Pitch and the Perceptual Separation of Simultaneous Vowel Sounds

M.T.M. Scheffers
Eindhoven, the Netherlands

1. Introduction

Two experiments were carried out investigating identification of simultaneous vowel sounds by listeners. Our research is inspired by the intriguing question - first posed by Colin Cherry as the 'Cocktail Party Problem' (Cherry, 1953) - of how listeners are able to perceive the speech of a single speaker separately from a background of interfering voices. Cherry mentioned voice pitch as one of the factors possibly facilitating the separation. Much earlier, Stumpf (1890) had reported that the sounds of two musical instruments tended to fuse into a single percept when both instruments played exactly the same note, but were separately audible when different notes were played. More recently, Brokx and Nooteboom (1982) found that speech sounds presented in a background of speech from another or even the same speaker, could be identified considerably better when there was a difference of more than 1 semitone between the pitches of the two sounds. These observations prompted us to investigate the role of differences in pitch between simultaneous vowels in the perceptual separation process.

Identification of pairs of unvoiced vowels was investigated in the second experiment. This experiment was conducted in order to determine to what extent listeners could use information derived from the spectral envelope of the sound for identifying the vowels.

2. Experiment 1

The stimuli of the first experiment consisted of two different voiced vowels. The waveforms of the vowel sounds were computed using a software five-formant speech synthesizer. Eight vowels were used viz. the Dutch /i/, /y/ /I/, /ɛ/, /ə/, /ɑ/, /ɔ/, and /u/. Formant structures were taken from Govaerts' study of Dutch vowels (Govaerts, 1974). The duration of each vowel was 220 ms including cosine-shaped onset and offset ramps of 20 ms. The vowels were added with no temporal onset difference, starting in zero-phase. They had about equal subjective loudness. Six F_0 differences were used: 0, ¼, ½, 1, 2 and 4 semitones. The average F_0 was 150 Hz. For each pair of vowels with unequal F_0 two stimuli were made, one in which one vowel had the lower and one in which the other had the lower F_0. The waveforms of the 308 different combinations were digitally stored on disk.

Twenty subjects took part in Exp. 1. They had normal hearing and were familiar with synthesized speech sounds and psychoacoustic experiments. They were tested individually. The subjects were seated in a sound-treated booth and received the signals diotically through TDH 49-P headphones. The signals were band-pass filtered from 50 Hz to 5 kHz and presented at a level of about 60 dB SL.

A minicomputer controlled the presentation of the stimuli and recorded the responses. The subjects were instructed to respond to each stimulus by pushing two buttons on a panel of eight, each button representing one of the eight vowels used. All vowels were played to them before the experiment started. No feedback was given on their responses. The subjects attended the experiment in four sessions held on consecutive days. In every session, each of the 308 stimuli was presented once, in a random order that differed for each subject and for each session. A session lasted about half an hour.

3. Results

A synopsis of the results is presented in Fig. 1. The solid line in this figure gives the percentage correctly identified combinations (both vowels correct), averaged over the 28 combinations. The dashed line depicts the average percentage of individual vowels correct. No significant difference was found between the performance on the stimuli in which one vowel had the higher F_0

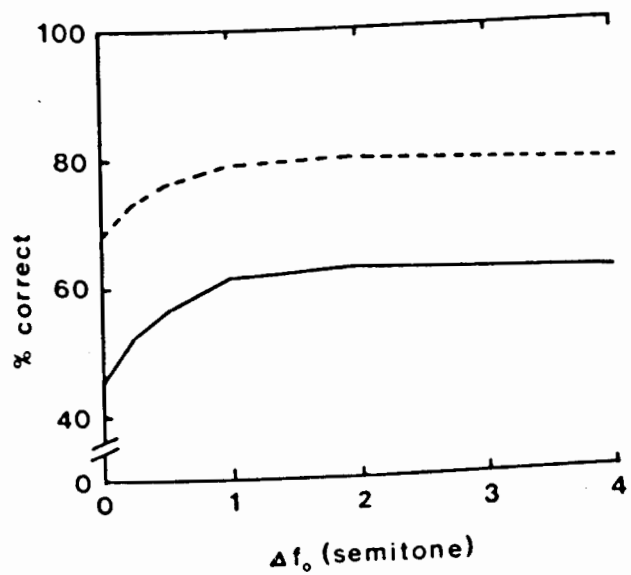

Figure 1: Percentage correct identification of two simultaneous voiced vowels as a function of the difference between the fundamental frequencies of the vowels. The solid line shows the average identification scores on the 28 combinations used (both vowels correctly identified) and the dashed line the average scores on individual vowels.

and on those in which it had the lower F_0. The results are therefore averaged over 'positive' and 'negative' F_0 differences.

The scores differed much for different combinations. They were in general lowest for combinations of similar vowels, such as two front vowels or two back vowels, and were highest for dissimilar combinations such as a front and a back vowel. The scores were down to around chance level (4%) for only a few combinations of vowels with equal F_0. It can be seen from Fig. 1 that the scores increased with increasing F_0 difference up to 1 or 2 semitones.

4. Experiment 2

When it was found that identification scores even on pairs of vowels with identical F_0's were generally well above chance level, a second experiment was devised. Stimuli in this experiment consisted of two different unvoiced vowels. They were constructed in the same way as the stimuli for Exp. 1. The unvoiced vowels had the same spectral envelopes as the voiced ones. The stimuli were D-A converted, band-pass filtered from 50 Hz to 5 kHz and recorded on magnetic tape with an ISI of 3 s. The tape contained every stimulus eight times in random order.

Eighteen subjects with normal hearing took part in this experiment. They were asked to identify the two vowels in a stimulus and to write down a phonemic transcription of both vowels on an answer form. The test method was the same as in Exp. 1 except for the use of a tape and written responses.

5. Results

Performance on the unvoiced vowels was significantly lower than for voiced vowels with equal fundamentals ($p < .01$). The identification score on combinations was 26% for the unvoiced stimuli and 45% for the voiced stimuli and the average score on individual vowels was 56% and 69% respectively. The same tendency for pairs of vowels with dissimilar formant structures to be better identified than pairs with similar structures was also found here.

6. Discussion and Conclusions

The most surprising result of the experiments is that identifiability of two simultaneous vowels was far above chance level even if both vowels had the same fundamental frequency or when they were unvoiced. The result that simultaneous unvoiced vowels were less well identifiable than simultaneous voiced vowels with equal F_0's cannot yet be explained. This was contradictory to what one would expect from the fact that formants are more sharply defined in unvoiced than in voiced vowels, although this is only true for the long-term spectrum. Identification scores on voiced pairs increased by about 18% on average when the F_0 difference between the two vowels was increased

from 0 to 2 semitones. It is noteworthy that at least one vowel was correctly identified in 95% of the voiced stimuli and in 86% of the unvoiced ones.

Identification scores on combinations of vowels with strongly differing spectral envelopes like /i/ and /a/ were much higher than the scores on vowels with relatively similar spectral shapes like /i/ and /y/. This supports our theory of a 'profile' analysis (cf. Spiegel and Green, 1981) in the recognition process. A profile is considered to be a relatively simple image of the envelope of the spectral representation of the sound in the peripheral ear. Recognition is then a process of matching reference profiles to the one of the present spectrum and identifying the sounds on basis of the best fitting profiles. The profile is probably best defined around the first two formants of the vowel. The shape of the profile near the frequencies of these formants apparently weighs most in the matching (cf. Klatt, 1982; Scheffers, 1983). If there is a great difference between the profiles of the composing vowels, identifiability of the combination is relatively high and little influenced by F_0 differences. If the profiles are rather similar, however, F_0 differences can aid to separate the profile of the combination in parts belonging to one of the vowels and parts belonging to the other or maybe to both. Separation is supposed to be guided by the harmonic fine structure of the spectrum. This is only possible for relatively low frequencies because high harmonics are not separately detectable in the auditory system (e.g. Plomp, 1964). The theory is supported by the results of a pilot experiment in which it appeared that two pitches could be perceived when the F_0 difference was greater than 1 semitone, while for smaller differences only one (beating) pitch was heard. We may therefore expect little further improvement of the performance when the F_0 difference is increased beyond 1 semitone. A decrease in performance can even be expected for harmonic intervals between the two F_0's such as a major third (4 semitones) and especially for an octave because many harmonics of both vowels will then coincide. A clear decrease in performance for the 4-semitone difference was indeed found in the results for 8 combinations.

Acknowledgement

This research was supported by a grant from the Foundation for Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.), grant nr. 15.31.11.

References

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

Cherry; E.C. (1953). Some experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Am.* **25**, 975-979.

Govaerts, G. (1974). *Psychologische en fysische structuren van perceptueel geselecteerde klinkers, een onderzoek aan de hand van Zuidnederlandse klinkers* Doctoral thesis, University of Louvain.

Klatt, D.H. (1982). Predictions of perceived phonetic distance from critical-band spectra: a first step. *Proc. ICASSP* **82 (2)**, 1278-1281.

Plomp, R. (1964). The ear as a frequency analyzer. *J. Acoust. Soc. Am.* **36**, 1628-1636.

Scheffers, M.T.M. (1983). Identification of synthesized vowels in a noise background. In preparation.

Spiegel, M.F. and Green, D.M. (1981). The effects of duration on masker profile analysis. *J. Acoust. Soc. Am.* **70 (1)**, S86(A).

Stumpf, C. (1890). *Tonpsychologie*. Lizensausgabe des S. Hirzel Verlages, Leipzig. Republished in 1965 by Knef-Bonset, Hilversum-Amsterdam.