Proceedings of the Institute of Acoustics

PATTERNS OF CONFUSION MADE BY A MODEL OF DOUBLE-VOWEL IDENTIFICATION: A COMPARISON WITH HUMAN DATA

S. J. Makin and G. J. Brown

Sheffield University, Dept. of Computer Science, 211 Portobello St., Sheffield, S1 4DP
{s.makin,g.brown}@dcs.shef.ac.uk

1. INTRODUCTION

1.1 Background

In a real acoustic environment, signals reaching the ears of a listener are usually mixtures of several sound sources. In order to make sense of such complex auditory "scenes", the perceptual system must analyse the mixture to recover the constituent sources. This problem has come to be known as "Auditory Scene Analysis" (ASA), [4]. In ASA, acoustic mixtures are thought to be decomposed into "elements" and groups of these elements recombined to form auditory "streams", on the basis of the likelihood of them having arisen from the same source. There are numerous cues which have been shown to play a part in this process, such as onset/offset synchrony, common AM, common spatial origin, etc. A particularly important factor is fundamental frequency (f0). Harmonically related components tend to perceptually "fuse", whereas differences in f0 promote segregation. There have been numerous studies showing that listeners are better able to identify two simultaneous speech sources if they have different f0's. This was shown for continuous speech sentence material by Brokx and Nooteboom [6]. Also, a widely used technique for investigating this phenomena, the double-vowel paradigm, was introduced by Scheffers [16]. A pair of steady-state, synthetic vowels are presented simultaneously, with identical onset and offset, at the same amplitude, to the same ear. Subjects are required to identify both vowels. Performance is significantly above chance when the vowels have the same f0 and improves when a difference in f0 is introduced, improving rapidly up to about 1 semitone and asymptoting between 1 and 2 semitones. This finding has been used to suggest that listeners use an f0-guided segregation strategy in identifying two vowels that differ in f0 [16],[3],[11]. A number of physiologically-motivated computational models of double-vowel identification have been proposed in the past, such as: (a) Assmann and Summerfield [3], "non-linear, place time" spectral pattern-matching model; (b) Meddis and Hewitt [11], autocorrelation-based model; (c) Culling and Darwin [7], model exploiting waveform-interactions produced by low-frequency beating; (d) Meyer and Berthommier [12], AM map model; (e) de Cheveigne [8], harmonic interference cancellation model; (f) Brown and Wang [5], Neural oscillator model. Models (b)-(f) have all been shown to demonstrate the pattern of increasing overall percentage identification up to 1 semitone which humans display, although quite different schemes are employed by most of them. A more discriminative measure of the models' performances is therefore necessary.

1.2 Patterns of Confusion

One possibility is to investigate the pattern of confusions made by each model and compare this with human listeners' confusion data. This will enable us to distinguish between the models, not in terms of their overall identification performance (which is not very informative), or in terms of their ability to reproduce the pattern of improvement with $\Delta f0$ (which many seem to achieve), but in terms of the particular mistakes which they make and how closely this matches human behaviour. This paper reports the preliminary results of a pilot study examining the extent to which the Meddis and Hewitt model can reproduce some human confusion data. Section 2 outlines the structure of the model, section 3 describes the experimental procedure used to obtain the human data, section 4 presents the results and initial comparison of human and model data and section 5 contains a discussion of some pertinent methodological considerations and some of the work planned to extend this study in the near future.

2 THE MODEL

The processing stages of the Meddis and Hewitt model are illustrated schematically in fig.1. They are:

1. A model of the auditory periphery (100 channel Gammatone filterbank and Meddis [10] inner hair cell models).
2. Correlogram (running autocorrelation function computed for each channel).
3. Dominant pitch estimation via a summary autocorrelation function (SACF), formed by summing the channels of the correlogram. The pitch estimate is the frequency corresponding to the period of the highest peak in the "pitch region" of the SACF. The pitch region is defined as the portion of the SACF between 4.5 and 12.5 ms (80 to 222 Hz).
4. Formation of two mutually exclusive subsets of channels, based on the dominant pitch estimate. This is achieved by assigning each channel containing a peak at the pitch period estimate to one subset, and all other channels to the other subset.
5. Formation of two "partial" SACF' s by summation of each "partial" correlogram. These are then truncated to "timbre regions", defined as the portion of each SACF between 0.1 and 4.5 ms.
6. Vowel identification using a template matching procedure (an inverse Euclidian distance measure) applied to these timbre regions. The vowel templates are formed by averaging timbre regions for each vowel at each f0.

A criteria level (80%) is set, such that if this amount or more channels are found to have a peak at the dominant pitch period, it is decided that one pitch is present. In the case of a single pitch being found, the original SACF for all channels is used in the matching procedure. If the best template matching score is more than a certain amount (5x) greater than the second best, it is decided that only one vowel is present.

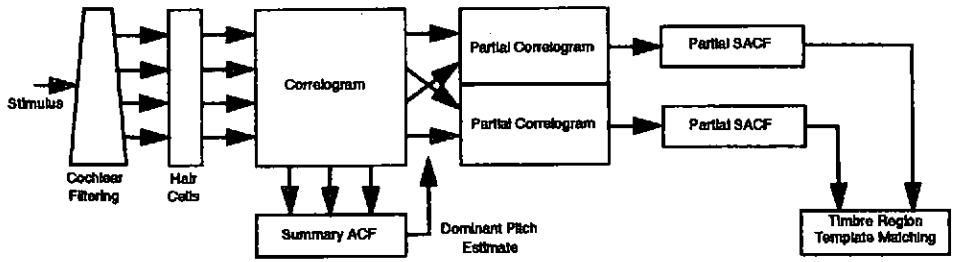

Figure 1: Model Schematic

Note that component vowels are characterised using information from one of two *mutually exclusive* subsets of channels, identified on the basis of a *single* pitch estimate. The model appears to reproduce the human performance observed for this task because the segregation of channels into two sets only becomes possible as the pitches of the two vowels diverge, and the segregation becomes more reliable as the $\Delta f0$ increases. Also note that the timbre region template matching uses *only* periodicity information. The templates and derived timbre regions do not necessarily show any pronounced peaks at periods corresponding to formants.

3 EXPERIMENT

3.1 Pretest

3.1.1 Stimuli. The purpose of the pretest was to verify that subjects could identify the individual synthetic vowels used in the main experiment at a sufficiently high level of accuracy. The vowels used were the same as those used in the double-vowel identification experiment conducted by Assmann and Summerfield [3], and in the original modelling study of Meddis and Hewitt [11]. They were exemplars of the five English monophthongal vowels: /ɑ/, /ɛ/, /i/, /u/ and /ɔ/, with fundamental frequencies of 100, 101.45, 102.93, 105.95, 112.25 and 125.99 Hz (corresponding to Δf0's of 0, 0.25, 0.5, 1, 2 and 4 semitones). Each vowel was presented at each of the f0's, giving 30 different stimuli. All stimuli were equalised for rms level.

3.1.2 Subjects. Subjects were 1 female and 12 male volunteers. None reported having a hearing disorder. They were either native English speakers (10) or of mixed nationality but fluent in English, which they spoke on a day-to-day basis (3). Most had some experience of listening to synthesised sounds (11), and some had extensive experience with the actual stimuli set (3).

3.1.3 Method. The experiment was run by a MATLAB program. Stimuli were presented over headphones. Subjects responded by choosing one of the five response categories. Each stimuli was presented once only, but subjects were allowed to practice listening to the 100 Hz f0 vowels. They were informed that they needed to attain a criteria performance level before progressing to the main experiment. The entire stimulus set was presented five times with a different random order each time, for a total of 150 trials. No feedback was given.

3.1.4 Results. Criteria performance level was 90%. Only 7 of the subjects achieved this. This is an interesting result in itself and reflects the "unnaturalness" of these synthetic stimuli. Four of the subjects who failed the pretest were native English speakers. The two who scored 100% were the two with the most experience listening to this stimulus set. It would appear that training plays a significant part in the identification of these stimuli.

3.2 Main Experiment

3.2.1 Stimuli. Double-vowels were created by adding pairs of single vowels and equalising for rms level. In each pair, one vowel always had an f0 of 100 Hz, otherwise every combination was generated, including same-vowel pairs and reverse f0 order pairs (i.e. /ɑ/ at 0 Δf0 + /ɛ/ at 0.5 semitones Δf0, and /ɑ/ at 0.5 semitones Δf0 + /ɛ/ at 0 Δf0), giving a total of 150 different stimuli.

3.2.2 Subjects. The subjects were the 7 who passed the pretest. All were male; 6 were native English speakers.

3.2.3 Method. The same program was used to run the experiment, but now the subjects were required to make two responses. Stimuli were presented once only and no feedback was given. More practice was allowed between pretest and the main experiment. Subjects were informed that they could choose the same vowel twice. The entire stimulus set was presented four times with a different random order each time, for a total of 600 trials.

4 RESULTS

4.1 Identification Performance

The overall performance of both subjects and model is plotted in fig. 2. As can be seen, the commonly reported finding is replicated. There are some differences in detail compared to the Assmann and Summerfield [3], and the Meddis and Hewitt [11] data however. Average human performance is lower here than reported by Assmann and Summerfield. This can most likely be attributed to degree of training of the subjects. Also, the Meddis and Hewitt modelling study under-predicted performance at zero Df0; these results are a closer match. This can be attributed to fine tuning of the single/double vowel decision criteria.

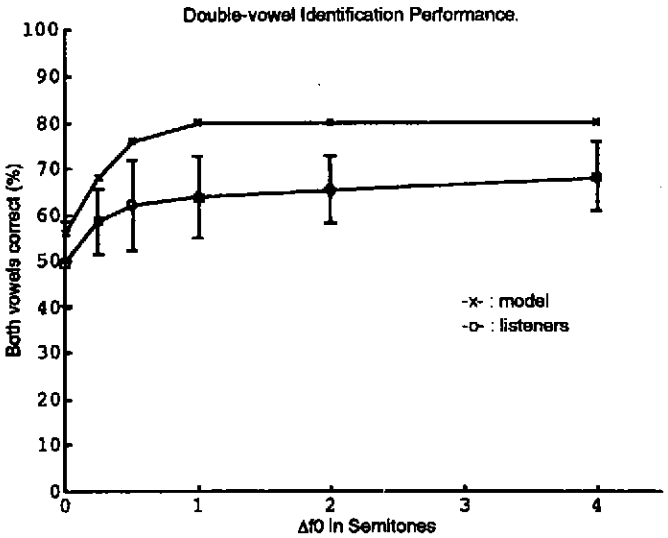

Figure 2: Double vowel identification performance for the computer model and pooled subject data. Error bars indicate one standard error.

4.2 Confusion Data

The confusion data is shown in fig. 3. Fig 3a shows the subject data and fig 3b shows the model data. Actual identity of the vowel-pair is indicated down the left side, response category across the top. Scores are in percentage of total observations. The confusion data for subjects and the model match qualitatively fairly well overall. There are some significant discrepancies however. Firstly, performance of the model for same-vowel pairs is consistently lower than human performance. Indeed, the model seems to find these stimuli the most difficult to correctly identify, whereas humans find them amongst the easiest to identify. The largest difference between human and model responses is a case of this and is highlighted in the figure. This is not surprising however, considering how the model works. In the case of a 0 Δf0 stimulus an arbitrary decision criteria is employed in the model to distinguish same-vowel and different-vowel stimuli in $\Delta f0 = 0$ conditions. The behaviour of the model can be quite significantly influenced at 0 Δf0 by altering this parameter. Also, in $\Delta f0 > 0$ conditions, two instances of the same vowel will have similarly shaped excitation patterns, so that the channels they maximally excite will be similar. Therefore, it is likely that when the channels associated with the dominant pitch are grouped together, little will be left of importance for identifying the second vowel if it is the same as the first. Secondly, and more crucially, there are some specific confusions in the human data which are not well reproduced by the model. The most significant of

these differences are highlighted in the figure. Reasons for these discrepancies could be sought in terms of the spectral composition of the double-vowels which is not well represented in the SACF timbre region.

	/ɑ/+/ɑ/	/ɑ/+/i/	/ɑ/+/ɛ/	/ɑ/+/u/	/ɑ/+/ɔ/	/i/+/i/	/i/+/ɛ/	/i/+/u/	/i/+/ɔ/	/ɛ/+/ɛ/	/ɛ/+/u/	/ɛ/+/ɔ/	/u/+/u/	/u/+/ɔ/	/ɔ/+/ɔ/
/ɑ/+/ɑ/	82	2	7	2	6	0	0	0	0	0	0	0	0	1	1
/ɑ/+/i/	4	92	1	0	1	0	1	0	1	0	0	0	0	0	0
/ɑ/+/ɛ/	14	2	30	4	43	0	0	0	0	0	0	2	0	1	2
/ɑ/+/u/	18	27	8	38	5	0	0	0	1	0	1	0	0	0	0
/ɑ/+/ɔ/	13	7	9	32	36	0	0	0	1	0	1	1	0	0	0
/i/+/i/	0	0	0	0	0	97	0	3	0	0	0	0	0	0	0
/i/+/ɛ/	0	0	1	0	0	0	91	0	0	6	1	1	0	0	0
/i/+/u/	0	1	0	0	0	7	3	84	4	0	1	0	0	0	0
/i/+/ɔ/	0	1	0	0	0	1	3	17	69	0	1	4	0	1	2
/ɛ/+/ɛ/	0	0	1	0	0	1	1	0	0	81	4	11	0	0	3
/ɛ/+/u/	0	0	4	0	1	0	2	1	1	27	24	37	0	1	3
/ɛ/+/ɔ/	0	0	3	0	3	0	3	0	1	9	10	55	0	5	11
/u/+/u/	0	0	0	1	0	0	0	0	0	0	1	0	97	1	0
/u/+/ɔ/	0	0	0	1	0	0	0	3	1	0	7	2	20	43	23
/ɔ/+/ɔ/	0	0	0	0	1	0	0	1	1	0	1	0	20	25	52

Figure 3a: Confusion matrix for listeners, pooled across pitch conditions, presentations and subjects.

	/ɑ/+/ɑ/	/ɑ/+/i/	/ɑ/+/ɛ/	/ɑ/+/u/	/ɑ/+/ɔ/	/i/+/i/	/i/+/ɛ/	/i/+/u/	/i/+/ɔ/	/ɛ/+/ɛ/	/ɛ/+/u/	/ɛ/+/ɔ/	/u/+/u/	/u/+/ɔ/	/ɔ/+/ɔ/
/ɑ/+/ɑ/	50	17	17	0	17	0	0	0	0	0	0	0	0	0	0
/ɑ/+/i/	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
/ɑ/+/ɛ/	8	0	75	0	0	0	8	0	0	0	8	0	0	0	0
/ɑ/+/u/	25	0	0	75	0	0	0	0	0	0	0	0	0	0	0
/ɑ/+/ɔ/	0	0	17	8	75	0	0	0	0	0	0	0	0	0	0
/i/+/i/	0	0	0	0	0	50	0	50	0	0	0	0	0	0	0
/i/+/ɛ/	0	0	0	0	0	0	92	0	0	0	8	0	0	0	0
/i/+/u/	0	0	0	0	0	25	0	67	8	0	0	0	0	0	0
/i/+/ɔ/	0	0	0	0	0	0	0	8	92	0	0	0	0	0	0
/ɛ/+/ɛ/	0	0	0	0	0	0	50	0	0	33	0	17	0	0	0
/ɛ/+/u/	0	0	8	0	0	0	0	0	0	17	67	8	0	0	0
/ɛ/+/ɔ/	0	0	0	0	0	0	0	0	0	17	8	75	0	0	0
/u/+/u/	0	0	0	17	0	0	0	0	0	0	0	17	17	50	0
/u/+/ɔ/	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
/ɔ/+/ɔ/	0	0	0	0	17	0	0	0	0	0	0	33	0	0	50

Figure 3b: Confusion matrix for model, pooled across pitch conditions

4.3 Partitioning the Data by Δf0 Condition and by Vowel-Pair

A problem arises when attempting to partition the data in order to make more detailed comparisons. For instance, the pattern of confusions for listeners varies with Δf0 condition. In some cases, not only does correct identification improve with increasing Δf0, but the most common confusion shifts from one category to another. It would be interesting to make comparisons with the model in this respect but the categorical nature of the output of the model means insufficient data exists to make such comparisons meaningful.

The data was also partitioned by each individual vowel-pair. Differences between vowel-pairs are evident in

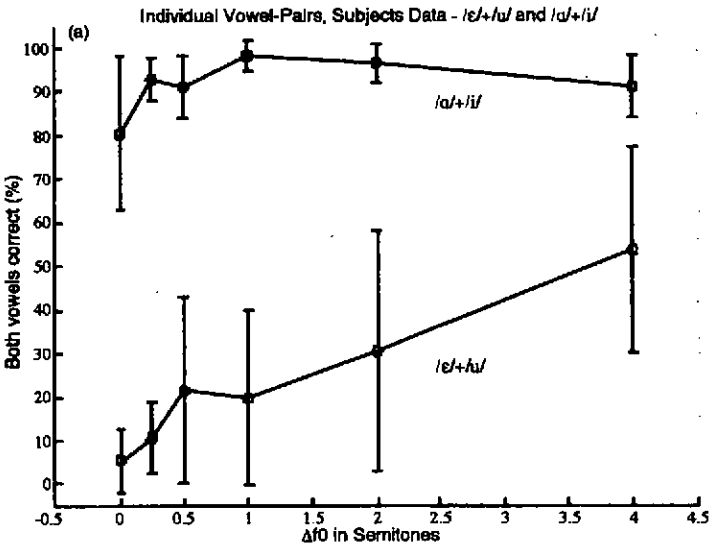

Figure 4: Identification performance for individual vowel pairs, showing floor and ceiling effects for subjects. Error bars indicate one standard error

the subject data and some of this data is shown in fig. 4. Floor and ceiling effects can be seen to be playing a part in the form of the overall results. Solutions to both ceiling effects and the lack of model data are discussed in the next section.

5 DISCUSSION

5.1 Spectral and Timbre Region Representations.

A possible explanation of the particular confusions made by listeners can be formulated in terms of the formant relationships among the vowels. Consider the f_1 frequencies of the vowels: /ɑ/ = 650Hz, /ɛ/ = 450Hz, /ɔ/ = 350Hz, /i/ = 250Hz, /u/ = 250Hz. Referring to fig. 3a; in the context of /ɑ/: /ɛ/ is frequently identified as /ɔ/, /ɔ/ is frequently identified as /u/, and /u/ is frequently identified as /i/. This corresponds to shifting the f_1 down by 100Hz in the first two cases, and, although there is no vowel with an f_1 lower than 250Hz to be reported in the third case, the incorrect vowel that is most commonly reported (/i/) has a much higher f_2, corresponding to an increase in f_2/f_1 ratio. It would be useful to apply a statistical test to establish whether the f_n relationships among the vowels predicts the pattern of confusions. This will be the first step in extending this study. An explanation for the different confusion pattern of the Meddis and Hewitt model may lie in the matching procedure. The hypothesis is that the lack of explicit formant frequency information in the timbre region is responsible. This could be tested empirically by comparison with a version of the model using a different "spectral" matching procedure applied to spectra constructed from the partial correlograms, using a technique which placed emphasis on spectral peaks and shoulders, such as WN2DM [2]. If the resulting confusion matrix was a closer match to the human data this would be strong evidence for the hypothesis. A difficulty here however, is the fact that such spectra will contain large gaps. Spectral templates with corresponding gaps could be used but such partial data may produce poor results. However, it may be possible to exploit work on "missing data" techniques (e.g. [13]).

5.2 Lack of Model Data

A major problem at present is a lack of sufficient model data to allow comparisons involving partitioning the data in any way. A solution to this would be to generate a larger number of observations using a stimuli set consisting of a number of allophones of each vowel. This would enable us to generate confusion data for the model at different Δf0's, allowing comparison with the human data, and also to generate more meaningful individual vowel-pair data for the model.

5.3 Ceiling Effects

The ceiling effects observed in the subject data is caused by differential masking of vowels by other vowels. i.e when two vowels have similar formant frequencies, corresponding harmonics in the (say) f_1 region which are not resolved will be of similar magnitude and resulting double-vowel spectra may contain peaks which do not correspond to harmonics or formants of either vowel, whereas two vowels with well separated formants will be well represented in the composite spectra and hence more easily identified. A procedure for dealing with ceiling effects in double-vowel identification tasks has been developed by de Cheveigne [8]. This involves introducing a systematic amplitude imbalance to reduce ceiling performance for the non-dominant vowel. Various studies have shown that one vowel is perceived as dominant and virtually always correctly identified. It is therefore the other vowel which gives rise to the improvement in performance [9], but if this non-dominant vowel is poorly masked by the dominant vowel and hence well identified even at 0 Δf0, the effect of increasing Δf0 will be obscured. Reducing the amplitude of the dominant vowel in a pair will therefore maximise the sensitivity to changes in f0. We intend to incorporate this method in future studies.

5.4 Same-Vowel Stimuli and 0 Δf0 Conditions

The reasons behind the differing behaviour of model and listeners for same-vowel pairs are somewhat distinct from the considerations discussed in section 5.1. A related point is that the model behaves differently at 0 Δf0 because it uses a single complete correlogram for matching rather than two partial ones. These factors are incidental to the issues we wish to investigate, and we intend to exclude same-vowel stimuli and 0 Δf0 conditions from future experiments. Phonetically identical vowel pairs and exactly coincident fundamental frequencies are somewhat unlikely to occur in real acoustic environments in any case.

5.5 Beating Effects

A complication of the issue of double-vowel identification is that, with steady-state synthetic vowel pairs, the small constant differences between frequencies of unresolved corresponding harmonics results in beating which causes fluctuations in the spectral envelope of the composite stimulus, particularly in the f_1 region. These fluctuations could be exploited by listeners. This is in contrast with the f0-guided segregation strategy previously assumed to be responsible for the improvement in performance. This effect was successfully modelled by Culling and Darwin [7]. Both the beating effect and the poor performance of many subjects in the pretest, suggest a move towards more natural stimuli. There is much perceptually important information in a natural vowel which is absent in these synthetic vowels, such as the dynamic properties of vowels. Reliable, vowel-specific formant movement has been found, even for monophthongs [14], and this has been shown to be perceptually relevant [1],[15]. We intend to employ LPC analysis and resynthesis to generate double-vowel stimuli, which will reduce the steady-state nature of the stimuli, in that, although monotone fundamentals will produce constant harmonic frequencies, the relative amplitudes of the harmonics will vary, making it more likely that at any one time, a harmonic from one voice will dominate the neighbouring harmonic from the other voice. It may also be possible to generate vowel stimuli with realistic f0 trajectories, either using a constant f0-offset between two vowels with the same f0 trajectory, or using the mean difference in f0 throughout the stimulus duration of two vowels on different

trajectories.

5.6 Other Models

The eventual goal of this investigation is to extend this comparison to other models of double-vowel identification. A similarity measure for two confusion matrices could be calculated using a Euclidian distance metric. This could be used to assess the degree of similarity between human and model confusion patterns, which could then be compared between models to determine which most closely mimic human behaviour. Implications for theories of concurrent sound segregation can then be considered.

6 REFERENCES

[1] P F ASSMANN, T M NEAREY & J T HOGAN 'Vowel identification: Orthographic, perceptual and acoustic aspects', J. Acoust. Soc. Am., 71 p975-989 (1982)

[2] P F ASSMANN & Q SUMMERFIELD 'Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency', J. Acoust. Soc. Am., 85 p327-338 (1989)

[3] P F ASSMANN & Q SUMMERFIELD 'Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies', J. Acoust. Soc. Am., 88 p680-697 (1990)

[4] A S BREGMAN 'Auditory scene analysis', Cambridge, MA: MIT Press, (1990)

[5] G J BROWN &D WANG 'Modelling the perceptual segregation of double vowels with a network of neural oscillators', Neural Networks, 10(9) p1547-1558 (1997)

[6] J P L BROKX & S G NOOTEBOOM 'Intonation and the perceptual separation of simultaneous voices', J. Phon., 10 p23-26 (1982)

[7] J F CULLING & C J DARWIN 'Perceptual and computational separation of simultaneous vowels: cues from low-frequency beating', J. Acoust. Soc. Am., 95 p1559-1569 (1994)

[8] A DE CHEVEIGNE 'Concurrent vowel segregation III: A neural model of harmonic interference cancellation', J. Acoust. Soc. Am., 101 p2857-2865 (1997)

[9] J D MCKEOWN 'Perception of concurrent vowels: The effect of varying their relative level', Speech Comm. 11 p1-13 (1992)

[10] R MEDDIS 'Simulation of mechanical to neural transduction in the auditory receptor', J. Acoust. Soc. Am., 79 p702-711 (1986)

[11] R MEDDIS & M J HEWITT 'Modelling the identification of concurrent vowels with different fundamental frequencies', J. Acoust. Soc. Am., 91 p233-245 (1992)

[12] G F MEYER & 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, p212-215 (1996)

[13] A MORRIS, M COOKE & P GREEN 'Some solutions to the missing feature problem in data classification, with application to noise-robust ASR', Proc. ICASSP, (ii) p737-740 (1998)

[14] T M NEAREY 'Static, dynamic, and relational factors in vowel perception', J. Acoust. Soc. Am., 85 p2088-2113 (1989)

[15] T M NEAREY & P F ASSMANN 'Modelling the role of inherent spectral change in vowel identification', J. Acoust. Soc. Am., 80 p1297-1308 (1986)

[16] M T M SCEFFERS 'Sifting vowels: Auditory pitch analysis and sound segregation', Ph.D. Thesis, Rijksuniversiteit te Groningen, (1983)