SEGREGATION OF CONCURRENT SPEECH WITH THE REASSIGNED SPECTRUM

G.F. Meyer[1], F. Plante[2] and F Berthommier[3]

1) Dept of Computer Science, Keele University, Keele, Staffs., ST5 5BG, UK, georg@cs.keele.ac.uk
2) Dept of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, UK, F.Plante@liverpool.ac.uk
3) Institut de la Communication Parleé, INPG, F-38031 Grenoble, France, bertho@icp.grenet.fr

ABSTRACT

Modulation maps provide an effective method for the segregation of voiced speech sounds from competing background activity. The maps are constructed by computing modulation spectra in a bank of auditory filters. If the modulation spectra are computed using a conventional DFT, windows of 200ms duration are necessary. The reassigned spectrum, a new time frequency representation [1,2], allows a reduction in the window size to 50ms without loss of performance.
The algorithm is tested on a 'double vowel' identification task that has been used extensively in psychophysical experiments [3,4].

1. INTRODUCTION

Human speakers are remarkably good at understanding speech in noise backgrounds. Psychophysical data suggests that listeners group features in complex 'auditory scenes' into streams which allow selective listening. These streams are formed by grouping all segments in an auditory scene that share features, such as harmonicity (F0 grouping), start or end time (common fate) or spatial location (binaural cues).

To be able to segregate the streams in a signal it is necessary to expand it into a rich representation. The approach used here is to split the signal into 32 band-pass filtered channels, using an auditory filterbank. Each channel is further expanded by Fourier transforming the half-wave rectified and low-pass filtered output. The map codes amplitude modulation frequency against channel frequency. A typical AM map for a single vowel /y/ at 126 Hz F_0 is shown in figure 1. Energy is concentrated in stripes, corresponding to the harmonics of the fundamental frequency (F_0). The representation allows two simultaneous voiced sounds to co-exist as separate (but interleaved) patterns, provided the patterns have different F_0s . Spectra can be recovered by estimating the F_0 for a window of speech and sampling the output of all filters at the first five harmonics of the F_0.

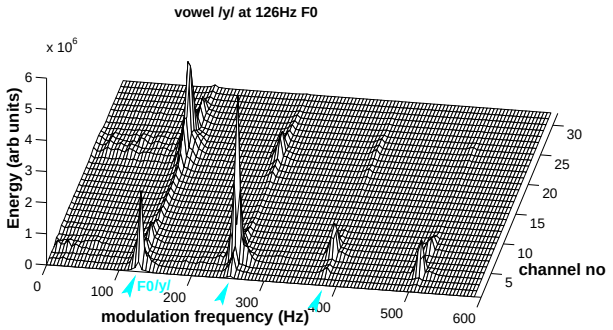

Figure 1: Amplitude modulation map for a single vowel /y/ at 126Hz F_0. The map shows energy as a function of modulation frequency and channel number. Channel centre frequencies range from 100Hz to 4.7kHz. The map was computed using a conventional FFT with 204.8ms windows. Voiced speech sounds form characteristic patterns where each harmonic is expressed as a ridge in the map. Spectra can be recovered by summing energy for the initial five harmonics.

Modulation maps are a good model for human perceptual data, which shows that human listeners are able to segregate vowels, provided their F_0s are one or two semitones (6-12%) apart.

A key requirement for the system is the need to operate on relatively short windows to deal with the changeable structure of speech. The map above is constructed with 204.8ms windows, which is much too long to be useful for spontaneous speech. This paper shows how the re-

This Work is supported by the EU-HCM Sphere Project and EPSRC Grant GR/L 05655.

assigned spectrogram allows a similar frequency resolution as given above with windows of only 51.2ms duration.

2. THE REASSIGNED SPECTRUM

The reassignment method was first proposed by Kodera et al. [1] in order to improve the resolution of the spectrogram. The method assigns the value of the spectrogram to the centre of gravity in the analysis window, rather than the centre of the window as in the normal Fourier Transform. When applied to the spectrogram, the point of assignment is moved in time and frequency. However, when time information is not required, only the frequency displacement is computed, leading to the definition of the reassigned spectrum. The frequency displacement uses the phase of the Fourier spectrum, and can be easily computed using a ratio of Fourier Transforms [2]:

$$\nu_r = \nu - \Im m\left\{\frac{STFT_{dh}(\nu).STFT_h^*(\nu)}{\left|STFT_h(\nu)\right|^2}\right\}$$

where

ν_r is the reassigned frequency point,

ν is the frequency point using the normal Fourier Transform, $STFT_h$ is the Short Time Fourier Transform using the window *h*, *dh* is the time derivative of the window *h*.

The frequency resolution of the reassigned spectrum is unlimited. However, the frequency points are no longer uniformly spaced. To facilitate processing the reassigned spectrum is resampled at regular points. The value of the spectrum at each of this point is computed by summation of all of the points falling into the same bin. Bins containing no frequency points are set to 0.

2.1 Comparison with conventional pre-processing

Contour plots of amplitude modulation maps for the concurrent vowels /a/ (100Hz F_0) and /i/ (142Hz F_0) are shown in figure 2. The top panel shows a conventional DFT, computed for 51.2ms windows, the lower panel shows the same representation computed with the reassigned spectrum. The contours show iso-energy lines.
The vowel F_0s are optimally separated at 100/142Hz. While both representations are able to resolve the representation in the AM domain, it is clear that the frequency resolution of the conventional DFT is not sufficient to localise energy in the representation. The reassigned spectrum is able to localise AM energy much more precisely. This is particularly evident at 400Hz AM where the DFT fails to resolve two separate peaks.

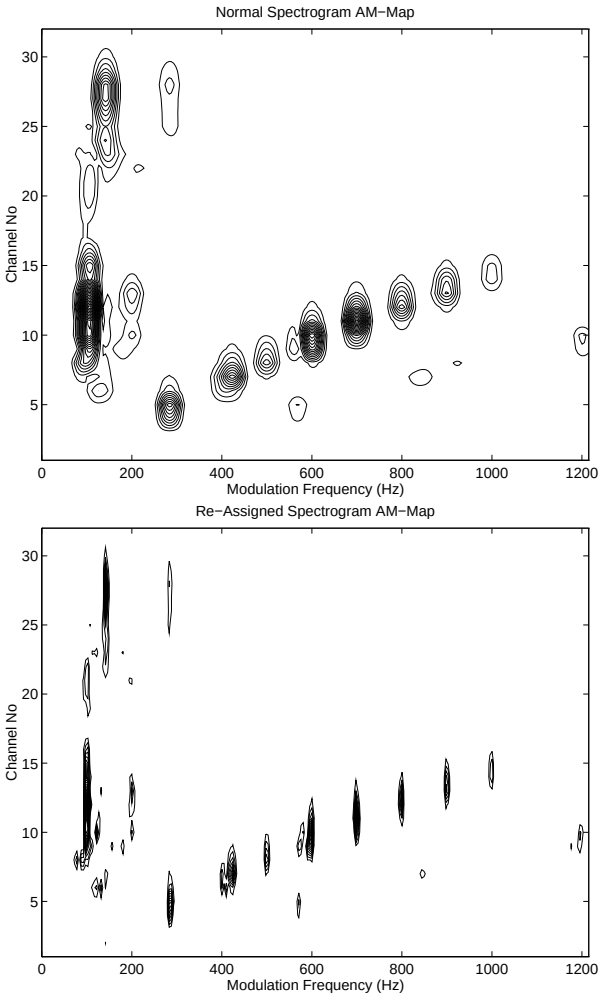

Figure 2: Contour plots of amplitude modulation maps for pairs of synthetic vowels /a/ and /i/ at 100Hz and 142Hz fundamental frequency respectively. Both maps were constructed using 51.2ms Hanning windows. The top AM map was constructed using a conventional FFT, the lower map is based on the reassigned spectrum.

2.2 Pulling it all together

A key problem for the application of amplitude modulation maps is to maintain good spectral separation of the harmonics with small FFT window sizes. If 204.8ms windows are used, each FFT bin is 4.9Hz wide, which is sufficient to model human segregation performance [3,4]. To be effective for spontaneous speech, much smaller window sizes must be used. The reassigned spectrogram allows this. Figure 3 shows the percentage of energy that can be recovered by interpo-

lating the energy at each harmonic. If 204.8ms windows are used 42% of the total energy is located at the bin positions, this reduces to 28% if 51.2ms windows are used. For comparison the same data for a conventional FFT is plotted: if 51.2ms windows are used, less than 10% of the energy is located at the harmonic positions. The reassigned spectrum is four times oversampled, to compare results the data obtained with the conventional FFT was presented in a 51.2ms Hanning window, within a 204.8ms zero-padded frame.

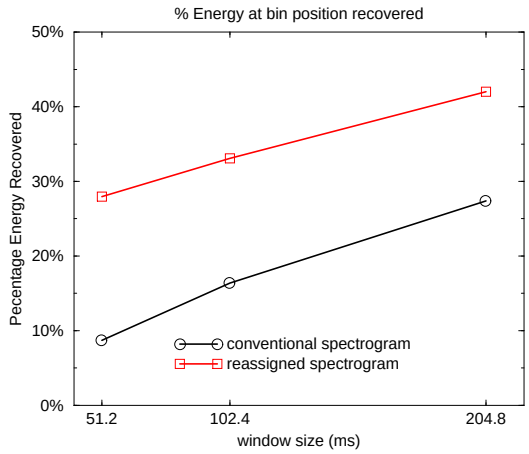

Figure 3: Percentage of energy that can be recovered by extracting the harmonics as a function of window size for amplitude modulation maps computed with conventional and reassigned spectrograms.

3 Recognition experiments:

The aim of the AM-maps is to use the reassigned spectrogram to segregate concurrent vowels. Performance is evaluated on synthetic vowel pairs. One vowel, /a/, is fixed at 100Hz F0, the second vowel, /i/, F0 is varied between 100 and 130Hz. The spectrum at the target F0 is recovered and compared against templates for isolated vowels. A Euclidean distance measure was used. The vowels /a/ and /i/ were chosen because they have formants in almost complementary positions at 650, 950 and 2950Hz (a) and 250, 2250 and 3050Hz (I). The first and second formant of the /a/ will leak into the mid frequency (low intensity) region of the vowel /i/ and vice versa. Both vowels have the same rms energy. When both vowels have the same fundamental (100Hz), the vowel /a/ is dominant - i.e. the difference between the extracted spectrum to the template for /a/ is smallest. As the target vowel (/i/) F_0 increases, the distance to the /i/ template reduces while the distance to the /a/ template increases. If 204.8ms windows are used, both the conventional and reassigned spectrogram require around 6 Hz separation to correctly identify the target. This point is identified by the cross-over point of the graphs, the further to the left it is, the better. If 51.2ms windows are used, the conventional spectrogram requires 25Hz F_0 separation to identify the second vowel while the reassigned spectrogram is still able to separate vowels with only 6% F_0 difference.

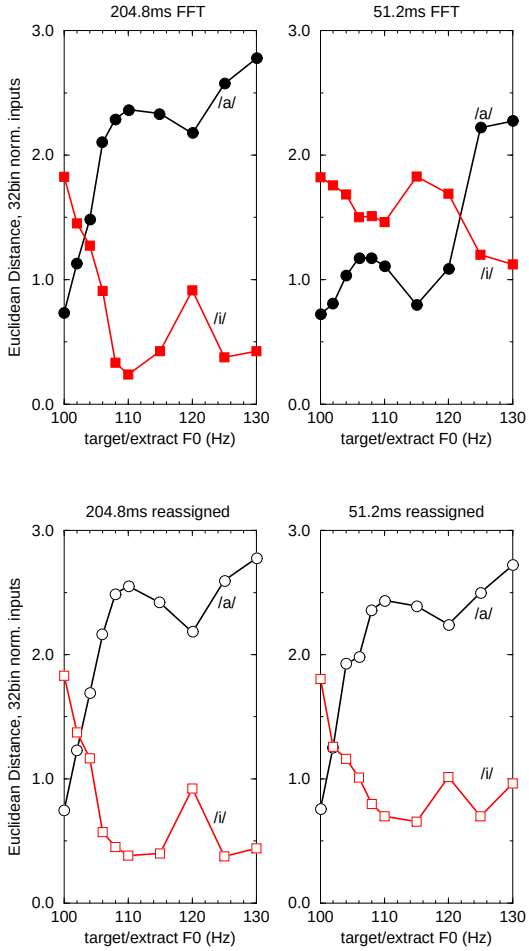

Figure 4: Euclidean distance measured between normalised extracted spectra of the vowel /i/ in the background of /a/ at 100Hz at the given frequencies and a template for isolated vowels /a/ and /i/. A detailed description is in the main text.

Double vowel recognition experiments

A classical set of experiments carried out in psychophysics measures the identification of simultaneously

played vowel pairs by human listeners. These experiments were adapted to show the degree to which simultaneous vowels interact within the representation as the analysis window size reduces.

To evaluate the degree of interaction between vowels, the five long British vowels /*a*,ɛ,*i*,*o*,*u*/ were synthesized at fundamental frequencies between 100 and 142 Hz and the corresponding spectra were recovered from the map and compared against a set of templates. The system performance is quantified as the average recognition performance for all possible vowel combinations. Chance performance is 4%, but pairs of equal vowels are always recognised correctly (20%). The data is plotted against fundamental frequency for the second vowel for 204.8ms, 102.4ms and 51.2ms (DFT) and 51.2ms, 25.6ms and 12.8ms (reassigned spectrum) windows, figure 5.

The performance of the reassigned spectrum based algorithm is roughly equivalent to that of the DFT with windows two to four times as large. If only 12.8ms windows are used the reassigned spectrum fails to segregate the vowels.

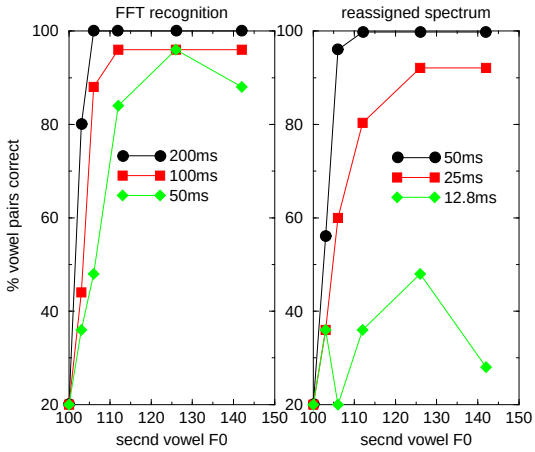

Figure 5: Recognition performance for pairs of simultaneous vowels. The left panel shows the percentage of correctly recognised vowel pairs using DFT decomposition, the right panel shows the data for the reassigned spectrum. Note that the window sizes for the DFT are 200, 100 and 50ms while the reassigned spectrum uses windows of 50, 25 and 12.5ms respectively.

Comparison with other techniques:

The proposed algorithm shares many features with harmonic selection strategies where long term spectra are computed and sampled at multiples of the fundamental frequency [5]. The technique allows a precise estimation of the spectrum, provided the fundamental frequency of both stimuli is steady within the analysis window and that the pitch is precisely estimated. For spontaneous speech these restrictions cause problems, particularly in the high frequency region where small F_0 changes or pitch estimation errors cause severe problems since the estimation errors multiply as the harmonic number increases. The AM map algorithm differs from harmonic selection in that spectra are recovered by sampling partial spectra at only at the first five harmonics. Spectral information is recovered in the filterbank channel domain. This means that resolution is limited to 32 channels, but this resolution is more than adequate for speech recognition where 20 channel MEL-scale filterbanks are commonly used.

Conclusions:

Modulation maps are an elegant way to separate simultaneous voiced speech sounds. The main problem in the application of the algorithm, the need for long term spectra, can be overcome with the reassigned spectrum. Synthetic vowel sounds were chosen because they are easy to control and allow a comparison with psychophysical data, but the technique will be applied to real speech in the future.

References:

[1] K. Kodera, RE. Gendrin, C. de Villedary, “Analysis of time-varying signals with small BT values” IEEE Trans. ASSP, Vol. 34, pp.64-76, 1978.

[2] F. Auger, P. Flandrin, “Improving the readability of Time-Frequency and Time Scale representations by the reassignment method” IEEE Trans SP, Vol.43, pp. 1068-1089, 1995.

[3] G.F. Meyer and F. Berthommier, “Vowel segregation with amplitude modulation maps: a re-evaluation of place and place-time models” Proc ESCA workshop on auditory basis of perception, Keele, pp. 212-215, 1996.

[4] G.F. Meyer “Expanded signal representations for auditory scene analysis” Proc Inst of Acoustics V18(9), pp. 3-10, 1996.

[5] T.W. Parsons, “Separation of Speech from interfering speech by means of harmonic selection” J Acoust Soc Am, Vol. 98 pp.1866-1877, 1976