Proceedings of the Institute of Acoustics

MINIMAL SPECTRAL CONTRAST FOR VOWEL RECOGNITION AS A FUNCTION OF SPECTRAL SLOPE

Andrew Lea and Quentin SUMMERFIELD

MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, UK.

1. INTRODUCTION

When two talkers speak concurrently, a formant in one voice may be partially masked by a more intense formant in the competing voice. The partially masked formant may be displayed only as an irregularity on one of the skirts of the more intense formant in the internal representation of the spectrum. Assmann and Summerfield (A&S) [1] called such irregularities "shoulders" and offered a formal definition which is included in Section 4.1, below. A&S found that they could improve the accuracy of predictions of listeners' identification responses, to pairs of concurrently spoken vowels, by estimating the frequencies of formants from shoulders as well as peaks. It is possible, therefore, that listeners also use shoulders to locate formants.

However, A&S's result can be explained in another way. Listeners may use only peaks, not shoulders, to locate formants, but may improve their performance by deploying a knowledge of other aspects of the spectral shape of particular vowels. A&S's model of vowel identification did not possess this additional knowledge. However, it may have compensated by using shoulders as well as peaks to locate formants.

Indirect evidence that formants can be specified by shoulders was provided by Chistovich and Lublinskaya [2]. They determined the position of the phoneme boundary between a central vowel and a back vowel by altering the intensity of F1 in relation to a fixed F2. In some conditions, F1 was represented by a shoulder rather than a peak in the boundary stimulus.

The aim of the experiments described in this paper was to provide a direct test of the hypothesis that a formant can be specified by a shoulder in the internal representation of a vowel.

2. METHODS

2.1 Stimuli

Stimuli consisted of two types of sound, "maskers" and "target vowels". Both were generated digitally by additive harmonic synthesis (10,000 samples/s, 12-bit amplitude quantisation). Each was 400ms in duration. Onsets and offsets were shaped by the halves of a 10-ms Hanning window.

Each of the 5 maskers consisted of the first 50 harmonics of 100Hz. The same randomly generated phase spectrum was used for each masker. The amplitudes of the harmonics were chosen such that the excitation patterns [3, 4] of the maskers had slopes of -1, -0.5, 0, +1, and +2 dB/erb, pivoted about the 1000-Hz harmonic. Fig. 1 shows the excitation patterns of the five maskers. In their Fourier spectra, the maskers had slopes ranging from approximately -5dB/octave to +10dB/ocatave. These values fall within the range from -6 to +12dB/octave over which Dijkhuizen et al [5] found little variation in speech-reception thresholds for sentences in noise, implying that normal processes of speech perception are possible.

The target vowels were steady-state approximations to five monophthongal vowels of British English, /i/, /a/, /u/, /ɒ/ and /ɜ/. Each consisted of 6 harmonics of 100Hz, chosen to straddle the frequencies of the first three formants listed in Table 1. Five sets of targets were synthesized, one for each masker. The targets in each set were created by retaining 6 harmonics from the corresponding masker, and setting the amplitudes of the remaining harmonics to zero.

MINIMAL SPECTRAL CONTRAST FOR VOWEL RECOGNITION AS A FUNCTION OF SPECTRAL SLOPE

2.2 Procedure

If a target vowel is added to its parent masker without attenuation, the levels of the 6 harmonics are raised by 6dB relative to the remaining harmonics. We define the "spectral contrast" of the formants in the composite sound as the difference in level between the 6 harmonics and the remaining harmonics. Thus, the spectral contrast in this case is 6dB. It is illustrated in Fig. 2A which shows the physical spectrum of the sound that results from adding /ɜ/ (+1dB/erb, 0dB/octave) to its parent masker without attenuation. The aim of the psychophysical procedure was to adjust the intensity of the targets so as to determine the minimum spectral contrast required for each vowel to be identifiably different from the other four vowels.

	Slope				
Formant	/i/	/a/	/u/	/ɔ/	/ɜ/
F1	250	650	250	350	450
F2	2250	950	850	750	1250
F3	3050	2950	1950	2850	2650

Table I: Formant frequencies in Hz of the five vowels. The formant values were chosen to fall between harmonics in a 100Hz harmonic series.

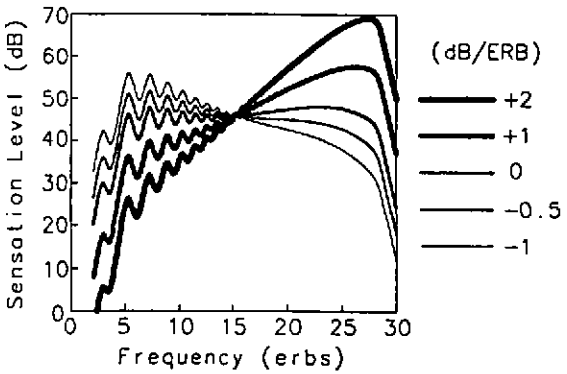

Fig. 1: Excitation patterns, generated by analysing the waveforms of the 5 maskers with a bank of linear overlapping bandpass "auditory" filters. The gain of the filters reflected the normal variation in absolute sensitivity with frequency. The rms output of the filters has been plotted as a function of their centre frequency. The frequency axis has been scaled in units of the equivalent rectangular bandwidths (erbs) of the filters. One erb corresponds to a distance of approximately 0.85mm along the cochlear partition. The 0 dB level represents the absolute threshold of normal listeners.

The procedure was an adaptive 2-interval, 5-alternative forced-choice task (2I5AFC). A trial consisted of two intervals. A masker was presented in both intervals. In one interval, chosen at random, one of the five targets with the same slope as the masker was also presented. The listener's task was to indicate which interval contained the target and what its identity was. The intensity of the target was adjusted adaptively to estimate the spectral contrast required for 71% correct responses [6]. Five adaptive procedures were run concurrently, one for each vowel. Trials from the five procedures were interleaved randomly. Thus, each run provided separate, but concurrent, estimates of threshold contrast for each vowel. These estimates were averaged to provide an overall threshold from each run. At least four runs were completed by each subject in each condition.

2.3 Presentation Levels

Stimuli were presented on-line (Tandon PC-AT) through a CED1401 interface. They were low-pass filtered at 4250Hz (Kemo VBF8, -135dB/octave), attenuated, mixed, and presented through the left ear-piece of a pair of Sennheiser HD 414X headphones. The 1000-Hz component of all maskers was presented at a fixed level of 46dB(A), giving levels of 52, 50, 51, 58, and 67 dB(A) for the maskers with slopes of -1, -0.5, 0, +1, and +2 dB/erb, respectively.

2.4 Subjects

Subjects were native speakers of British English and included the first author. Their ages ranged from 21 to 26 years. Their audiometric thresholds were within 15dB of the ANSI standard [7] at audiometric frequencies between 0.25 and 8kHz in both ears.

3. EXPERIMENT 1

Vowels are generally identified from the frequencies of the first two formants [e.g. 8]. However, Table I shows that, potentially, the five vowels could be distinguished solely from their different F2s. Experiment 1 was run to ensure that listeners perform the 2I5AFC task phonetically by confirming that they use evidence of both F1 and F2 to identify the target vowels.

3.1 Methods particular to Experiment 1

The masker and targets with slopes of +1dB/erb were used. Four conditions were run. For the 1st condition, a new set of targets was created by boosting by 5dB the levels of the pairs of harmonics defining the F1s of the vowels. For the 2nd condition, the F2s were boosted by 5dB. For the 3rd condition, the F3s were boosted by 5dB. For the 4th condition, the baseline, the original unaltered targets were used. Note that boosting a formant by 5dB does not increase its contrast by this amount after the target vowel has been added to its parent masker. Comparison of Figs. 2A and 2B shows that boosting F2 by 5dB increases the contrast of F2 by only 2.9 dB when the target vowel is added to its parent masker without attenuation. With attenuation, the increase in contrast is reduced.

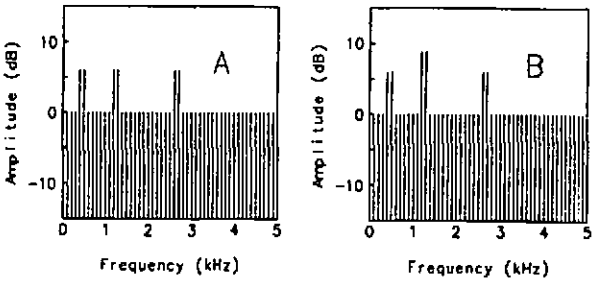

Fig. 2A: Frequency spectrum of a +1 db/ERB masker and an /ɜ/ target mixed without attenuation, giving 6 dB of contrast. B: Spectrum as in A, but with F2 boosted by 5 dB, giving a contrast of 8.9 dB.

The rationale for the experiment is as follows. Suppose that a listener performs the 2I5AFC task by discriminating the frequency of F2, ignoring F1 and F3. In which case, the threshold contrast for F2 will be constant across the four conditions. In particular, threshold contrast will be the same in the condition where F2 is boosted as in the baseline condition. Conversely, if the listener performs the task by discriminating the frequency only of F1, contrast for F2 will be greater in the condition where F2 is boosted than in the baseline condition. Thus, the contribution that a formant makes to the vowel identification task can be estimated by determining how little threshold contrast is increased, relative to the baseline, when that formant is boosted.

3.2 Results of Experiment 1

The mean contrast at threshold in each condition, averaged over four subjects, has been plotted in Fig. 3. Error bars, where they extend beyond the plotting symbols, show plus/minus one standard deviation computed from the within-subjects variance. The left most point shows contrast in the baseline condition where none of the formants was boosted. The lower horizontal line, which intersects this point, indicates the threshold that would occur when a formant is boosted, if that formant alone controlled performance. The upper horizontal line indicates the threshold that would be found if the boosted formant made no contribution. The four mean thresholds differ significantly ($F_{3,9}$=14.8, $p<0.01$) and *post-hoc* tests show the F1 and F2 conditions differ significantly from the other two. The

results are clear. Overall, F1 and F2 control performance, while F3 makes no contribution. This pattern was shown by /a/, /i/, /3/, and /ɔ/ individually. With /u/, all three formants played a role.

3.3 Discussion of Experiment 1

The finding that listeners use both F1 and F2 to perform the vowel-identification task suggests that they performed phonetically, despite the format of the psychophysical procedure and the unusual spectral structure of the stimuli. The outcome is compatible with the experience of listening to the stimuli, namely that they sound like vowels, even when close to threshold.

The baseline threshold contrast of 1.2dB is somewhat lower than estimates of about 2dB obtained with other methodologies [9, 10, 11]. It is similar to the estimate of 1.3dB reported by Henn and Turner [12] as the contrast created by the minimum detectable increment in intensity of the 1-kHz member of a 200-Hz harmonic series. Thus the minimal spectral contrast that defines a formant is close to the minimum that can be detected.

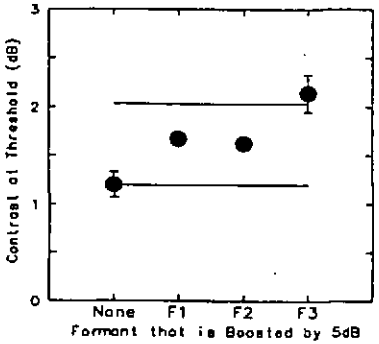

Fig 3: Results of Experiment 1, averaged over subjects and vowels. "None" signifies no formant was boosted.

4. EXPERIMENT 2

Experiment 2 was designed to test the hypothesis that formants can be specified by shoulders in the excitation pattern of a vowel. Thresholds were measured separately with each of the five maskers using target vowels with no formants boosted. The 5 threshold stimuli estimated with each masker were then synthesized, and their excitation patterns were examined to establish whether formants were defined by peaks or shoulders.

4.1 Results of Experiment 2

The mean contrast at threshold, averaged over four subjects, has been plotted in Fig. 4 as a function of the spectral slope of the masker. The five means differ significantly ($F_{4,12}$=9.82, p<0.01). *Post-hoc* tests showed that the threshold at +2dB/erb was significantly higher than the other four thresholds which did not differ from each other.

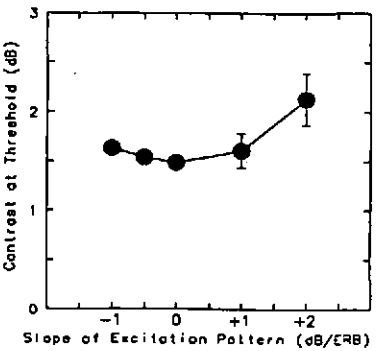

Fig. 4: Results of Experiment 2 averaged over subjects and vowels.

The 25 stimuli corresponding to the subjects' mean thresholds were synthesized. Excitation patterns were computed using the entire 400-ms duration of each stimulus. Peaks and shoulders were located using the following strategy which was derived from Scheffers [13] and Assmann and SUMMERFIELD [1].(i) The excitation pattern was sampled at integer multiples of the fundamental frequency (f_0) to avoid confusing resolved harmonics with low-frequency formants. (ii) Peaks were located as negative-going zero-crossings in the 1st differential of the sampled excitation pattern computed with respect to frequency. (iii) Shoulders were located as positive-going zero-crossings in the 3rd differential. The thin line in Fig. 5A shows the excitation pattern of a simplified vowel containing a peak at 17.5 erbs and a shoulder at 20.5 erbs. The f_0 is 100Hz. The thick line shows the sampled excitation pattern. Fig. 5B shows that the peak is marked by a

negative-going zero-crossing in the 1st differential. Fig. 5C shows that both the peak and the shoulder are marked by positive-going zero-crossings in the 3rd differential.

Table II summarises the results of applying this analysis to the threshold stimuli. There are three entries for each vowel and slope, corresponding to F1, F2, and F3, from left to right. A ✓ indicates that the formant was specified by a peak, a ✖ indicates that the formant was specified by a shoulder. A − indicates that no zero-crossing could be found within ±100Hz of the nominal formant frequency. With the extreme slopes of -1dB/erb and +2dB/erb, several formants were specified by shoulders rather than by peaks.

4.2 Discussion of Experiment 2

Table II shows that vowel identification generally requires formants to be specified as spectral peaks. However, when the overall spectrum slopes steeply, some formants of some vowels need only be specified as shoulders. Does this mean that listeners interpret shoulders as evidence of formants? Possibly not. Listeners could use the following strategy of "spectral subtraction" [14] to perform the 2I5AFC task: (i) subtract the internal spectrum of the masker from the internal spectrum of the masker plus target to compute a "difference pattern"; (ii) treat peaks in the difference pattern as formants.

If listeners used this strategy and the auditory system were linear, vowel identification thresholds would not vary with spectral slope. Fig. 4 shows that thresholds do vary with spectral slope. However, other factors including the deterioration of frequency selectivity at high intensities and the difficulty of detection near absolute threshold, may raise thresholds when the overall slope is extreme. Accordingly, Experiment 3 was run to determine whether thresholds are raised if subjects are prevented from benefiting from the strategy of spectral subtraction.

	Slope				
Vowel	-1	-0.5	0	+1	+2
/i/	✓✓✓	✓✓✓	✓✓✓	✓✓✓	−✖✓
/a/	✖✖✖	✓✓✓	✓✓✓	✓✖✓	✖✖✖
/u/	✓✓✓	✓✓✓	✓✓✓	✓✓✓	−✓✖
/ɒ/	✓✖✖	✓✓✓	✓✓✓	✓✓✓	−✖✖
/ɜ/	✖✓✓	✓✓✓	✓✓✓	✓✓✓	✖✖✖

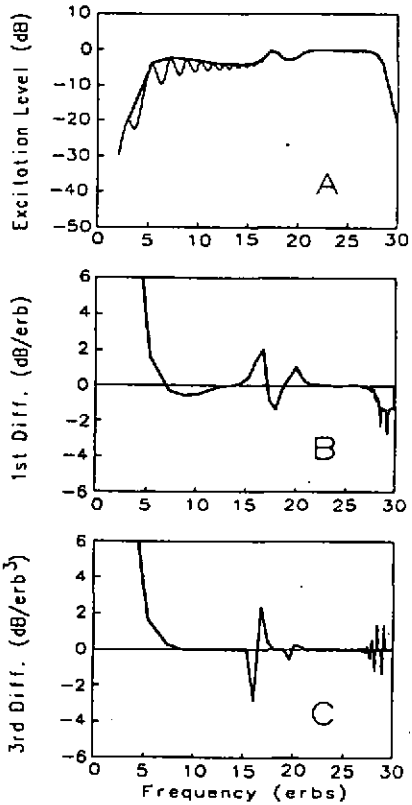

Fig. 5A: Excitation pauern and sampled excitation pattern of a simplified vowel. B: first differential. C: third differential of the sampled excitation pattern.

Table II: Results of formant analysis for Experiment 2. Formants are specified as peaks (✓), shoulders (✖) or not present (−), for all formants (F1, F2 and F3 from left to right), slopes and vowels. See text for details of analysis.

MINIMAL SPECTRAL CONTRAST FOR VOWEL RECOGNITION AS A FUNCTION OF SPECTRAL SLOPE

5. EXPERIMENT 3

There was a major procedural difference between Experiment 3 and Experiment 2. In Experiment 2, thresholds were estimated for each vowel separately with each slope. In Experiment 3, thresholds were estimated with slope varying randomly (among the 5 slopes) from trial to trial within the adaptive procedure for each vowel. There were then four conditions, distinguished by the aspects of the masker that differed between the two intervals of the forced-choice procedure, as follows: (i) intensity and slope did not differ; (ii) intensity was fixed, slope varied randomly (among the 5 slopes); (iii) slope was fixed, intensity varied randomly over a 10 dB range; (iv) both slope and intensity varied randomly. Spectral subtraction would be a beneficial strategy in Condition (i), but would be of little or no use in the other three conditions.

5.1 Results of Experiment 3

The mean contrast at threshold in each condition, averaged over four subjects, has been plotted in Fig. 6. The letters next to the condition numbers indicate whether level (L) or slope (S) varied randomly within trials. The four thresholds do differ significantly ($F_{3,9}$=6.3, $p<0.05$). Listeners benefited from spectral subtraction when it was a useful option in Condition (i) and suffered from the inability to use it in Condition (iv). It is possible, therefore, that listeners used the strategy in Experiment 2 to convert shoulders in the excitation pattern into peaks in the difference pattern.

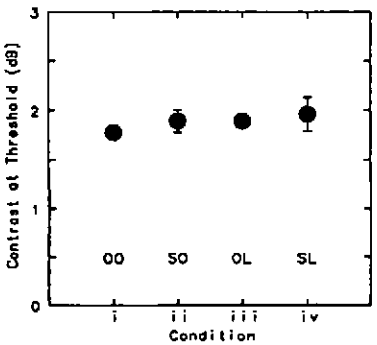

Fig. 6: Results of Experiment 3. S signifies that slope varied and L signifies that level varied between the two intervals of the task. 0 signifies no variation.

It is not possible to verify that formants at threshold were specified by shoulders rather than peaks in Experiment 3, because a "composite" threshold was obtained for each vowel with slope randomly varying. Therefore, Experiment 4 was designed to estimate thresholds for each vowel separately with each slope, in a condition were spectral subtraction would not be beneficial.

6. EXPERIMENT 4

In this experiment, slope and level were varied randomly between the intervals of the forced-choice procedure as in Condition (iv) of Experiment 3. However, 25 adaptive procedures were now interleaved to allow a separate threshold to be estimated for each vowel with each slope. Two subjects, AL and IL, were tested. Each had taken part in Experiment 2.

6.1 Results of Experiment 4

Mean thresholds for each slope, averaged over the 5 vowels, have been plotted as the squares in Figs. 7A and 7B for AL and IL, respectively. Error bars show plus/minus one standard deviation of the mean. For comparison, the circles show the same subjects' thresholds from Experiment 2. Thresholds were somewhat higher in Experiment 4 and again differ significantly as a function of slope for both subjects (AL: $F_{4,12}$=21.9, $p<0.01$; $F_{4,12}$=9.9, $p<0.01$), with higher thresholds at the two extreme slopes.

The 25 threshold stimuli for each subject were synthesized and analysed to establish whether formants were specified by peaks or shoulders using the procedure described in Section 4.1. The results are shown in Tables III (AL) and IV (IL), in the same format as Table II. (The meaning of "✚" symbols is explained below.) Some formants of some vowels were specified only by shoulders.

MINIMAL SPECTRAL CONTRAST FOR VOWEL RECOGNITION AS A FUNCTION OF SPECTRAL SLOPE

Vowel	Slope -1	-0.5	0	+1	+2
/i/	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓
/a/	✖✓✓	✓✓✓	✓✓✓	✓✚✓	✖✖✖
/u/	✓✚✓	✓✓✓	✓✓✓	−✓✚	−✖✖
/ɪ/	✓✓✓	✓✓✓	✓✓✓	✓✓✓	−✖✚
/ɜ/	✖✓✓	✓✓✓	✓✓✓	✓✓✓	✖✖✚

Vowel	Slope -1	-0.5	0	+1	+2
/i/	✓✓✚	✓✓✓	✓✓✓	✓✓✓	−✖✚
/a/	✖✓✓	✓✓✓	✓✓✓	✓✚✓	✖✖✖
/u/	✓✚✓	✓✓✓	✓✓✓	−✓✚	−✚✖
/ɪ/	✓✖✚	✓✓✓	✓✓✓	✓✓✓	−✚✚
/ɜ/	✖✓✓	✓✓✓	✓✓✓	✓✚✓	✖✖✚

Table III (top) and Table IV (bottom): Results of subjects AL and IL respectively from Experiment 4. ✓s indicate peaks, ✖s indicate shoulders, ✚s indicate a shoulder is transformed into a peak when suppression is accounted for and −s indicate no feature could be found.

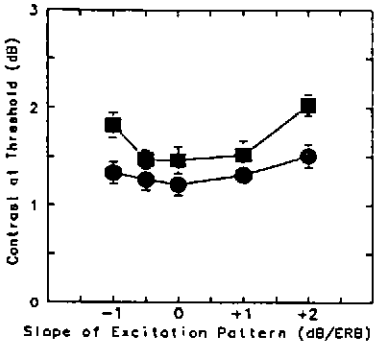

Fig 7: Squares show the results of subject AL from Experiment 4. Circles show the results for the same subject from Experiment 2.

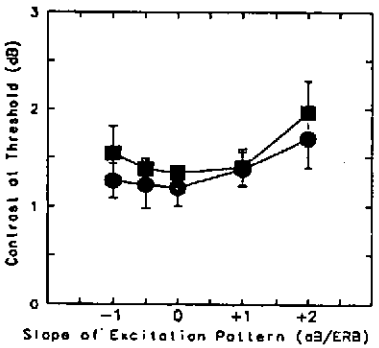

Fig. 8: Squares show the results of subject IL from Experiment 4. Circles show the results for the same subject from Experiment 2.

6.2 Discussion of Experiment 4

The results of Experiment 4 confirm A&S's hypothesis that listeners can detect formants that appear only as shoulders in the excitation pattern of a vowel. This outcome could imply that listeners can detect formants from shoulders in their own internal spectrum. However, this is not certain because excitation patterns, as computed here, do not include effects of non-linearities, such as lateral suppression, which might convert shoulders into peaks.

Effects of suppression are hard to predict precisely. However, an approximation can be achieved by reducing the bandwidths of the filters in the filter bank by 20%, since suppression reduces the bandwidths of auditory filters by about this amount [15]. This change was made and the threshold stimuli from Experiment 4 were re-analised. Some formants which appeared as shoulders in the previous analysis now appear as peaks. They are marked by ✚s in Tables III and IV. However, several others remain as shoulders.

7. CONCLUSIONS

These experiments have shown that when a vowel has a flat or gently sloping spectrum, listeners with normal hearing can detect a formant from a spectral peak that creates little more than 1dB of spectral contrast. This figure is an order of magnitude smaller than the spectral contrast generally found when vowels are spoken naturally in quiet. It provides

the necessary tolerance to sustain accurate vowel identification when spectral contrast is reduced by noise, reverberation, or impaired frequency selectivity.

When a vowel has a steeply sloping spectrum, listeners can detect some formants from spectral shoulders rather than spectral peaks. This ability may be useful when more than one talker is speaking and a formant in one voice is partially masked by a more intense formant in the competing voice. It may also be useful when a single speaker's formants are close together in frequency.

In natural, and thus echoic, environments, harmonic amplitudes are distorted by cancellation and reinforcement from echoes [e.g. 16] creating spurious peaks and shoulders in the spectrum. In such circumstances, the criterion for accepting a peak or shoulder as evidence of a formant would have to be set considerably more conservatively than was possible here when signals were presented in quiet through headphones.

8. REFERENCES

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

[2] L A CHISTOVICH & V V LUBLINSKAYA, 'The "center of gravity" effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli,' Hearing Research 1 p185-195 (1979).

[3] B C J MOORE & B R GLASBERG, 'Suggested formulae for calculating auditory-filter shapes and excitation patterns,' J. Acoust. Soc. Am. 74 p750-753 (1983).

[4] B C J MOORE & B R GLASBERG, 'Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns,' Hearing Research 28 p209-225 (1987).

[5] J N van DIJKHUIZEN, P C ANEMA & R PLOMP, 'The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences,' J. Acoust. Soc. Am. 81 p465-469 (1987).

[6] H LEVITT 'Transformed up-down methods in psychoacoustics,' J. Acoust. Soc. Am. 49 p467-477 (1971).

[7] ANSI, 'Specifications for audiometers,' ANSI S3,6-1969 (1969).

[8] P DELATTRE, A M LIBERMAN, F S COOPER & L J GERTSMAN, 'An experimental study of the acoustic determinants of vowel colour: observations on one- and two- formant vowels synthesized from spectrographic patterns,' Word 8 p195-210 (1952).

[9] M R LEEK, M F DORMAN, & A Q SUMMERFIELD, 'Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners,' J. Acoust. Soc. Am. 81 p148-154 (1987).

[10] A Q SUMMERFIELD, A SIDWELL, & T NELSON, 'Auditory enhancement of changes in spectral amplitude,' J. Acoust. Soc. Am. 81 p700-708 (1987).

[11] C W TURNER, & D J van TASELL, 'Sensorineural hearing loss and the discrimination of vowel-like stimuli,' J. Acoust. Soc. Am. 75 p562-565 (1984).

[12] C C HENN, & C W TURNER, 'Pure tone increment detection in harmonic and inharmonic backgrounds,' J. Acoust. Soc. Am. 88 p126-131 (1990).

[13] M T M SCHEFFERS, 'Sifting vowels: auditory pitch analysis and segregation,' Doctoral Dissertation, Groningen University, The Netherlands (1983).

[14] A Q SUMMERFIELD, & P F ASSMANN, 'The perception of concurrent vowels: effects of harmonic misalignment and pitch period asynchrony,' Submitted to J. Acoust. Soc. Am. (1990).

[15] B C J MOORE, & B J O'LAUGLIN, 'The use of non-simultaneous masking to measure frequency selectivity and suppression,' In: Frequency selectivity in Hearing, p251-308, Ed. B C J MOORE, Academic Press, London (1986).

[16] R PLOMP, & DUQUESNOY, 'Room acoustics for the aged,'J. Acoust. Soc. Am. 68 p1616-1621 (1980).