Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Vowel perception: The relative perceptual salience of selected acoustic manipulations

Carlson, R. and Granström, B. and Klatt, D.

KTH Computer Science and Communication

journal: STL-QPSR
volume: 20
number: 3-4
year: 1979
pages: 073-083

http://www.speech.kth.se/qpsr

B. VOWEL PERCEPTION: THE RELATIVE PERCEPTUAL SALIENCE OF SELECTED ACOUSTIC MANIPULATIONS **

R. Carlson, B. Granström & D. Klatt*

Abstract

A set of 66 vowels acoustically similar to /ae/ have been synthesized by adding together sinusoidal harmonics of the appropriate frequencies, amplitudes, and phases. Subjects were asked to estimate the psychophysical distance between each stimulus and a reference vowel in a 300-trial randomized test. Normalized subjective distance scores have been computed for stimulus manipulations involving formant frequencies, formant bandwidths, spectral tilt, phase relations among harmonics, vocal tract length, and filtering passband/stopband. The data can be used to develop better spectral measures of psychophysical similarity among vowels. In addition, large perceptual effects of phase manipulations (in particular, the observation that random phase lead to a harsh "aperiodic" sensation) implicates temporal processing of neural spike information as the most likely mechanism for pitch estimation during speech perception.

Introduction

The objective of the preliminary research described here is to quantify and compare the psychophysical importance of a number of acoustic parameters related to vowel perception. This research is an extention of earlier work on the discriminability of spectral slope changes in several synthetic vowels (Carlson & Granström, 1976). We have generated 66 different versions of a vowel similar to /ae/ by manipulating several acoustic dimensions related to the voicing source and other dimensions related to the vocal tract transfer function. Most of the stimuli to be generated could have been produced by an existing digital formant synthesizer. However, in a few cases, it is necessary to have direct control over the amplitude and phase of each harmonic in the voicing source. Therefore, we have designed a new vowel synthesizer (the "additive harmonic synthesizer") that can generate roughly the same waveform as a formant synthesizer by adding together sinusoids of the appropriate frequencies, amplitudes, and phases (Klatt, 1979a).

Stimulus Definitions

The reference stimulus is defined in Table IV-B-1. As a control the harmonic synthesizer was compared to a conventional cascade formant synthesizer. There is little difference between the results

* Rm 36-523, Mass.Inst. of Technology, Cambridge, MA 02139, USA

** This is a slightly revised version of a paper publ. Speech Communication Papers, 97th ASA-Meeting, June 1979 (eds Wolf & Klatt).

of the two methods of synthesis, either in terms of waveform or spectra. Characteristics of the remaining 65 comparison stimuli are described below when discussing the results.

Time-Varying Control Parameters:

Time =	0	70	140	300	msec
F0 =	110	125	125	100	Hz
Time =	0	20	200	300	305 msec
AV =	48	60	60	54	0 dB

Constant Control Parameters:

F1-F5 = 700, 1800, 2500, 3300, 3700 Hz

B1-B5 = 60, 140, 150, 200, 250 Hz

DB0 = -6 dB (Spectral tilt in dB per octave)

Table IV-B-I. Control parameter values for the reference stimulus. Time varying parameters are updated every period by linear interpolation between time-value points given here.

Listening Test

A 300-trial randomized listening test was prepared in which subjects listened to pairs of stimuli and were asked to estimate how different the pair of sounds were, using a scale from 0 to 10. The reference stimulus was always one of the pair of sounds. A familiarization sequence was played before the test began. Subjects were instructed to respond to the amount of difference, no matter what type of change was heard.

Eight subjects listened to the test tapes in two different orders. The data for each subject were first analyzed separately and an f-ratio of cross-stimulus variance to within stimulus variance was used to see if the data obtained from any subject should be discarded. All subjects produced significant f-ratios, and their data looked superficially similar (an analysis of variance is pending).

Results

The results of the perceptual distance judgement experiment are summarized in Fig. IV-B-1. The plots present perceived distance averaged across subjects after the data for each subject have been normalized such that the total responses from each subject have a mean of 2.73 and a variance of 5.78. These (arbitrary) scale factors were chosen so as to give the reference stimulus an average distance score of zero when it is paired against itself, and to give the most different stimulus of the test an average distance score of 10. Preliminary estimates of the standard deviations of the means plotted in Fig. IV-B-1 suggest that differences in mean subjective distance greater than about 0.5 are significant at the 0.05 level.

Formant Frequency

The panels in the first two indicate the results of changing one, two, or all formant frequencies by 4 and 8 percent. The just-noticeable difference (JND) for a change to the frequency of a single formant (with F0 held constant) is about 3 percent (Flanagan, 1957). Thus it is not surprising that a 4 percent change to F1 or F2 results in a small subjective distance from the reference (see Panel A). A given percentage change to F1 or F2 produces about the same (nearly linear) change in normalized subjective distance -- an 8% change corresponding to a distance of about 3.

The results of varying F1 and F2 simultaneously are presented in Panel B of Fig. IV-B-1. For a given percentage change, subjective distance is somewhat greater when two formants are changed together -- an 8% change corresponding to a distance of about 4. It is perhaps surprising that it makes little difference whether F1 and F2 increase and decrease together or in opposite sign. Presumably this result would not hold for larger changes in formant frequency or for other vowels where the formants are closer together than for /ae/.

The results of varying all formant frequencies are summarized in Panel C of Fig. IV-B-1. Simultaneous increases or decreases in all formants cause the spectrum to tilt more upward or downward, while changes to vocal tract length shift the frequencies of formant peaks without introducing an additional spectral tilt cue.

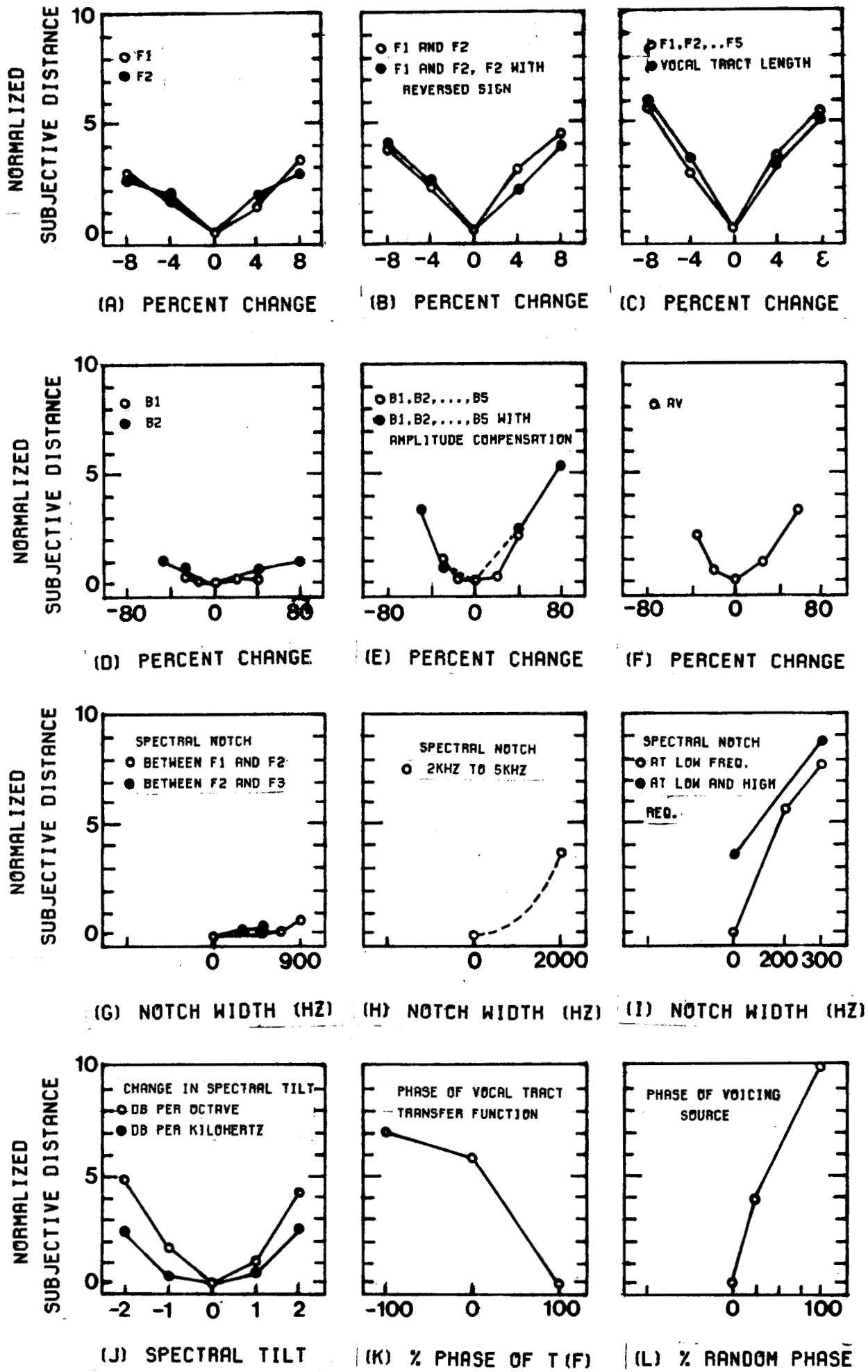

Fig. IV-B-1. Results of the perceptual experiments.

The similarity of the two curves suggest that the auditory system is more interested in formant frequency shifts than in changes to spectral tilt (see more on this subject below). An 8% change in all formant frequencies results in a subjective distance of about 5.5, which is greater than that obtained by shifting only one or two formants.

Formant Bandwidths

The results of manipulating formant bandwidths are summarized in Panels D-E in Fig. IV-B-1. In terms of percent change, the perceptual system does not rank bandwidth changes as very important. We employed bandwidth changes that were up to ten times greater than formant frequency changes, as can be seen by the horizontal scale used in Panels D and E. Yet, for a 40% change to B1 or to B2, the subjective distance from the reference is less than 1. The auditory system seems to be at least 20 times more sensitive to changes in formant frequency than to changes in formant bandwidth in an experiment where F0 varies and there are many types of acoustic manipulations to be compared -- even though under ideal conditions, the JND for first formant bandwidth is 10%, or only about three times the JND for formant frequency (Flanagan, 1957). (We chose an area of the vowel space where small changes in formant frequencies shouldn't affect phonemic category of the vowel. However, perceptual contrast effects invoked by hearing the reference were often strong enough to shift perceived vowel quality. Thus, linguistic experience may have increased the psychophysical distance for formant frequency changes relative to bandwidth changes.)

If all formant bandwidths are shifted together, greater perceived distances are obtained, as shown in Panel E. This acoustic manipulation increases (or decreases) the amplitudes of all formant peaks by a percentage change about equal to the percentage change in bandwidth, while leaving the harmonic amplitude unchanged in the vicinity of the valleys between formant peaks. A shift in the bandwidth of a formant thus has two effects on the spectrum: the amplitude of the formant peak is changed and the width of the energy concentration changes.

It is of interest to compare the bandwidth-change data of Panel E with the data of Panel F where only overall vowel amplitude has been changed. A change in overall amplitude produces essentially the same subjective distance as a comparable percentage change in all formant bandwidths. All harmonics change in amplitude if AV changes, while only those harmonics very near a formant peak change in amplitude when bandwidths are changed. It is perhaps surprising that overall amplitude of the spectrum would induce a greater change in subjective distance than a change to spectral shape, but this particular spectral shape change is really quite small when viewed through a set of critical band filters (see below).

To test whether a bandwidth change could really be thought of, in perceptual terms, as only affecting formant peak amplitudes, we synthesized a set of stimuli in which the bandwidth changes were accompanied by a comparable change in overall amplitude. Thus formant peak amplitudes were unchanged, while the valleys between formants moved up and down. The results, shown as the solid triangles in Panel E, indicate that this attempt to cancel amplitude cues induced similar changes in spectral distance; i.e. although a greater range of bandwidth variation was employed in anticipation of reduced perceptual distance, data points at +40% and minus 29% fall almost exactly on top of the uncompensated data. Data to be presented below suggest that the valleys between formants are not as important as formant peaks, but that the auditory system weights changes at very low frequencies (near the first one or two harmonics) heavily, and this probably accounts for the similarity between the amplitude compensated and uncompensated data of Panel E.

Spectral Notches

Spectral notches were created in the stimuli by defining a stop band where the amplitudes of harmonics in selected frequency regions were clamped to zero. To avoid transients that would be introduced by a discontinuous change in the amplitude of a harmonic as it moved into or out of a stop band, harmonic amplitudes were always linearly interpolated to desired values over the duration of each period. The results of introducing a spectral notch between

F1 and F2 and between F2 and F3 are shown in Panel G of Fig. IV-B-1. There is no measurable change in distance until the notch is 900 Hz wide and thus captures harmonics very close to a formant peak. Psychoacoustical data to be discussed below indicate that masking makes it difficult but not impossible to detect these notches -- in fact, the auditory system appears to have learned to pay less attention to a notch.

A spectral notch at high frequencies results in a vowel that has been low-pass filtered. We do not have enough data points to assess the details of subjective distance growth with low-pass cut-off frequency, but the single point plotted in Panel H of Fig. IV-B-1 indicate that removal of energy from 3000 to 5000 Hz has only a moderate effect.

Placement of a spectral notch at low frequencies, however, makes a much greater difference. In Panel I, results are plotted for the case where a notch has been located so as to remove only the first harmonic (notch width = 200 Hz) and only the first and second harmonic (notch width = 300 Hz). Attenuation of the lowest one or two harmonics has the greatest effect on subjective distance of any manipulation described thus far. There is little change in vowel quality, but the vowel sounds much like it would over a telephone. Also plotted in Panel I are the results of including both a low-frequency notch and a high-frequency notch so as to produce a 300-to-3000 Hz bandwidth telephone-like channel (solid triangle at the right) and a subjective distance of 8.9.

Spectral Tilt

The reference spectrum has an average spectral falloff of about -6 dB per octave (-12 dB per octave attributed to the voicing source spectrum compensated by a +6 dB per octave rise attributed to the radiation characteristic (Fant, 1960)). This spectral tilt was manipulated in two ways to produce the data shown in Panel J of Fig. IV-B-1. Results for an incremental change in spectral tilt of 1 or 2 dB per kiloHertz are given by the solid triangles. Two dB per kiloHertz causes a 6 dB change in the amplitude of the 4th or 5th formant peak relative to the first formant peak (the tilt was adjusted to leave the amplitude of F1 unaffected). The relatively

small score, about 2.5, associated with this tilt, is in agreement with previous results (Carlson & Granström, 1976) and with the minimal effect of bandwidth manipulations discussed above.

The second method of changing spectral tilt, that of using units of dB per octave, produced results shown by the open circles in Panel J. Greater subjective distance scores are obtained even though the formant peak amplitude changes are somewhat less (the amplitudes of F4 and F5 change by about 5 dB re F1 for a 2 dB per octave change in spectral tilt). The reason for the greater scores is clearly to be found in the greater changes at low frequencies induced by employing units of dB per octave.

Relative Phases among Harmonics

The reference stimulus was synthesized to have a combined voicing source and radiation characteristic that produced a magnitude spectrum falling off at -6 dB per octave and a phase spectrum of zero (cosine phase). The vocal tract transfer function then imposed its phase spectrum on the source, resulting in a waveform that appears to decay between glottal excitation pulses. The perceptual results of several perturbations to these phase relations are shown in Panels K-L of Fig. IV-B-1. If the phase associated with the vocal tract transfer function is removed, the resulting waveform has zero phase, all harmonics add in phase at one point in a period, and the waveform has a large peak factor (ratio of maximum to rms energy in a period). This resulted in a perceived distance of 5.9, a value that is perhaps larger than one might expect based on earlier efforts to measure the perception of phase in simpler harmonic complexes. A second phase manipulation that was tried concerned imposing the negative of the correct vocal tract transfer function phase lag, which is the same as turning the waveform around in time. During each period, the waveform appears to grow exponentially. This resulted in even a larger subjective distance score of 6.9.

A final phase manipulation concerned adding values from a uniformly distributed random variable to the initial phase of each harmonic at the beginning of the stimulus. For the particular set of random numbers that were chosen, the resulting waveform had a very

small peak factor and a harsh quality that resulted in the largest normalized subjective distance of any manipulation employed in this experiment, i.e. 10. In a second quasi-random-phase stimulus, the random variable was restricted in range to from zero to pi/2, resulting in a distance score of 4.

It has been said that the auditory system is insensitive to phase unless two harmonically related components fall within a critical band. Our results are not inconsistent with this view, but it seems that speech waveforms are characterized by harmonic complexes that frequently fall within a critical band. The harsh almost "aperiodic" quality of the random phase conditions that we have synthesized suggest strongly that there is a temporal component to the perception of pitch in speech (since the magnitude spectra of the stimuli are not changed at all, and combination-tone generation could shift the relative amplitudes of components, but this cannot account for the strong aperiodicity sensation). One might conjecture that the detailed motions of the basilar waveform at positions most sensitive to the lower formants of a vowel must have a particular pattern to ensure that a certain fraction of the neural spikes are time locked to a single large displacement peak in a period.

Discussion

Speech perception modeling often begins by assuming an input representation consisting of the outputs of a set of critical bandwidth analyzing filters (Zwicker & Feldtkeller, 1967; Klatt, 1979c). It is often convenient to employ energy detectors that average over about 10 msec so as to obtain an output representation that does not fluctuate over a period. Such a representation clearly cannot account for our phase results, but it is of some interest to determine the extent to which such a critical band filter bank produces magnitude spectra consistent with the remaining psychophysical distance estimates.

The commercially available third-octave analyzers are not suitable to this task because their filter skirts are not as sharp as the psychophysically determined auditory filter skirts (Paterson, 1976), and this is a serious shortcoming when analyzing harmonically rich sounds. Therefore, we have designed a digital spectral analyzer

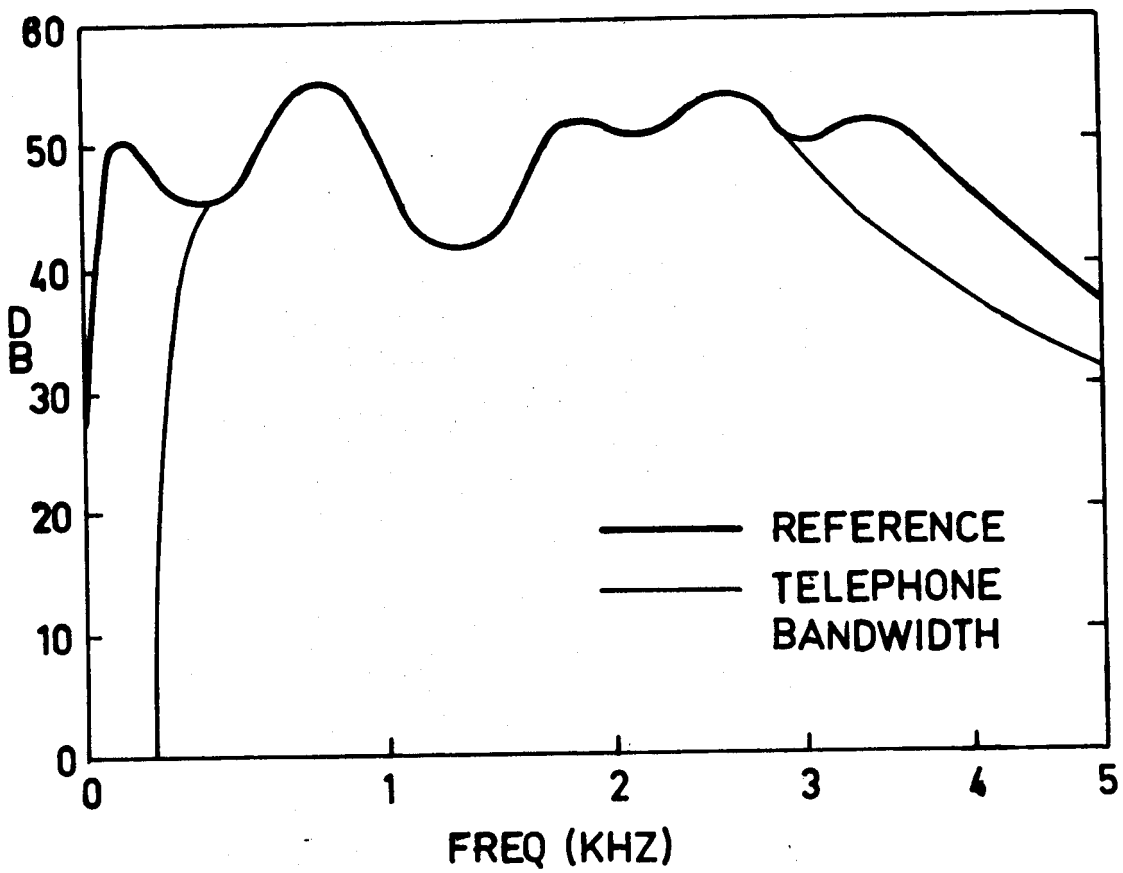

Fig. IV-B-2. Critical band spectra of two stimuli.

having filter parameters modeled after the Paterson data (Klatt, 1979b). Examples of stimuli analyzed by this filter are shown in Fig. IV-B-2.

We have examined the extent to which the area between two curves representing vowel stimuli to be compared (or area squared, Lindblom, 1979) correlates well with perceptual distance (after appropriate normalization for overall stimulus loudness as a separate factor contributing to distance (Klatt, 1976). The results are encouraging, although at least three departures from a single area calculation seem desirable:

(1) area discrepancies in a valley between two energy concentrations are less important,

(2) area discrepancies at high frequencies are less important than at low and mid frequencies, at least for vowels,

(3) area discrepancies at very low frequencies are more important than one might predict from the critical band representation.

Acknowledgments

Research supported in part by an NIH grant.

References

CARLSON, R. & GRANSTRÖM, B. (1976): "Detectability of changes in level and spectral slope of vowels", STL-QPSR 2-3/1976, pp. 1-4.

FANT, G. (1960): Acoustic Theory of Speech Production, Mouton, The Hague.

FLANAGAN, J.L. (1957): "Estimates of the maximum precision necessary in quantizing certain dimensions of vowel sounds", J.Acoust. Soc.Am. 29, pp. 533-534.

KLATT, D.H. (1976): "A digital filter bank for spectral matching", Proc. IEEE ICASSP, IEEE Catalog No. CH1067-8 ASSP, pp. 537-540.

KLATT, D.H. (1979a): "Documentation for additive harmonic synthesis using the Fortran program HARSYN", unpubl. memo.

KLATT, D.H. (1979b): "Documentation for critical band spectral analysis using the Fortran program PLOTS", unpubl. memo.

KLATT, D.H. (1979c): "Speech perception: A model of acoustic-phonetic analysis and lexical access", J. Phonetics 7, pp. 279-312.

LINDBLOM, B. (1979): "Phonetic aspects of linguistic explanation", Studia Linguistica, in press.

PATERSON, R.D. (1976): "Auditory filter shape derived with noise stimuli", J.Acoust.Soc.Am. 59, pp. 640-654.

ZWICKER, E. & FELDTKELLER, R. (1967): Das Ohr als Nachrichtenempfänger, Hirzel Verlag, Stuttgart.