Proceedings of The Institute of Acoustics

MODELLING THE PERCEPTION OF CONCURRENT VOWELS

Peter Assmann and Quentin Summerfield

MRC Institute of Hearing Research, University Park, Nottingham

INTRODUCTION

The effects of noise on speech communication have been understood in broad terms for some time (10). However, only relatively recently have speech researchers begun to address the question of how listeners extract individual acoustic cues to phonetic categories from speech signals corrupted by noise (6), and how that process can be modelled to be independent of the spectro-temporal structure of the noise (17). A particular problem, both for human speech perception and for automatic speech recognition, arises when the background noise is that of a competing voice. Three cues for segregating a single voice from a background of other voices have received attention: binaural information (9), fundamental frequency differences (3,15), and onset/offset asynchronies (6,16). In this paper we investigate a restricted case of the multi-source recognition problem, where none of the above cues are available: that of two simultaneous, synthetic vowels presented monaurally with the same, fixed fundamental frequency, and their pitch pulses in synchrony.

Scheffers (15) reported that listeners could identify both members of such a pair of vowels with better-than-chance accuracy. This rather surprising finding was confirmed in a later study by Zwicker (18) and was replicated in the data presented here. Scheffers suggested that the perception of simultaneous vowel sounds might be modelled as a template-matching process. The spectrum of an incoming vowel pair is compared with a set of stored reference profiles, and the two best-matching profiles determine the responses made to the presented vowels. Scheffers adopted a formant representation as the basis for his classification algorithm. In his model, formant peaks were extracted from auditory excitation patterns, and classification was based on a comparison of the frequencies of detected formant peaks with those of a set of stored reference profiles. Here we compare a formant-based representation of the perceptually salient information in vowel sounds with several alternative spectral representations (1,8,14), and evaluate their performance in predicting identification profiles for paired vowels.

THE EXPERIMENT

The data were obtained from a larger experiment investigating the 'enhancement effect' of spectral amplitude increments on the identification of simultaneous vowels (16). Listeners identified paired vowels in several precursor conditions and a control condition with no precursor. The data described below were from the control condition.

Stimuli were produced digitally (10,000 samples/sec, 12-bit amplitude quantisation) by summing the waveforms of isolated synthetic vowels from the set / i a u ɔ ɜ /. A total of 25 pairs was constructed, representing all possible combinations of these five vowels. Formant frequencies were based on estimates from natural tokens from a speaker of British English. The fundamental frequency was constant at 100 Hz. Two sets of vowels were produced using different synthesis procedures. In the first set, the vowels were produced by means of cascade formant synthesis [7]. We will refer to stimuli from this set as 'Klatt vowels'. In the second set, pairs of adjacent harmonics of a 100 Hz fundamental frequency were synthesised at frequencies straddling the centre frequencies of the lowest three formants. Harmonics were of equal amplitude, with random starting phases. Stimuli from this set will be referred to as '6-harmonic vowels'. Line spectra illustrating the two stimulus sets for the vowel /i/ are shown in Figure 1.

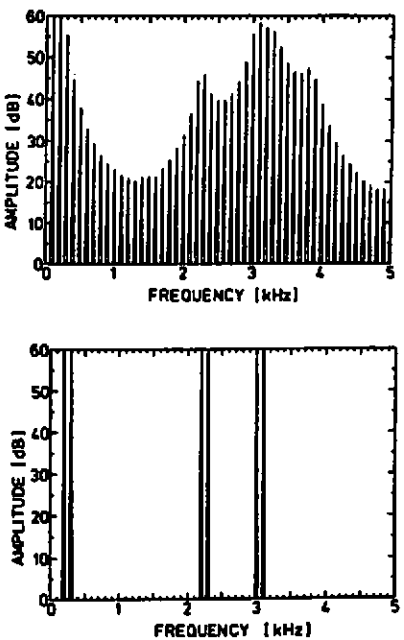

Figure 1. Line spectra of vowel /i/.
(a) 'Klatt vowels'
(b) '6-harmonic vowels'

Vowel onsets and offsets were shaped by a 10.7 ms Kaiser function. Vowels were 188.6 ms in duration between the -6 dB points. Stimuli were presented on-line by means of a DEC PDP-11/60 and LPA-11K, low-pass filtered at 4.25 kHz (KEMO VBF/8, -135 dB/octave), and presented to listeners over the left channel of a set of Sennheiser HD-414 headphones. Six listeners with normal hearing participated in the experiment. They were tested individually in a sound-attenuated room, instructed to give two responses to each stimulus, and responded by pressing VDU keys labelled with the orthographic representations of each of the vowels: /i/:EE, /a/:AH, /u/:OO, /ɔ/:OR, /ɜ/:ER.

Pooled identification rates are shown in Table 1. No significant differences were found between the two stimulus types, suggesting that formant amplitude and overall spectral shape may be relatively unimportant in the identification of paired vowels with the same fundamental frequency, at least for the vowel set investigated here.

Table 1. Identification rates (% correct) for single vowels and for paired vowels (both identified correctly). Standard deviations (6 listeners) shown in brackets.

	Klatt vowels	6-harmonic vowels
single vowels:	99.0 (1.1)	98.3 (3.2)
paired vowels:	49.2 (9.5)	54.2 (16.6)

SPECTRAL INFORMATION AND THE IDENTIFICATION OF PAIRED VOWELS

Listeners typically reported hearing the paired vowels as stemming from a single sound source. They were aware of two distinct 'voices' only in the case of /i/ paired with /a/, where both vowels could be 'heard out' and were identified with a high degree of accuracy (better than 80%). For the remaining pairs, listeners reported hearing only a single vowel quality, 'coloured' by another. As reported by Scheffers [15] and Zwicker [18], one member of the pair was dominant, and listeners felt that they frequently had to guess at the identity of the second vowel. Since the paired vowels were generally heard as stemming from a single sound source, no attempt was made explicitly to separate the composite vowel into two components in modelling the results.

As a first step toward predicting the identification responses to paired vowels, we computed measures of the degree of dissimilarity or 'distance' between each paired vowel and a set of reference patterns representing the response alternatives. The resulting distance vectors were compared with observed response vectors. The response vector for each paired vowel was defined as the proportion of responses assigned to each category, pooled across trials, listeners and the two responses given on each trial. No attempt was made to incorporate a classification rule to assign the paired vowels to response categories on the basis of their distances. It was assumed that the probability of selecting a particular response alternative for a paired vowel depends only on the degree of similarity to the reference pattern, and not on the response to the other member of the pair, as it might if a process of partitioning the composite spectrum was involved.

It is generally considered that information specifying the identity of vowels is encoded in the short term amplitude spectrum. To represent this information, we computed auditory excitation patterns according to the model of Moore and Glasberg [11]. Four spectral distance metrics were applied to the excitation patterns to determine the degree of similarity of paired vowels and single vowel prototypes. Three of these metrics were modified versions of Klatt's weighted spectral slope metric [8,13]. The fourth was a formant peak frequency metric (PEAK) similar to the model proposed by Scheffers [15].

Klatt's weighted spectral slope metric (WSM)

Klatt [8] proposed a weighted spectral slope metric (WSM) to account for perceptual judgements of phonetic quality involving pairs of isolated vowels. Phonetic dissimilarity judgements [4,5,8] were found to be greatly influenced by changes in formant peak frequencies, but were little affected by changes in overall spectral tilt, formant amplitude, or formant bandwidth. WSM compares pairs of spectra, or auditory excitation patterns, in terms of differences in amplitude level between adjacent spectral channels. This metric highlights local contrasts in spectral amplitude, but is relatively insensitive to changes in overall spectral balance. A set of weighting coefficients adjusts the contribution of a given frequency channel depending on: (a) the output of the channel relative to the global maximum in the spectrum (k_{GMAX}); (b) the output of the channel relative to the nearest local spectral maximum (k_{LMAX}); (c) overall level differences between the two spectra (k_E). The values of these coefficients are selected to optimise prediction performance. The distance according to WSM between two spectra, S_1 and S_2 is given by:

$$d_{WSM} = k_E [E_{S1}-E_{S2}] + \sum_{1}^{Q} k_1(i) [S_1'(i)-S_2'(i)]^2 \qquad (1)$$

where E_{S1} and E_{S2} are the overall energy levels of S_1 and S_2;

S_1' and S_2' are their spectral slopes, computed as the first difference:

$S_1'(i) = S_1(i) - S_1(i+1)$ and $S_2'(i) = S_2(i) - S_2(i+1)$;

$k_s(i) = k_{LMAX} / [k_{LMAX} + D_{LMAX}(i)] [k_{GMAX} + D_{GMAX}(i)]$;

$D_{GMAX}(i)$ is the dB difference between the level of the ith spectral channel and the global maximum of the spectrum; $D_{LMAX}(i)$ is the dB difference between the ith channel and the nearest local maximum of the spectrum; and Q is the number of spectral channels.

Weighted level metric (WLM)

Two modified versions of the WSM were also investigated. The first was a weighted level metric (WLM) which replaced the spectral slopes, S_1' and S_2', with spectral amplitude levels, S_1 and S_2. Metrics similar to the unweighted form of WLM have been used to predict vowel similarity judgements [2,14]. Although Klatt [8] reported that these metrics appeared to be overly sensitive to spectral tilt and formant amplitude manipulations, we felt that comparisons of WLM and WSM might help to establish whether a slope-based representation provides a better characterisation than the level-based representation of the identification of paired vowels.

Negative portion of the second differential (N2D) metric

The weighted N2D metric (WN2DM) replaced the spectral slopes $S_1'(i)$ and $S_2'(i)$ with $S_1''(i)$ and $S_2''(i)$, the absolute values of the negative portion of the second differential of the spectrum, setting the positive portion to zero. S_1'' was computed as follows:

$$S_1''(i) = MAX [- S_1(i-1) - 2S_1(i) + S_1(i+1) , 0] \qquad (2)$$

This metric is not sensitive to differences in overall spectral tilt, assigns greatest weight to differences in peak locations, but also emphasises spectral regions involving 'shoulders' and poorly resolved peaks. The detection of shoulders is likely to be of particular importance for the perception of paired vowels, where the formants of different vowels can merge to form a single peak.

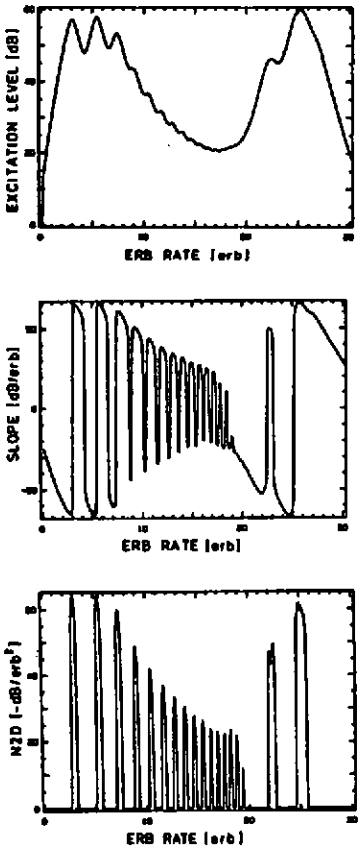

Figure 2 illustrates the spectral representations which underly each of the three metrics. In the top panel, excitation level (as used by WLM) is plotted as a function of erb-rate [11]. This profile shows a series of peaks in the low frequency region of the excitation pattern, representing prominent harmonics near the first formant. Peaks in the high frequency region reflect the presence of higher formants; individual harmonics are not resolved. In the spectral slope representation used by WSM (middle panel), zero crossings correspond to maxima and minima of the excitation pattern. In the N2D representation (bottom panel) peaks correspond to either peaks or shoulders in the excitation pattern. Since only the negative portion of the second differential is used, the resulting profile appears as a 'sharpened' version of the excitation pattern.

Figure 2.
Spectral representations used by (a) WLM (b) WSM (c) WN2DM for the vowel /i/ (cascade synthesis)

Formant peak metric

The PEAK metric is similar to Scheffers' model [15] which represents vowel quality in terms of the estimated frequency locations of formant peaks in the spectrum envelope. The frequency locations of the lowest two or three formants are generally considered to be the primary determinants of vowel quality. In the low frequency region of the excitation pattern, a formant may be defined by several peaks corresponding to strong harmonics, while in the high frequency region, several harmonics may combine to form a single peak.

Following Scheffers, we sampled the excitation pattern at integer multiples of the fundamental frequency. This step was necessary to avoid confusion between harmonic peaks and formant peaks. Six dB/octave preemphasis was applied to the spectrum to obtain better formant frequency estimates. The frequencies of all peaks and shoulders in the sampled excitation pattern were estimated by a multiple differentiation technique.

To establish a set of reference patterns with which to compare the peaks found in the excitation patterns of the paired vowels, estimates of the frequencies of F1, F2 and F3 were obtained, by the same technique, from excitation patterns of the isolated vowels, / i a u ɔ ɜ /. These estimates were in all cases very similar to the synthesised formant frequencies. For each formant peak of the reference pattern, the distance to the nearest spectral peak or shoulder in the paired vowel was computed. The formant peak distance to the ith response vowel was computed as:

$$d_{PEAK} = W_j \ln [F_{ij} - F_n]^2 \qquad (3)$$

where F_{ij} is the estimated frequency of the jth formant of the reference pattern for the ith vowel, F_n is the frequency of the nearest estimated peak or shoulder in the paired vowel, W_j is a weighting coefficient reflecting the relative contribution of the jth formant, selected to optimise the performance of the metric.

COMPARISON OF DISTANCE PROFILES WITH IDENTIFICATION RESPONSES

For each paired vowel, a response vector was computed as described above. Separate analyses were carried out for the Klatt vowels and the 6-harmonic vowels. Distances between the paired vowel spectrum and each of the corresponding reference patterns were computed for each metric, as described above, for a total of 125 distances (25 vowels by 5 response alternatives).

For each of the four metrics and three stimulus types, a correlation coefficient was computed between the distance profiles and response profiles (Table 2). 'Unweighted' analyses were carried out for models WLM, WSM, and WN2DM by setting k_E to 0, and k_{LMAX} and k_{GMAX} to 1×10^6. For the PEAK model, W_1, W_2 and W_3 were set equal to 1. A general function minimisation routine [12] was used to find values of the weighting coefficients which optimised correlations with the perceptual data.

For each stimulus type, the largest correlations were obtained using models WN2DM and PEAK. Both models resulted in highly significant correlations. It can be seen that optimisation of the weights for the WN2DM and PEAK models resulted only in small increases in the correlation coefficients, while more substantial increases were observed for the WLM and WSM models. Since the optimisation was carried out independently for each vowel set, it is significant that unweighted versions of the WN2DM and PEAK metrics perform nearly as well as the optimised versions. It would appear that these metrics give superior results because they highlight spectral peaks and shoulders. Together with the finding that the performance of listeners was very similar for two synthesis types, these findings confirm that spectral peaks and shoulders are important in the perception of concurrent vowels.

Table 2. Pearson product moment correlations (d.f.=123) between spectral distance vectors and response vectors from listeners.
Set 1: Klatt vowels Set 2: 6-harmonic vowels

	unweighted metrics		'optimised' metrics	
	Set 1	Set 2	Set 1	Set 2
WLM	.56	.73	.68	.75
WSM	.79	.76	.85	.81
WN2DM	.88	.83	.88	.85
PEAK	.88	.83	.89	.85

These high correlations were found without partitioning the composite spectrum. Such a strategy is appropriate when there are no cues to source segregation, such as a difference in fundamental frequency between the vowels, and the vowels are generally perceived as coming from a single source. However, when paired vowels have different fundamental frequencies, two sources may be heard and accuracy of identification improves [15]. Thus, to model the general case, it may be necessary to include a 'perceptual grouping' stage in which the composite spectrum is partitioned. Scheffers suggested that listeners might be able to partition the excitation pattern of a paired vowel on the basis of spectral fine structure. In his model, two fundamental frequencies are estimated from the peaks in the excitation pattern, which is then partitioned by sampling through a pair of 'harmonic sieves'. However, this approach did not succeed in modelling the improvement in performance with increasing difference in fundamental frequency, possibly because the strategy can only crudely segregate the higher-frequency part of the excitation pattern where individual harmonics are not resolved. An alternative strategy that might avoid this limitation would be to group and segregate the outputs of an array of auditory filters on the basis of common periodicity in their temporal fine structure [17]. Our present efforts are directed at incorporating spectral distance metrics of the sort discussed in this paper into models of the perception of paired vowels differing in fundamental frequency and degree of pitch-pulse asynchrony using time-domain and frequency-domain grouping strategies.

REFERENCES

[1] Assmann, P.F. (1985). "The role of harmonics and formants in the perception of vowel quality," Unpublished Ph.D. thesis, University of Alberta.

[2] Bladon, R.A.W. and Lindblom, B. (1981). "Modelling the judgement of vowel quality differences," J. Acoust. Soc. Am. 69: 1414-1422.

[3] Brokx, J.P.L. and Nooteboom, S.G. (1982). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics 10: 23-26.

[4] Carlson, R. and Granstrom, B. (1979). "Model predictions of vowel dissimilarity," Speech Transmission Laboratory Quarterly Progress Report STL-QPSR 3-4/1979: 84-104, Royal Institute of Technology, Stockholm, Sweden.

[5] Carlson, R., Granstrom, B., and Klatt, D. (1979). "Vowel perception: The relative perceptual salience of selected acoustic manipulations," Speech Transmission Laboratory Quarterly Progress Report STL-QPSR 3-4/1979: 73-83, Royal Institute of Technology, Stockholm, Sweden.

[6] Darwin, C.J. (1983). "Auditory processing and speech perception," In H. Bouma and D.G. Bouwhuis (eds.) Attention and Performance, Hillsdale, N.J.

[7] Klatt, D.H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67: 971-995.

[8] Klatt, D.H. (1982). "Prediction of perceived phonetic distance from critical-band spectra: A first step," Proc. ICASSP 82: 1278-1281.

[9] Licklider, J.C.R. (1948). "The influence of interaural phase relations upon the masking of speech by white noise," J. Acoust. Soc. Am. 20: 150-159.

[10] Miller, G.A., Heise, G.A. and Lichten, W. (1951). "Intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. 41: 329-335.

[11] Moore, B.C.J. and Glasberg, B.R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. 74: 750-753.

[12] Nelder, J.A. and Mead, R. (1965). "A simplex method for function minimization," Computer J. 7: 308-313.

[13] Nocerino, N., Soong, F.K., Rabiner, L.R. and Klatt, D.H. (1985). "Comparative study of several distortion measures for speech recognition," Speech Communication 4: 317-331.

[14] Plomp, R. (1970). "Timbre as a multidimensional attribute of complex tones," In: Plomp, R. and Smoorenburg, G.F. (eds.) Frequency analysis and periodicity detection in hearing. (A.W. Sijthoff: Leiden).

[15] Scheffers, M.T.M. (1983). "Sifting vowels: Auditory pitch analysis and sound segregation," Unpublished Ph.D. thesis, University of Groningen.

[16] Summerfield, A.Q. and Assmann, P.F. (1986). "Auditory enhancement and speech perception," NATO Advanced Research Workshop on the Psychophysics of Speech Perception, Utrecht, June 30-July 4, 1986.

[17] Weintraub, M. (1985). "A theory and computational model of monaural auditory sound separation," Unpublished Ph.D. thesis, Stanford University.

[18] Zwicker, U.T. (1984). "Auditory recognition of diotic and dichotic vowel pairs," Speech Communication 3: 265-277.