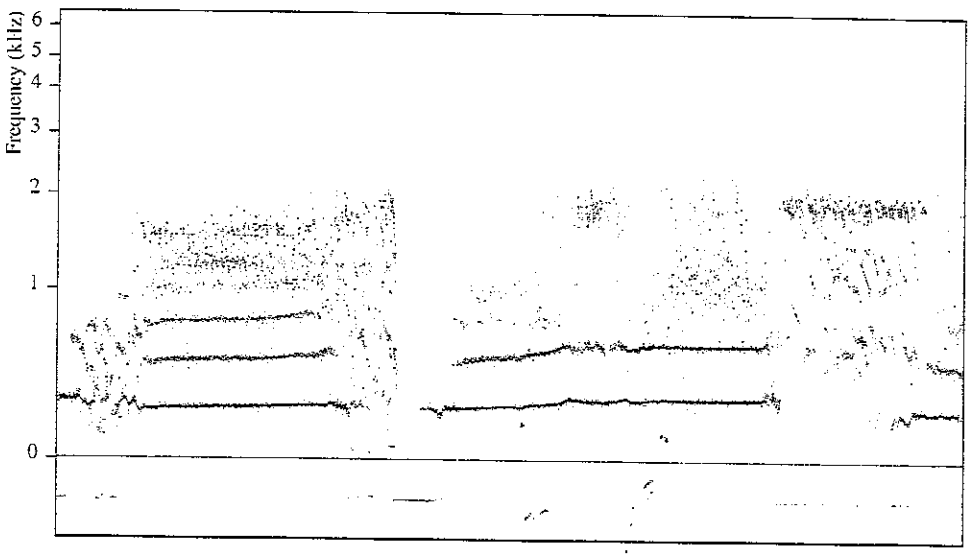


Fig. 6 Instantaneous frequency AWS of 'clean.syl'.

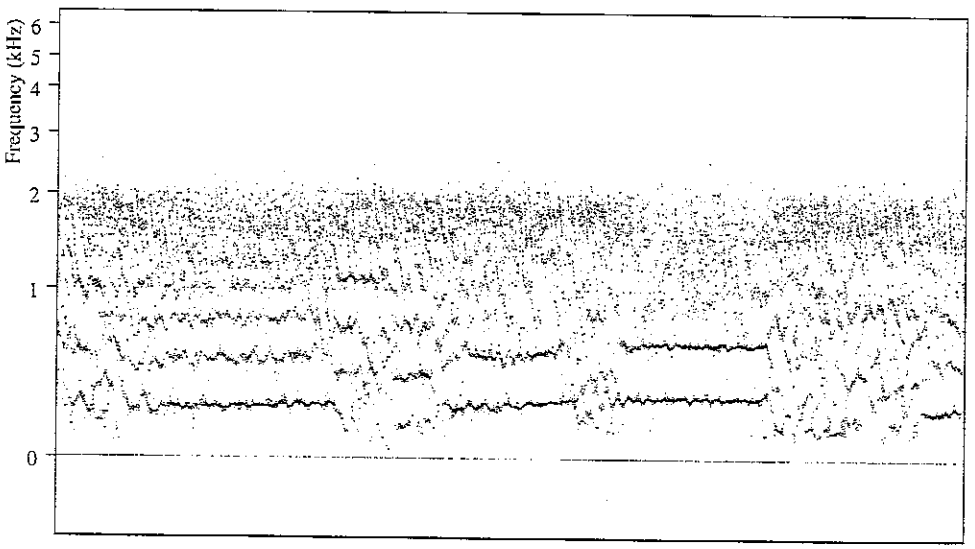


Fig. 7 Instantaneous frequency AWS of 'dirty.syl'.