Proceedings of the Institute of Acoustics

# EXPANDED SIGNAL REPRESENTATIONS FOR AUDITORY SCENE ANALYSIS

**G F Meyer**

**Dept of Computer Science, Keele University, Keele, Staffs., ST5 5BG.**
**email: georg@cs.keele.ac.uk**

## 1. INTRODUCTION

Auditory models have been tested as front-ends for speech recognition systems, but typically do not improve recognition significantly compared to conventional systems. The key problem in interfacing auditory front-ends to recognition systems is the high data rate these models generate. A common solution is to down-sample the data generated by the models (e.g. Ainsworth and Meyer, 1994). This process, while making the front-ends compatible with the pattern matching stages, removes all fine detail from the representation. Psychophysical data suggests that it is precisely this fine detail, which is used by human listeners to deal with situations where conventional recognition systems fail. An area, which received particular attention is the recognition of vowels in a background of a second vowel.

Experimental data shows that human listeners use two main cues to segregate simultaneously presented vowels: One cue is the fundamental frequency difference ($\Delta F_0$) between the two vowels (Assmann and Summerfield, 1990; Berthommier and Meyer, 1995), the other is the relative onset time of the two vowels (McKeown and Patternson, 1995). The aim of this work is to investigate the utility of signal representations, which expand conventional signal representations along the axis of perceptually relevant cues. Model performance on a vowel-vowel recognition task is compared with human performance.

## 2. GROUPING BY FUNDAMENTAL FREQUENCY

Human recognition performance for pairs of simultaneous vowels improves as the fundamental frequencies ($F_0$) of the two vowels diverge from each other. It follows that there must be some internal representation or mechanism which segregates them. A number of models performing this segregation computationally have been proposed. The most direct solution is to employ a 'harmonic selection' strategy. A typical example is proposed by Parsons (1976). Fourier transforms are applied to 51.2ms windows of the waveform, the fundamental frequencies in the signal are estimated and harmonics belonging to each of the estimates are extracted. This strategy has two important draw-backs: harmonic selection requires very precise pitch estimation to recover high frequency harmonics. It also does not explain human performance because auditory filters do not resolve harmonics in the speech pitch range above about 1kHz. While auditory filters are not selective enough to resolve single harmonics in the frequency domain, precise information about the signal is represented in the time domain for each filterbank channel.
A number of existing models use autocorrelation analysis of the temporal discharge pattern to explain human performance (e.g. Weintraub, 1986; Assmann and Summerfield, 1990 - "place-time model"; Meddis and Hewitt, 1992). Autocorrelation models extract energy at given periodicities across all channels or select channels according to dominance of a particular $F_0$ in a given channel. Both strategies distort the signal and are sensitive to relative level differences.

A robust alternative to the time-lag representation is to perform a frequency decomposition of the discharge pattern in each channel. This type of representation, modulation frequency against channel frequency, has been shown physiologically at the level of the inferior colliculus (Schreiner and Langner, 1988; Langner and Schreiner, 1988).

### 2.1 Experiment 1: Human Performance for $F_0$ guided vowel segregation

The experiment is a repetition of earlier experiments by Assmann and Summerfield (1990). Subjects are presented with pairs of synthetic long English vowels */ɑ, ɜ, i, ɔ, u/*. The synthesiser and parameters are the same as in Assmann and Summerfield (1990). Each stimulus consists of two different vowels with identical rms energy, one with a $F_0$ of 100Hz, the other with a $F_0$ ranging from 100 to 200Hz, only a subset of the data is shown. Both components of the pair start and end simultaneously and have a duration of 200ms with 25ms Hanning windows at the start and end. The subject's task was to identify the vowel pair. Signals are presented dichotically at around 50dB(A) using a 16bit D/A converter and Sennheiser HD435 headphones.

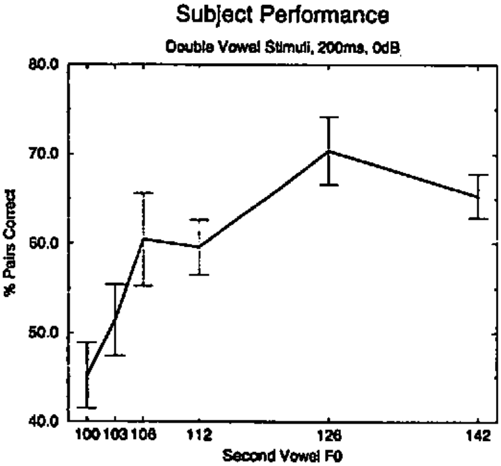

*Figure 1: Human performance for the double vowel recognition task.*

Without any segregation cues subjects recognise around 50% of the pairs correctly. As the $F_0$ difference rises, performance increases to 70%. One vowel in the pair is recognised in 98% of the trials. The data is consistent with previous experiments, where similar performance increases were found for Dutch (Scheffers, 1983), German (Zwicker, 1984), English (Assmann and Summerfield, 1990) and French listeners (Meyer and Berthommier, 1996).

### 2.1 The Model

The discharge pattern in each channel of an auditory filterbank codes both the average energy in the channel (as average discharge rate) and fine timing information (as the discharge pattern) (reviews: Evans, 1978; Langner, 1992). To illustrate the point the response to a complex stimulus consisting of two amplitude modulated tones is shown in fig. 2. Each tone consists of a carrier frequency, 1kHz and 2kHz in this example, which is 100% amplitude modulated with an envelope of 100Hz and 142Hz respectively. A DFT of the signal is shown in the leftmost panel. Each of the tones activates a different place of the auditory model. Channels with centre frequencies of 1.05kHz and 2kHz show the highest average activity, right panel. The discharge pattern, middle, shows both the carrier and envelope of the stimulus clearly as a function of time. As both tones activate different regions in the filterbank, they could be separated by grouping channels or activity within channels into separate streams, which would be perceived separately. This is the basis of $F_0$ guided stream segregation.

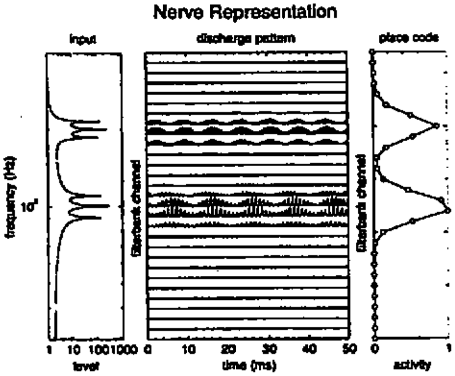

*Figure 2a: Signal representation in an auditory model for two 100% amplitude modulated sines. The carrier frequencies are 1kHz and 2kHz, modulated at 100Hz and 142Hz respectively. A DFT of the stimulus is given on the left, the temporal discharge pattern in a bank of filters in the centre panel and the average activity in the filterbank on the right. Note that the filterbank is unable to resolve the harmonics.*

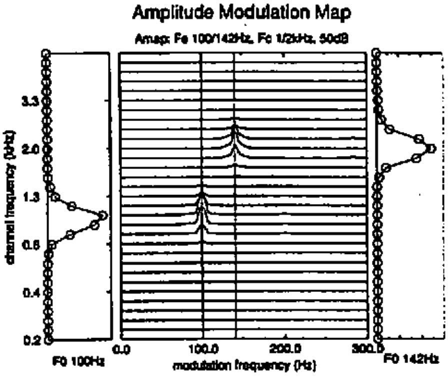

*Figure 2b: Amplitude modulation map, constructed by applying a DFT to 204.8ms windows of the filterbank output. The map shows two clear peaks, corresponding to the carrier and modulation frequencies of the two signal components. Streams can be extracted by sampling the map along the modulation frequency axis. The recovered spectra (left 100Hz, right 142Hz) correspond to the components of the perceptual streams.*

### Auditory Preprocessing

The first stage in the information processing is an auditory filterbank. The signal is split into 32 0.5 Bark spaced channels with characteristics frequencies ranging from 100Hz to 4.7kHz. Each filter is a linear fourth-order recursive gamma-tone filter (Darling, 1991; Kortekaas and Meyer, 1994). The output of each channel is scaled to approximate human hearing thresholds. The hearing thresholds are calculated for each channel from a polynomial regression model for data reported in Fay (1988)[1].

As shown in figure 2a, the discharge pattern in auditory filterbanks codes both the carrier frequency and the envelope of amplitude modulated signals. The envelopes are extracted by half-wave rectification and low-pass filtering (T=2ms) of the output of each filter. This process is analogous to that seen at the hair-cell transduction stage. Modulation frequencies are computed by Fourier transforming a window of activity in each nerve channel. The resultant map for 32 channels is shown in figure 2b - centre panel. To remove any DC component and low frequency beating a high pass filter with a time constant of 4ms is applied.

[1] The equation reads: $Tf_d = 4.08.c_f^{-1} + 17.47 - 45.23.c_f + 45.76.c_f^2 - 19.59.c_f^3 + 4.11.c_f^4 - 0.41.c_f^5 + 0.02.c_f^6$. Where $c_f$ denotes the channel centre frequency in kHz and $TH_d$ is the absolute hearing threshold in dB SPL.

The map shows energy in the modulation spectrum against modulation frequency for each of 32 channels in the auditory model. Carrier and envelope frequency for each object in the scene can be read off the two axes. Spectra for isolated tones can be recovered by sampling the map along target $F_0$. The spectra recovered in this way are shown to the left (100Hz) and right (142Hz) of the map, figure 2b.

Speech sounds have many characteristics of simple amplitude modulated sine waves. Each of the harmonics is modulated by an envelope, which is directly related to the fundamental frequency. If a voiced speech sound is used to drive the model, a characteristic striped pattern appears.

Why AM-maps?

Amplitude modulation maps expand the underlying place representation in the frequency domain which is the perceptually relevant domain. The representation in the maps is fully separated along a 'harmonicity' axis. Sounds, such as voiced speech, which contain multiple harmonicities have to be 'collected together' to recover their spectra. This can be achieved by estimating the fundamental frequencies of the signals to be extracted. Energy is recovered across all channels in the filterbank representation, so that it is sufficient to recover only the spectra corresponding to the five initial harmonics. The recovered spectra have the same frequency resolution as the filterbank.

### 2.2 Modelling Human Performance

The stimuli that were used in the psychophysical experiments were also used as the basis to a modelling study. The segregation stage was used to drive a pattern matching stage. Spectra are recovered by linearly interpolating the energy at the known target $F_0$ for the first five partial spectra in the map.

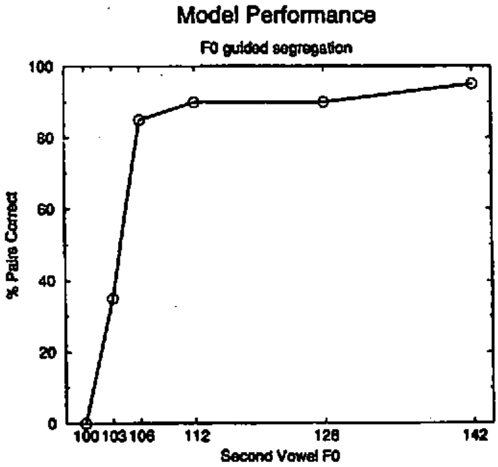

*Figure 3: Model performance for the double vowel recognition task. The model fails of both vowels have the same $F_0$, Performance increases with $\Delta F_0$. After one semi-tone the model recognises 90% of all vowel pairs correctly.*

A simple Euclidean distance metric between the recovered spectrum and a template generated by averaging the recovered spectra for isolated vowels with $F_0$s ranging between 100 and 200Hz was used. The proposed pattern matching stage could without doubt be refined, but the objective is to mimic human performance qualitatively, not to implement a speech recogniser.

The set of stimuli consisted of non-identical pairs of six vowels, leading to 15 possible combinations, and a chance level of performance of 6.67%. The $F_0$ guided segregation stage fails when both vowels have the same fundamental frequency because the recovery algorithm extracts the same data twice leading to the same recognised vowel. Subjects in this situation recognise around 50% of all vowel pairs correctly. As the $\Delta F_0$ between the two vowels increases, recognition performance increases to 90%. This level of performance is significantly higher than that achieved by human listeners, but shows the same qualitative behaviour, a minimum of one semi-tone $F_0$ difference is required to segregate the vowels. Larger differences are required if the window size used for analysis is reduced, as this reduces the frequency resolution of the frequency analysis.

A number of factors may account for the fact that the model, while showing the correct qualitative data performs significantly better than human listeners. The most obvious explanation is that the machine only 'knows' about five different vowels while humans are adapted to deal with a much broader range of signals. Performance of the model could be reduced by shortening the window size, thereby reducing frequency resolution in the AM domain, addition of noise to the representation, or a reduction in the number of filterbank channels. While little is known about human signal processing or pattern matching strategies, it is difficult to justify any of these steps.

## 3. VOWEL SEGREGATION BY COMMON FATE

In experiment 1 all vowels start and end together. Another strong cue used by human listeners to segregate simultaneous vowels is known as 'common fate'. As soon as one of the vowels precedes the other, the second vowel is much easier to identify, even if both vowels end together. A simple frame-based recognition model, without a mechanism allowing for persistence of previous frames cannot explain this performance increase.

### 3.1 Experiment 2: Vowels with different start times

The experiment uses the same vowels and presentation parameters as the previous experiment, but this time the $F_0$ of both vowels is fixed at 100Hz while one vowel precedes the other by up to 800ms. The second vowel duration for all stimuli is 200ms. If humans only used an instantaneous spectral estimate for vowel recognition then allowing one vowel to precede the other would not improve performance. The data shows that, as the time lag between the vowels increases, recognition performance for the vowel pairs increases from 50% to around 70%. The increase is not linear, but asymptotes at around 200ms. The 0ms time lag point is equivalent to the 100Hz V2 point in figure 1.

### 3.2 Modelling segregation by Common Fate

The proposed model uses a frame based representation of the signal. An intuitive model explaining the increase in performance, as the time lag between the vowels increases, is that successive frames are subtracted from each other and that only the remainder is used to drive the pattern matching stage. If both vowel onsets are coincident, both will appear simultaneously. If one vowel precedes the other, then the two vowels appear in sequence. To model this the signal was processed as a sequence of overlapping frames. Each frame was Hanning windowed with windows of 50, 100, 200 and 400ms

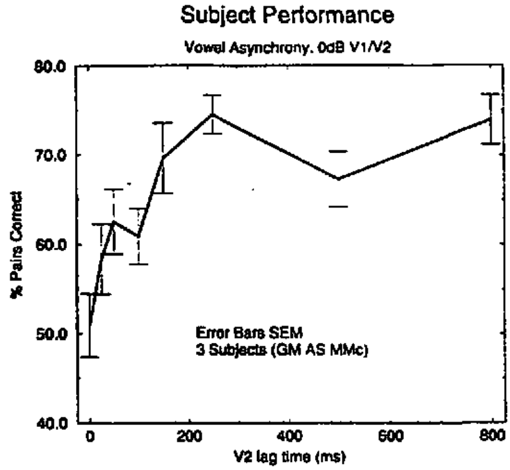

*Figure 4: Subject performance in the 'common fate' experiments. Subject's recognition performance increases as one vowel is allowed to precede the other. Both vowels have the same $F_0$ and end point.*

duration to model different integration times of the auditory system. As the model was used to process stimuli with the same $F_0$ the computation of amplitude modulation maps would not have had any benefit. The cochlear activation patterns were computed by integrating all auditory model output after windowing. We assume that the first window is triggered at the onset of the first vowel. The time-lag axis defines the offset between the windows. If the segregation cue is not present, here where both vowels are coincident, the model will use the composite signal and recognise the same vowel twice. Subjects, as in the previous experiment, recognise around 50% of all pairs correctly. As the offset between the vowels increases, recognition performance increases, again to around 90%. The slope of the increase depends on the size of the window. The model approximates human data if integration periods of 200ms are used.

**Integrating the models**

The common fate model described uses the average discharge rate in a window, not the amplitude modulation map. This is purely for computational convenience. Signals with identical $F_0$ overlap 100% in the AM map, so that segregation is not possible, the result signal would be identical to the place code.

Both stages can be integrated very easily by computing amplitude modulation maps and subtracting successive images. The segregation can then be performed on the image resulting from the subtraction step. The model predicts recognition rates of 100% where both cues are present.

## 4. RECOGNITION WITHOUT GROUPING CUES

Neither of the two proposed segregation model is able to explain the 50% recognition rate achieved by human listeners when no segregation cues are present. This is not surprising, but suggests that a third, high level, process is employed. Zwicker (1984) suggested that the recognition process might have a feedback component: Pattern matching is carried out on the vowel pair and a dominant vowel is recognised. After the 'dominant' vowel is recognised, it's template is subtracted from the recovered spectrum and the remainder is passed through the pattern matching stage again. This is consistent with human data. In only 2% of all trials is none of the component vowels recognised. Both segregation models recognise one of the two component vowels in all cases.

This process was implemented as a subtraction step and predicts a recognition rate of around 50% without any segregation cues. Both the $F_0$ guided segregation and the common fate model are able to cleanly segregate the vowel pairs, the spectral subtraction is a very crude method and relies on the availability of normalised templates and extracted spectra.

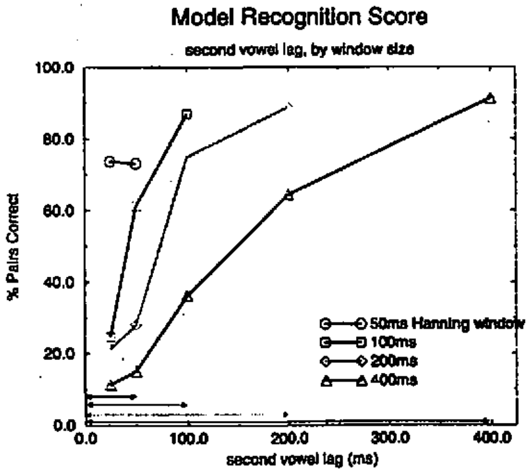

*Figure 5: Model performance for the common fate stimuli. The model integration time can be set arbitrarily. An integration window of 200ms gives a good qualitative approximation to human data. At zero time lag, none of the pairs is recognised.*

### 5. CONCLUSION

Humans use a range of simple cues to segregate concurrent vowels, and probably other phonemes, for recognition. Models based on these cues are able to replicate human performance increments but they cannot fully account for human performance. If sounds are presented without segregation cues, i.e. dichotically, with the same $F_0$, in $F_0$-phase, and with the same onset and offset, models using segregation cues predictably fail. A simple subtraction model is able to account for the baseline performance seen in human listeners.

**Segregation models for speech recognition**

So far the segregation models were used to drive very simple pattern matching stages on an almost trivial problem: recognising simultaneous vowels. Before auditory scene analysis can be used for speech recognition in the large, some fundamental problems have to be solved.

The models can be used in two modes, we use it by extracting spectra from the map, so that, assuming a pitch tracking stages that was able to deal with concurrent speakers existed, the model could be used directly to drive a speech recogniser. An alternative is to use the whole map as the input to a pattern matching stage, the computational overheads are much higher, but this procedure would be the only way to ensure minimal commitment during the preprocessing stages and allow an explicit model of selective attention. Dealing with more than a single speaker at a time is a formidable challenge for current speech pattern matching technologies.

### *Acknowledgements*

This work was supported by the EU-HCM SPHERE project.

### 7. REFERENCES

Ainsworth, W.A. and Meyer, G.F (1994) Recognition of plosive syllables in Noise: Comparison of an auditory model with human performance. J Acoust Soc Am, 96, 687-694.

Assmann, P.F. and Summerfield, A.Q. (1990) Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. J Acoust Soc Am, 88, 680-697.

Darling, A.M. (1991) Properties and implementation of the gammatone filter: A tutorial. Work in Progress (5), Dept of Phonetics and Linguistics, University College, London.

Evans, E.F. (1978) Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons. Audiology, 17, 369-420.

Fay, R.R. (1988) Hearing in vertebrates: A psychophysics data book. Hill-Fay Associates.

Kortekaas, R.W.L. and Meyer, G.F. (1994) Vowel onset detection using models of the auditory periphery and the nucleus cochlearis: Physiological Background, Institute for Perception Research, TU Eindhoven, Rapport no. 963/1994.

Langner, G. and Schreiner, C.E. (1988) Periodicity coding in the inferior colliculus of the cat. I. neuronal mechanisms. J Neurophysiol, 60, 1799-1822.

Langner, G. (1992) Periodicity coding in the auditory system. Hearing Research, 60, 115-142

McKeown, J.D. and Patterson, R.D. (1995) The time course of auditory segregation: Concurrent vowels that vary in duration. J Acoust Soc Am 98 1866-1877.

Parsons, T.W. (1976) Separation of speech from interfering speech by means of harmonic selection. J Acoust Soc Am, 60, 911-918.

Scheffers M.T.M. (1983) Shifting vowels: Auditory pitch analysis and sound segregation. PhD Thesis Groningen University.

Schreiner C.E. and Langner, G. (1988) Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. Journal of Neurophysiology, 60, 1823-1840.

Weintraub, M. (1986) A computational model for separating two simultaneous talkers. Proc ICASSP '86 81-84.

Zwicker, U.T. (1984) Auditory recognition of diotic and dichotic vowel pairs. Speech Communication, 3, 365-277.