Sifting vowels. Auditory pitch analysis and sound segregation.

Scheffers, Michaël Titus Maria

Publication date:
1983

FINALE

SUMMARY

The aim of this study has been to investigate the role of pitch in perceptual separation of vowel sounds from interfering background sounds. What have we learned? I will first give a summary of the main conclusions from this thesis, followed by some speculations regarding the implications for normal speech communication situations.

The accuracy with which listeners can perceive the pitch of periodic signals disturbed by background noise was investigated in the experiment described in Chapter I. It was found that this accuracy decreased when the signal-to-noise ratio was decreased. Differences in the fundamental frequency could nevertheless be discriminated rather accurately (6%) as soon as the signal level was just raised above masked threshold. This shows that pitch is at least a possible clue for separation of a periodic sound at low signal-to-noise ratios. It was also found in that experiment that, especially for vowel sounds, not only the residue pitch but also the pitch of a single strong signal component could be dominant, in particular at low S/N ratios.

A model of human pitch perception - the second line of study - was introduced in the second chapter. A simulated auditory spectrum was constructed in this model by convolving a high-resolution FFT spectrum of the sound with stylized pure-tone excitation patterns. A peak detector determined the frequencies of signal components that were resolvable in this spectrum. A harmonics sieve procedure determined which fundamental best fitted this set of resolved components. The model was tested on inharmonic signals and on the periodic signals in noise from the first chapter. The results of these tests have shown that the quantitative predictions by the model were in quite good agreement with the results obtained for human listeners. It appeared that the pitch of a single strong signal component had to be taken into account to obtain good predictions for the vowel sounds in noise.

In the third chapter we came closer to our main point of interest, viz. the identification of vowels in noise. The results of the listening experiments reported in Chapter III have shown that identification thresholds of vowels, masked by pink noise, were determined by the detectability of the first and second formants. Identification thresholds were found to be systematically lower when the onset of the noise preceded the onset of the vowel than when the noise was only present during the presentation of the vowel. The detectability of a formant could fairly well be predicted on the basis of the signal-to-noise ratio in a 1/3-octave band, roughly representing the bandwidth of the auditory frequency analysis. It was furthermore found that the identification score for a voiced vowel was often higher than that for its unvoiced version at the same S/N ratio. This might imply that the correlation that exists between the harmonics in the different formants of voiced vowels (i.e. the common fundamental) but not in unvoiced vowels, facilitated identification of the voiced vowels. A test of a simple vowel classifier that was linked to the model of pitch perception, showed promising performance in identifying the vowels in noise, though the test was rather limited.

Experiments on the identification of two simultaneous vowels were described in Chapter IV. The results of these experiments have shown a surprising ability of the listeners to identify both vowels in a stimulus correctly even if both vowels had the same fundamental frequency or when they were both unvoiced. The identification scores on voiced vowels clearly improved when the difference between their fundamentals was increased till over 1 semitone. It was, however, also found that best performance did not reach the values that could be expected on the basis of the identification scores on the individual vowels in quiet. An interpretation of perceptual separation in terms of a "profile analysis" of the spectra of the sounds was proposed to explain these results. Simultaneous vowels are supposed to be identified in this interpretation by a comparison of a gross image of the spectral envelope (the profile) of the compound signal with stored profiles of the individual vowels. Differences between the fundamental frequencies of the two vowels were supposed to lead to a separation of the profile into (mainly low-frequency) parts that belong to different vowels and parts that may belong to both vowels.

An attempt has been made to implement this concept of profile analysis in an extension of the model of pitch perception and of the vowel classifier. This attempt was hampered by the fact that no data were available on the multiple pitches that listeners can perceive in simultaneous voiced vowels. These data could also not be obtained within the scope of this research project. A comparison between the pitch estimates by the model and the fundamentals at which the signals were generated, has shown that the model could estimate one of the two fundamentals correctly in almost all stimuli. This and the not so poor performance in estimating both fundamentals correctly led me to assume that the approach to modelling pitch perception, used in this study, represents a promising line of research.

The vowel classifier could predict at least one vowel correctly in almost all stimuli. Performance on both vowels correct for unvoiced pairs was in good agreement with the average performance of the listeners. The scores on voiced vowels, however, were lower than those for the listeners and the increase in performance with increasing f_0 difference between the simultaneous vowels found for the listeners could not be predicted. A further test using a version of the model in which the presence of two sounds was not presupposed, yielded better results. This led me to assume that in the auditory system, identification is not performed after separation of the spectral information. In other words: if some combination of a subset of formants detected would yield a possible vowel sound, this vowel may be identified. This would not only explain the finding that the listeners could identify at least one vowel correctly in almost all combinations but also the better-than-chance score on the second vowel because the remaining extra formants would allow them an educated guess on that vowel. The role of pitch differences might then be to indicate to the identification process that the signal comprises more than one sound. Furthermore, pitch differences can aid to determine which formants belong to the same vowel and which do not. Pairs of unvoiced vowels were found to be more difficult to identify. This might mean that the profiles of such vowels are less well defined than those of voiced vowels or that it is more difficult in such signals to group formants together as belonging to one vowel.

Summarizing, the conclusion is drawn that formant detection and classification, and pitch processing are two important processes in perceptual separation. The two mechanisms are seemingly independent. The formant processor probably operates on the spectral envelope of the signal, while the pitch processor operates on the spectral fine structure. They can support each other to some extent. On the one hand, the formant detector can indicate to the pitch processor in which frequency regions harmonics from one fundamental can be found. On the other hand, regions with a different harmonic structure found by the pitch processor can be an indication to the formant classifier that components found in those regions probably belong to another sound. More in general, if the pitch processor detects two different pitches, this can be an indication to the formant classifier to search for more than one sound.

The results from the listening experiments could in general be fairly well predicted using a rather simple power-spectrum model. To predict the results of the experiments on perceptual separation of simultaneous vowels, however, another way to construct the spectral envelope of voiced vowels should be devised for the model. I expect that, in combination with another criterion for the detection of multiple pitches without using the assumption of the presence of one or more sounds in the signal that is analyzed, better prediction of the results for listeners could be obtained.

In normal communication situations, both processes can be supported by tracking mechanisms since both the formant frequencies and the fundamental frequency of speech sounds vary relatively slowly. This whole system can expected to be supervised by a mental monitoring function which can guide either process in some direction, following expectations based on, for instance, continuity constraints, syntactic rules and the context (see the introduction to this thesis). It was also found that when the onsets of two simultaneous sounds were different (Chapter III, see also Scheffers, 1979), perceptual separation was facilitated. A mechanism that can store the profile of the first sound and can compare that profile which that of the compound sound could explain this effect. Such onset differences will occur frequently in normal situations. The cooperation between all these various processes and expectations may explain the ease with which listeners can separate simultaneous sounds.

RECOMMENDATIONS FOR FURTHER RESEARCH

During the study reported in this thesis, I have come across a number of interesting questions, the investigation of which may cast additional light on the processes underlying perceptual separation of simultaneous sounds. Some of these issues fell beyond the scope of this study, others beyond its time schedule. A number of issues will be given below in conclusion to this thesis.

Pitches evoked by simultaneous vowels. This study will be necessary before a decision can be made on further modifications of the model. The best way to measure the multiple pitches that listeners may perceive in simultaneous vowels is probably by using a matching task. The residue pitches could be estimated by matching the pitch of such a sound to that of a pulse train. Pitches evoked by individual harmonics can be determined by matching with the pitch of a pure tone. The latter experiment would be a rather extensive one if we would like to get insight on the strength of each of these pitch percepts.