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

# Model predictions of vowel dissimilarity

Carlson, R. and Granström, B.

journal: STL-QPSR
volume: 20
number: 3-4
year: 1979
pages: 084-104


KTH Computer Science and Communication

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

## C. MODEL PREDICTIONS OF VOWEL DISSIMILARITY*

R. Carlson and B. Granström

Abstract

Several models of the peripheral auditory system have been studied. As test material we have used perceptual data on psychoacoustic and phonetic dissimilarity (Carlson, Granström, & Klatt, 1979; Klatt, 1979b. The predictive value of the models depends on type of stimuli and perceptual task. Type of metric used in the dissimilarity calculation is of small importance for the correlation between predicted and perceived dissimilarity. We find a strong support for the view that some kind of peak picking mechanism is involved in speech perception. A model is presented that includes such a feature.

### Introduction

Lately we have seen an increasing interest in representations of speech sounds that are more related to human perception than the conventional spectral representations used in spectrograms or FFT spectral sections. Basic research has resulted in several models of (peripheral) auditory processing. The elaboration of Zwicker et al (1967) of the loudness density concept has become more or less standard. Other models, including effects of lateral inhibition, have been created elsewhere, e.g. in Leningrad (Karnickaya et al, 1973; Chistovich et al, 1979) and in Eindhoven (Houtgast, 1974). The development of methods for similarity rating of spectra has been the interest of many research groups, notably the group in Soesterberg (Plomp, 1970; Pols, 1970). In Stockholm, we developed a model of local dominant frequency (Carlson, Granström, & Fant, 1970) for which our interest has been renewed in view of the work by Young & Sachs (1979), who, through a measure of average localized interval rate, managed to preserve frequency peak information despite rate saturation and two-tone suppression.

The practical need for sound representation related to auditory processing has been acknowledged in many quarters. For example, in evaluating the distortions in speech transmission or coding systems there is a need for objective measures of degradation (Schroeder et al, 1979; Viswanathan et al, 1976). In 1974 Zagoruiko & Lebedev published results from speech recognition experiments that convin-

---

* This paper will be presented to the NAS meeting of the Acoustic Society of Scandinavia in Turku, Finland, June 1980. Some of the work described in the paper was presented at the Gotland 1979 workshop on "Vowels: production and perception", in Sweden, Aug. 1979.

cingly demonstrated the virtues of using psychoacoustically related recognition parameters. Recently Klatt (1979a) has discussed the use of physiologically related spectral representations in models for lexical access.

Lindblom, Lubker, & Pauli (1977) have used a psychoacoustic distance metric to quantitatively evaluate abnormal speech and the effects of various clinical treatments.

When correlating the physical distance between speech sounds processed through some model and subjective similarity judgements, a correlation coefficient between .80 and .90 is reported (Plomp, 1970; Bladon & Lindblom, 1980). This rather high correlation might, to some extent, depend on the stimulus material. Often the stimuli are taken from a homogeneous ensemble such as vowels produced by a single vocal tract, with the same glottal source and all subject to the same kind of filtering. It is difficult to compare the merits of different models since the reported results depend on stimulus inventory and the subjective task, to mention only two factors.

In this study we have tried to relate several existing models of peripheral auditory analysis to the outcome of psychoacoustical and phonetic similarity judgements on vowels that have been manipulated in several ways.

## Data on psychoacoustical and phonetic distances between vowel sounds

In order to establish the perceptual importance of different aspects of vowel sounds, we synthesized 66 different versions of a vowel similar to /ae/. The vowels differed in some or all of their formant frequencies and bandwidths, in their overall amplitudes and phase characteristics, in the spectral slope of the glottal source, and in the amount of filtering. Some filtering affected only the lowest harmonics, some only the highest. Combined low-pass and high-pass filtering approximated the filtering in telephone networks. Spectral notches between formants were also created in some stimuli. An exhaustive description of the stimulus set is given in Carlson, Granström, & Klatt (1979). The average results from the "psychoacoustical" distance ratings of that study are seen in Table IV-C-Ia.*

* Subjects were asked to take into account any difference between the vowels.

The same set of stimuli was run in an experiment using "phonetic" rather than psychoacoustic distance criterion (Klatt, 1979b).* The results from the latter experiment are found in Table IV-C-Ib. The effect of changing decision criterion is quite marked. In the phonetic judgements, formant frequency adjustments are clearly the most important. All other changes are, in fact, relatively unimportant. It is also interesting to note that the change in all formants is less important than the F1 & F2 change as opposed to the case when general psychoacoustical distance was the criterion.

| AVERAGE PSYCHOACOUSTICAL DISTANCE | |
|---|---|
| 1. RANDOM PHASE | 20.0 |
| 2. HIGH-PASS FILTERING | 15.8 |
| 3. F1-F5 VARIATIONS | 10.7 |
| 4. DB/OCTAVE SPECTRAL TILT | 9.6 |
| 5. F1 AND F2 VARIATIONS | 8.2 |
| 6. LOW-PASS FILTERING | 7.6 |
| 7. F1 VARIATIONS | 6.5 |
| 8. DB/HZ SPECTRAL TILT | 5.2 |
| 9. F2 VARIATIONS | 5.0 |
| 10. OVERALL AMPLITUDE | 3.4 |
| 11. ALL BANDWIDTH | 2.4 |
| 12. 2:ND FORMANT BANDWIDTH | 2.4 |
| 13. 1:ST FORMANT BANDWIDTH | 1.1 |

Table IV-C-Ia.

| AVERAGE PHONETIC DISTANCE | |
|---|---|
| 1. F1 AND F2 VARIATIONS | 8.6 |
| 2. F1-F5 VARIATIONS | 8.1 |
| 3. F1 VARIATIONS | 6.2 |
| 4. F2 VARIATIONS | 6.1 |
| 5. RANDOM PHASE | 2.9 |
| 6. ALL BANDWIDTHS | 2.4 |
| 7. LOW-PASS FILTERINGS | 2.3 |
| 8. DB/HZ SPECTRAL TILT | 2.0 |
| 9. 2:ND FORMANT BANDWIDTH | 1.7 |
| 10. HIGH-PASS FILTERING | 1.6 |
| 11. DB/OCTAVE SPECTRAL TILT | 1.5 |
| 12. OVERALL AMPLITUDE | 1.5 |
| 13. 1:ST FORMANT BANDWIDTH | 1.0 |

Table IV-C-Ib.

Six models to predict our perceptual data

We now have at our disposal sets of both perceptual and acoustical data. What kind of models could be used to predict the perceptual results? We have tried six models of the peripheral auditory system beginning with a simple spectrum analysis approach. It is important to note that no higher level analysis in the form of decision-making has been included in the models presented.

---

* Subject instructions: Rate only changes that tend to influence vowel identity, disregard changes associated with harshness, speaker identity, or transmission channel.

All models have as their first component an FFT analysis of the stimuli. A 512-sample hamming window and a time constant of 20 ms has been used in the calculations. The spectrum was sampled at 20 Hz or 0.13 Bark intervals. To avoid sampling errors of the speech wave, the spectrum used as input to the models is an average computed over the total vowel stimulus. Thus the partials in Fig. IV-C-2a have more or less disappeared.

Figs. IV-C-1 and IV-C-2 illustrate the output from the models. In Fig. IV-C-1a, a sinusoid is used as input, and in Fig. IV-C-2a, two vowels. The thin line is the reference vowel in our experiments and the thick line is one of the test vowels with lowered F1 (8%). The levels are calculated in such a way that the plotted energy is constant. Hence, a broadening of the bandwidths in Figs. IV-C-1b and IV-C-2b reduces the peak amplitude.

Model 1: A band-pass spectrum analysis model

In this model (called "FFT") we regard the auditory system as a simple spectral analyzer with a linear frequency scale and amplitudes represented in decibels. A bandwidth or interval of 200 Hz is used to calculate the summed output from each point along the frequency axis (see Figs. IV-C-1b and IV-C-2b).

Model 2: A Bark-band model

Psychoacoustical data and physical measurements of the basilar membrane reveal that a linear frequency scale is not suitable as a perceptual dimension. A more reasonable approach, the Bark scale, defined by Zwicker & Feldtkeller (1967) is used in Model 2, the Bark model. A summation of energy within each critical band, 1 Bark, is also made. This model is similar to the third octave band analysis used by Plomp (1970). Figs. IV-C-1c and IV-C-2c show the findings obtained with this model. It should be noted that no masking effects except that introduced by the critical band concept is included in this model.

Model 3: A model that includes masking effects

Masking effects are well known both in psychoacoustic and phonetic experiments. Kakusho et al (1971) and Chistovich et al (1979)

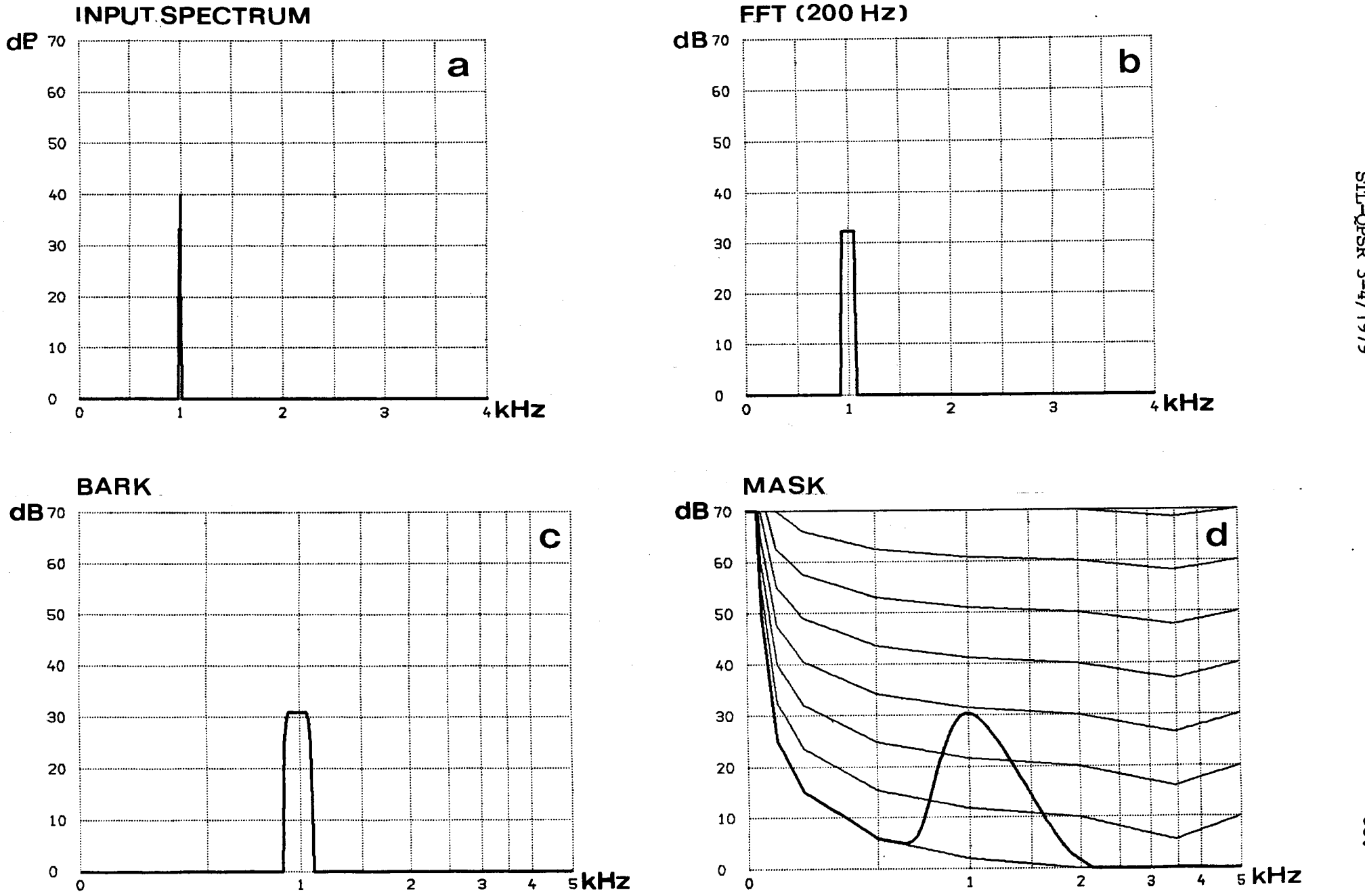


Fig. III-1(a-d) Representation of 1 kHz test tone in the models (see the text).

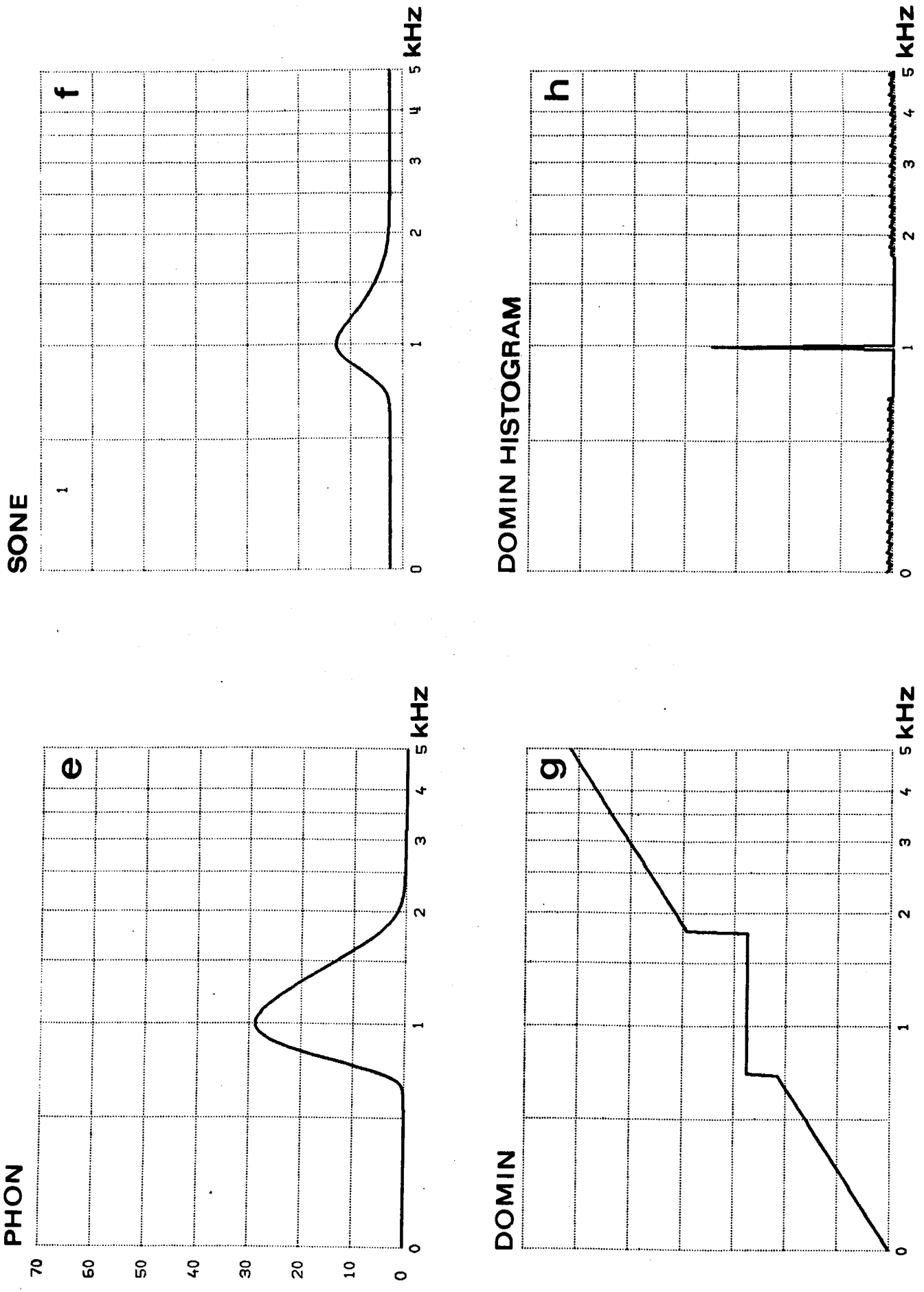


Fig. IV-C-1(e-h). Representation of 1 kHz test tone in the models (see the text).

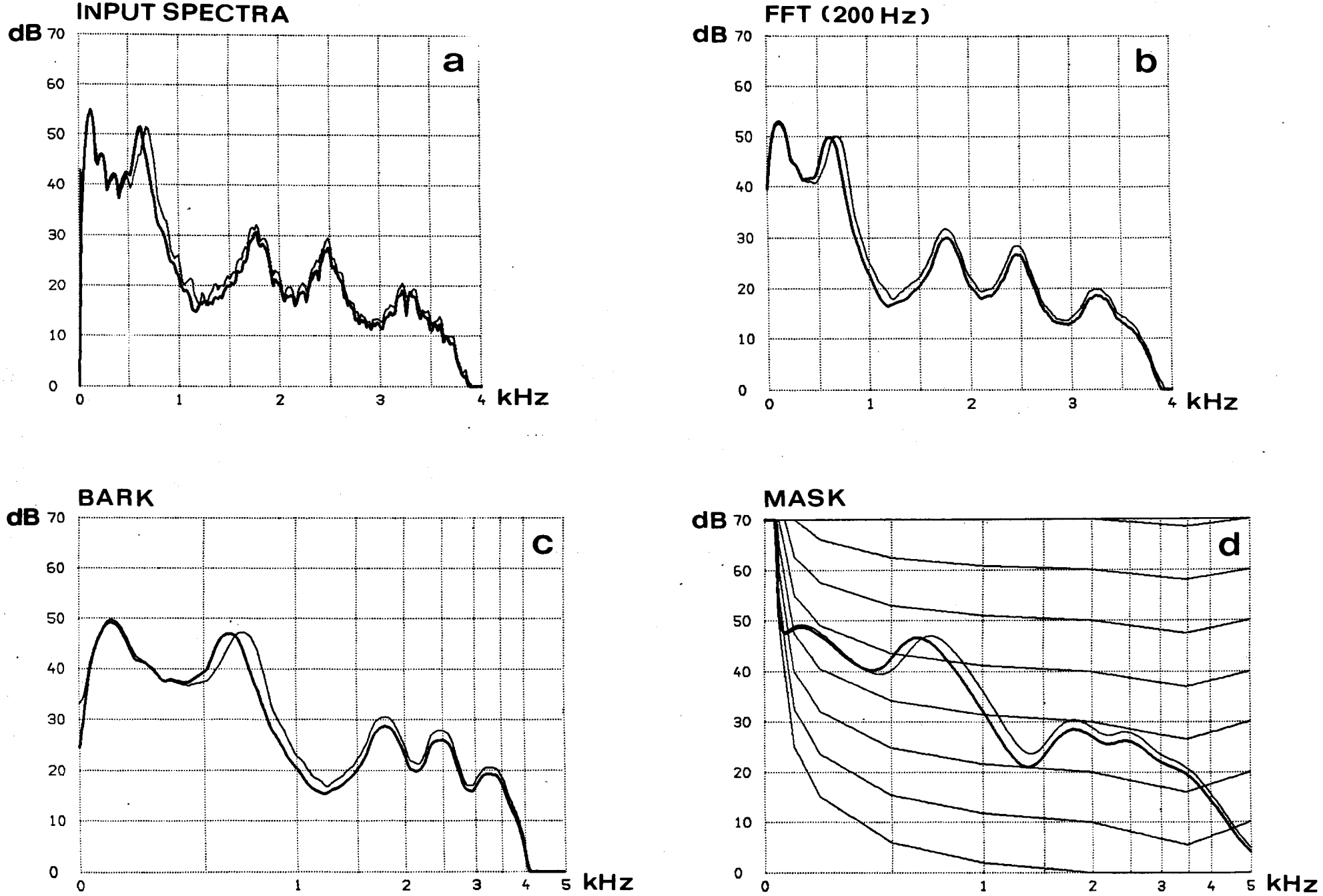


Fig. IV-C-2(a-d). Reference and test vowel in the models (see the text).

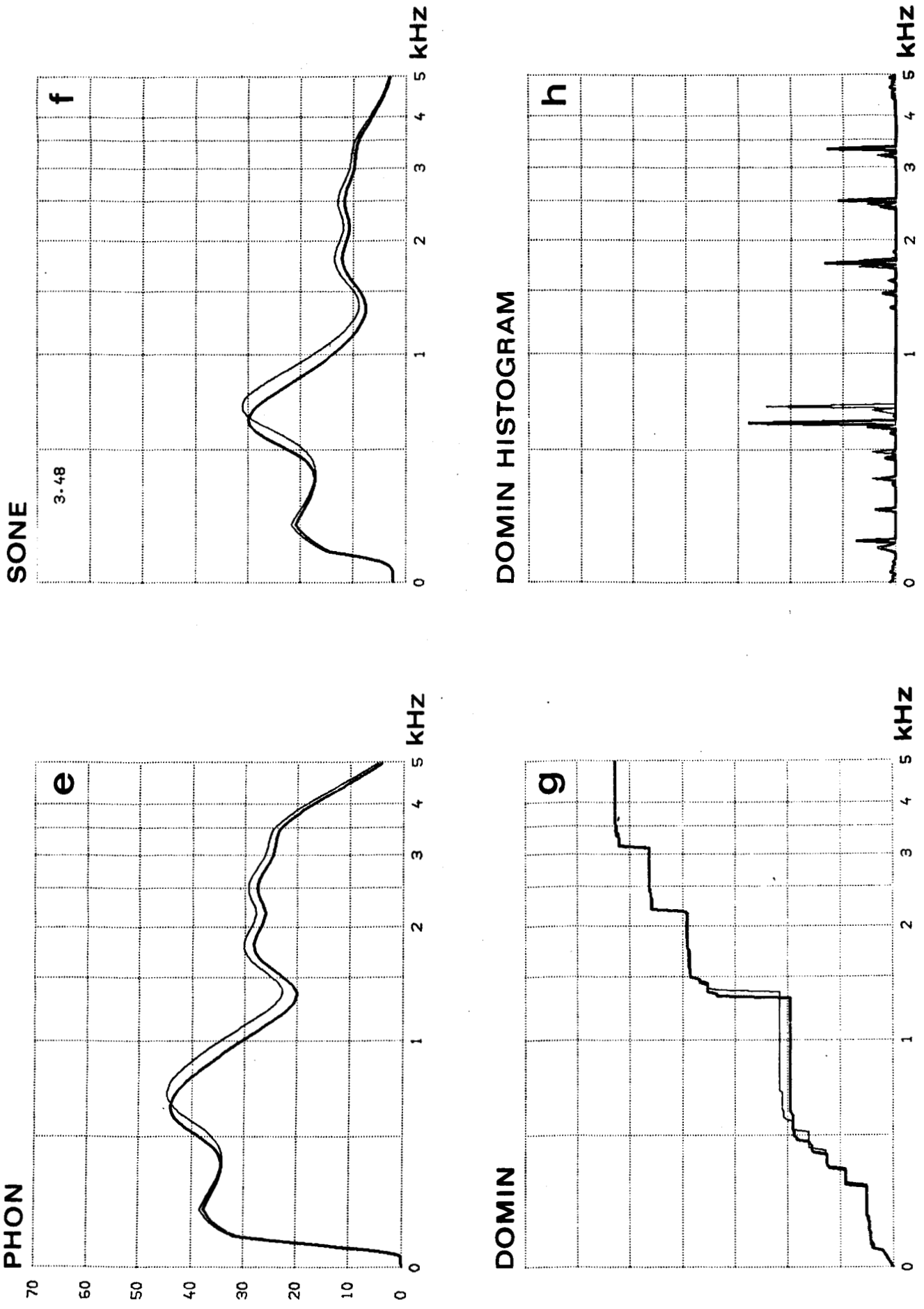


Fig. IV-C-2(e-h). Reference and test vowel in the models (see the text).

have shown how formant peaks mask the surrounding partials. Sometimes also a neighboring formant peak could be masked. Schroeder et al (1979) have proposed a masking filter to be used in evaluations of speech quality degeneration. This filter is meant to simulate typical results from Psychoacoustics. We have included this feature in one model, the Mask model (Figs. IV-C-1d and IV-C-2d).

Starting with the narrow band FFT, a convolution is made with the filter. Since the input signal in Fig. IV-C-1 is a sinusoid, the actual shape of the filter is shown in Fig. IV-C-1d. The auditory threshold is also included in the model. The plotted approximations of equal loudness curves are used in the transformation to the next model and have no meaning for the Mask model.

Model 4: A phon model

In this model, the Phon model, we take the next step by using the classical equal loudness curves (Fig. IV-C-1d). We use the output of Model 3 and transform it into a phon Bark space (see Figs. IV-C-1e and IV-C-2e).

Model 5: A sone model

The final step in this sequence of models is the Sone model. It makes a transformation from the phon/Bark space into the sone/Bark space, Figs. IV-C-1f and IV-C-2f. This is often regarded as being a good perceptual representation on which higher level processes and decision-making apply. The paper by Bladon & Lindblom (1980) explores this model.

Model 6: A local dominant frequency model

In our work with approximating vowels by using two-formant synthesis, we attempted to find an auditory model that could predict the outcome of these experiments (Carlson et al, 1970). A successful approach was to use fairly broad filters to simulate the basilar membran motion, and to make a zero-crossing analysis of the output from each filter.

Recent work by Young & Sachs (1979) on the representation of steady state vowels in terms of discharge patterns in auditory fibers

supports this approach. At low sound levels, well defined peaks in the discharge pattern could be found at formant frequencies. At high levels these peaks disappeared because of rate saturation and two-tone suppression. Analysis of the temporal pattern in separate neurons shows general phase locking to formant frequencies both at low and high levels. This supports the view that temporal analysis is involved in frequency perception.

Considering these results, it seems reasonable to reintroduce our old model. We have chosen to define it slightly differently this time.

Model 6, a local dominant frequency model, Domin, uses the filter discussed in Model 3 to find which frequency dominates each point along the "basilar membrane". The dominant frequency is plotted along the y-axis in Figs. IV-C-1g and IV-C-2g while the x-axis corresponds to the Bark scale. Fig. IV-C-1g shows how the sinusoid dominates the surrounding area while the auditory threshold dominates outside this area. The width is dependent on the shape of the filter. As could be seen in Fig. IV-C-2g, the formants form a staircase in this representation. It is obvious that formant shape and spectral notches have little importance for the output of this model.

A special and important feature of this model is that the degree of dominance can be quantified by combining the number of output channels that have the same dominant frequency. These channels can be interpreted as corresponding roughly to neurons. We have replotted the output from the model into a histogram representation, Figs. IV-C-1h and IV-C-2h. The dominant frequency is plotted along the x-axis and the degree of dominance along the y-axis. This has proved to be a useful representation both in the work by Young & Sachs and in our zero-crossing model.

Distance metric

The six models presented above generate two-dimensional patterns as outputs. Our next problem is to design a method to calculate the

the perceptual dissimilarity from two such patterns. Zwicker & Scharf (1965) have shown that the total loudness of a sound can be computed by a summation of the contribution of every critical band. Plomp (1970) hypothesized that the dissimilarity could be calculated by a summation of the level difference in each third octave band. The method is described by the following formula:

$$D_{ij} = \sqrt[p]{\sum_{n=1}^{m} \left| L_i,n - L_j,n \right|^p}$$

where

$D_{ij}$ = the distance between two stimuli, i and j
$L_i,n$= the level in band n
m = the total number of bands

If p equals 1 we get a city block model, i.e. a simple summation as in the model of Zwicker & Scharf. If p equals 2 the metric is Euclidean. We have employed this method to calculate the relative dissimilarity. In our case, the number of bands is the number of samples along the x-axis and the level is dB/Bark, phons/Bark, sones/Bark, or dominant frequency depending on the model.

## Correlation analysis

A correlation analysis between perceptual data and model prediction was conducted for all models and the two distance metrics. The phonetic and psychoacoustic data were handled separately. In Fig. IV-C-2a, a typical result is plotted. Perceptual psychoacoustic differences are plotted along the y-axis and model predictions along the x-axis. It could be questioned whether a straight line approximation is the best to our data. At small distances the perceptual threshold sets the limit, and with large distances, the response numbers tend to be constant. In our analysis we have not concerned ourselves with this problem and in our paper we have presupposed that a linear approximation is appropriate for our comparative purpose. The result of the correlation analysis can be seen in Table IV-C-II.

ALL VOWELS
PSYCHOACOUSTIC JUDGEMENTS

| | CITY-BLOCK | EUCL. |
|---|---|---|
| 1 FFT | .86 | .82 |
| 2 BARK | .85 | .74 |
| 3 MASK | .78 | .79 |
| 4 PHON | .78 | .76 |
| 5 SONE | .69 | .77 |
| 6 DOMIN | .77 | .74 |
| 7 DOMSON | .81 | |

Table IV-C-IIa.

ALL VOWELS
PHONETIC JUDGEMENTS

| | CITY-BLOCK | EUCL. |
|---|---|---|
| 1 FFT | .47 | .22 |
| 2 BARK | .23 | .09 |
| 3 MASK | .19 | .10 |
| 4 PHON | .13 | .05 |
| 5 SONE | .18 | .11 |
| 6 DOMIN | .69 | .64 |
| 7 DOMSON | .60 | |

Table IV-C-IIb.

FREQ.CHANGE VOWELS
PSYCHOACOUSTIC JUDGEMENTS

| | CITY-BLOCK | EUCL. |
|---|---|---|
| 1 FFT | .96 | .96 |
| 2 BARK | .95 | .94 |
| 3 MASK | .72 | .74 |
| 4 PHON | .72 | .75 |
| 5 SONE | .70 | .70 |
| 6 DOMIN | .94 | .88 |
| 7 DOMSON | .91 | |

Table IV-C-IIc.

FREQ.CHANGE VOWELS
PHONETIC JUDGEMENTS

| | CITY-BLOCK | EUCL. |
|---|---|---|
| 1 FFT | .62 | .66 |
| 2 BARK | .67 | .72 |
| 3 MASK | .67 | .72 |
| 4 PHON | .67 | .73 |
| 5 SONE | .67 | .73 |
| 6 DOMIN | .67 | .66 |
| 7 DOMSON | .70 | |

Table IV-C-IId.

By splitting the stimulus set into subgroups, we can examine whether a model makes any kind of systematic error. Figs. IV-C-3b-3e show such a separation. Only the stimuli that have a difference in formant frequency compared with the reference have been plotted in Fig. IV-C-3b. The bandwidth is the parameter in Fig. IV-C-3c, and amplitude or slope variations characterize the stimuli in Fig. IV-C-3d. The last figure, Fig. IV-C-3e, shows a more inhomogeneous group consisting of low-pass or high-pass and stop-band filtering between formants. These four kinds of plots have been used to derive Fig. IV-C-4, where general trends for subgroups have been plotted for each model.

Result of correlation analysis

The result of the correlation analysis is shown in Table IV-C-II. We can notice that the different models differ very little in the psychoacoustical experiment as compared with the phonetic experiment. Model Dowson is a special model based on the dominant frequency model but with the total loudness difference ($\Delta S$) as a separate parameter in the distance function.

$$D_{ij} = \sum_{n=1}^{m} \left| L_i,n - L_j,n \right| + K^* \left| \Delta S \right|$$

In Table IV-C-IIa the constant K has been optimized to give the highest correlation. The same constant is used in Table IV-C-IIb-d.

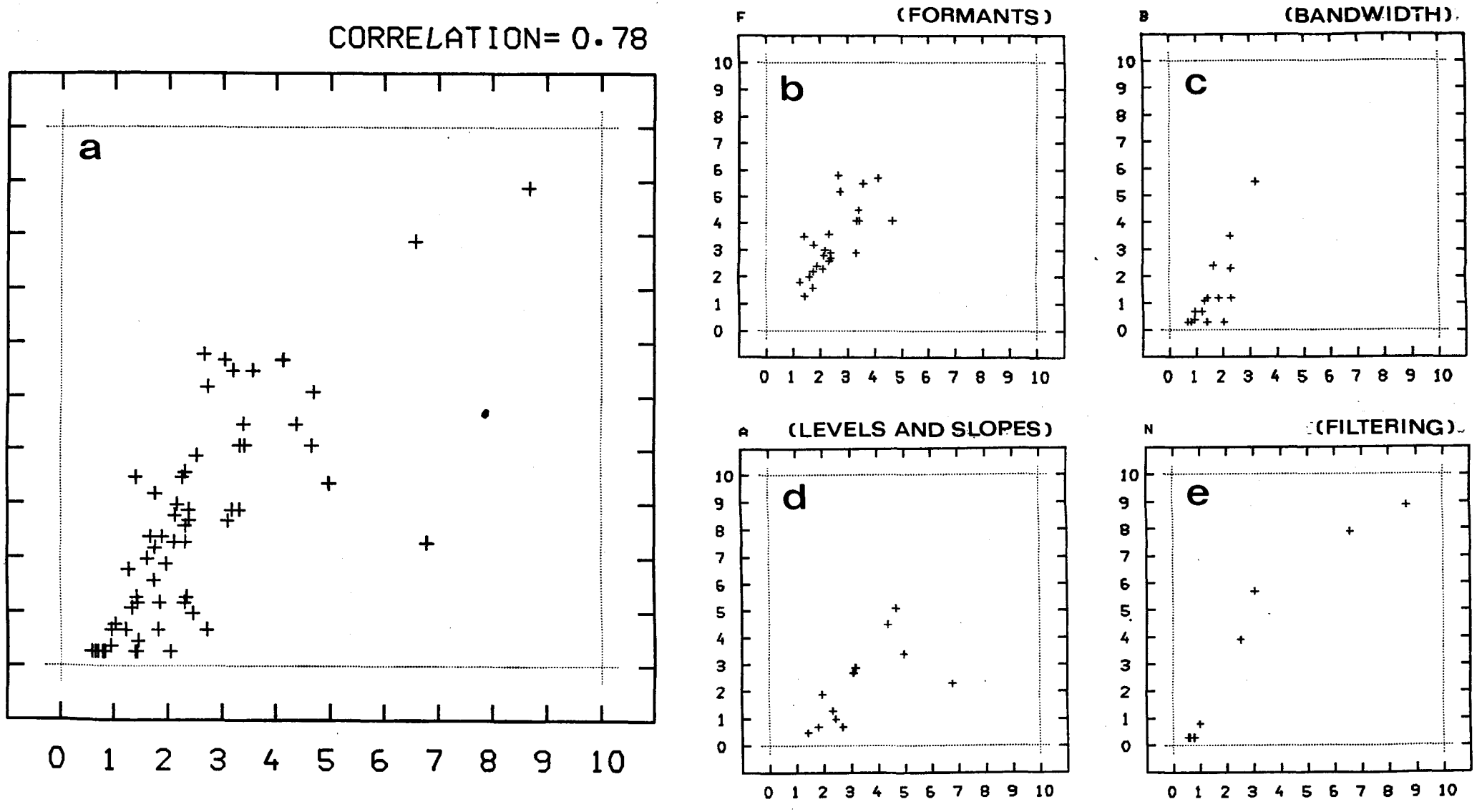


Fig. IV-C-3. Example of correlation of analysis of subjective (ordinate) and model (abscissa) data.

Evaluation of distance

In Table IV-C-II we present results for both Euclidean and city-block distances. Bladon & Lindblom (1980) found the Euclidean metric to be slightly superior for their data in a sone/Bark representation. Plomp (1970), using different kinds of sounds, also found some advantage for Euclidean distances when studying vowels in a dB/third-octave analysis. For other sounds, such as musical instruments, the correlation to a city-block measure, i.e. a simple sum of the distances in each frequency bands, correlated better to subjective data. Comparing the two columns in Table IV-C-IIa through IV-C-II$^{d}$, it is obvious that the merits of the different metrics depend on the model, the task, and the set of data. In our data, city-block is generally a better metric. The only situation where Euclidean distance is appreciably superior is for psychoacoustic criteria and the sone model. This is interesting to observe remembering that the loudness in each Bark band adds up linearly to a total loudness of the sound in the procedure elaborated by Zwicker & Feldtkeller (1967). Our data indicate that the same procedure is suboptimal in comparing loudness density patterns. Again the dependency on the type of stimuli is too important for it to be possible to draw firm conclusions. For phonetic judgements of formant frequency change, the Euclidean distance has some advantage. This situation is also closer to the situation in the Bladon & Lindblom experiment.

Model correlations with the psychoacoustical data

As a starting point for our discussion, let us take a look at Table IV-C-IIc. In this table we present correlation analysis of a subset of our data, namely formant variations. We get a high correlation for models 1, 2, 6, and 7. It is astonishing that the FFT-approach does so well, especially considering the fact that no knowledge from present-day ear model work has been included. One explanation could be that the maniuplations used are so small that non-linearities or other more sophisticated effects do not influence the result. The Domin model does give a good score since it has an inherent feature of extracting formant peaks.

If we take a look at Table IV-C-2a which includes all our data and compare the correlations for these models, we see a reduction of the scores by 10-15%. When different types of manipulations are included in an experiment, higher demands are put on a model. One could therefore question the degree to which perceptual experiments increase our understanding of the auditory system, if only one dimension, e.g., formant frequencies, is changed. Is it possible to start with a simple model like FFT-analysis and add special features to it, or do we have to employ a more general approach to enhance our knowledge? Irrespective of this methodological question we could conclude that models 1, 2, 6, and 7 behave rather well for our psychoacoustical data.

Let us once again turn back to Table IV-C-2c. Models 3, 4, and 5 have a lower correlation score of around 70%, which must be regarded as rather low. We have to reject these models as good predictors of our psychoacoustic distances. On the other hand, the correlations are not especially reduced when the total set of psychoacoustical data is included. This means that they present a different kind of performance compared to the FFT-model predicting the perception of more sophisticated changes than just formant variations. Needless to say, manipulations like filtering have a small effect on the output while amplitude differences have the same impact as for a simple spectral analysis model. This will be discussed in more detail later on in our paper.

Before turning to the discussion of phonetic distance, we would like to make some additional comments on the Masking model (Model 3). The filter shape used is an approximation by Schroeder et al (1979). This filter might not be optimal for our purpose. This could be one reason for the low scores in Table IV-C-2c. It is otherwise difficult to understand why the inclusion of such a filter should reduce the correlation compared to the Bark model. It should be noted that the shape of this filter has a consequence for the following models.

Model predictions of the phonetic data

Table IV-C-IIb gives the correlation between model prediction and phonetic distance. Only one model, Domin, presents a reasonably satisfactory result. All other models, for the most part, fail to give an accurate estimate. This result strongly supports the view by Chistovich et al (1979) that the perception of vowels includes some kind of formant or peak estimation algorithm and that levels and slopes have small importance. Even if this is included in the domin model, to some degree the correlation .69 is not especially good and we may turn to Table IV-C-IId to get some explanatory pointers. We see that all models present about the same low correlation, around 70%. Why do we not get a better result? To some extent it could be explained by the fact that the phonetic data are based on half the number of observations as the psychoacoustical. It might also be that it is more difficult to make a phonetic similarity judgement compared to a psychoacoustic one.

An interesting observation is that some changes have a more phonetic value compared to others of the same spectral magnitude. A change of all formant frequencies, for example, does not change the phonetic quality in proportion to the spectral change. Thus, we could suppose that higher level processes have been active in normalizing the perceived vocal tract length. This could clearly be seen in Table IV-C-Ib.

A related observation concerns the co-variation of F1 and F2. If we increase the first and second formant frequencies by the same percentage, we get a lower perceptual distance relative to the references than if the two formants are changed in opposite directions. This is not predicted by our models. Thus we have to conclude that our modelling of low level processes could not explain the perception of phonetic distances and that higher level processes must be included. We could only hypothesize that some representations are more relevant than others as intermediate patterns to be used in the decision process. In the next part of our paper we will discuss the general trends in the model predictions.

Model predictions according to subsets of stimuli

One reason for the rather low correlation between the subjective data and the model output is that different kinds of stimuli were used. If we perform the analysis on subgroups of the stimuli as described in relation to Fig. IV-C-3, there is generally a better within-group correlation. It is interesting to compare how the different models operate on these subgroups. In Fig. IV-C-4, the approximate regression lines for the different stimulus groupings are displayed. Since the scales are not readily comparable, only comparison of the relative slopes should be made. A steeper slope of a line indicates that the corresponding change is underestimated by the model. Let us first consider the relation to psychoacoustical distance (bold lines). As an example it could be seen that the simple FFT-analysis (Fig. IV-C-4a) tends to underestimate formant bandwidth changes and, to some extent, level manipulations compared to formant frequency modifications and filterings.

As a general trend in the models corresponding to Fig. IV-C-4a through Fig. IV-C-4e, levels and bandwidths are gaining more importance as the models are refined. In the sone representation, formant frequency, bandwidths and filtering have about the same weight. The rather poor overall correlation coefficient depends to some extent on a greater scatter within groups, but primarily on the model's overestimation of level changes. This is interesting to observe in relation to the study of Bladon and Lindblom who obtained the best correlation for a similar model but did not include stimuli with level variations.

In Fig. IV-C-4f, finally, it could be observed that, as could be expected, the local dominant frequency model underestimates bandwidth and level changes heavily and also, to some extent, filtering.

In Fig. IV-C-4, the thin lines represent the correlations between model and the subjective phonetic data. We noted generally low correlations in the overall correlation coefficients of Table IV-C-IIb. The reason for this seems from the analysis in Fig. IV-C-4a-f to be primarily a great difference between the four subgroups. The only moderately good correlation was found for the local

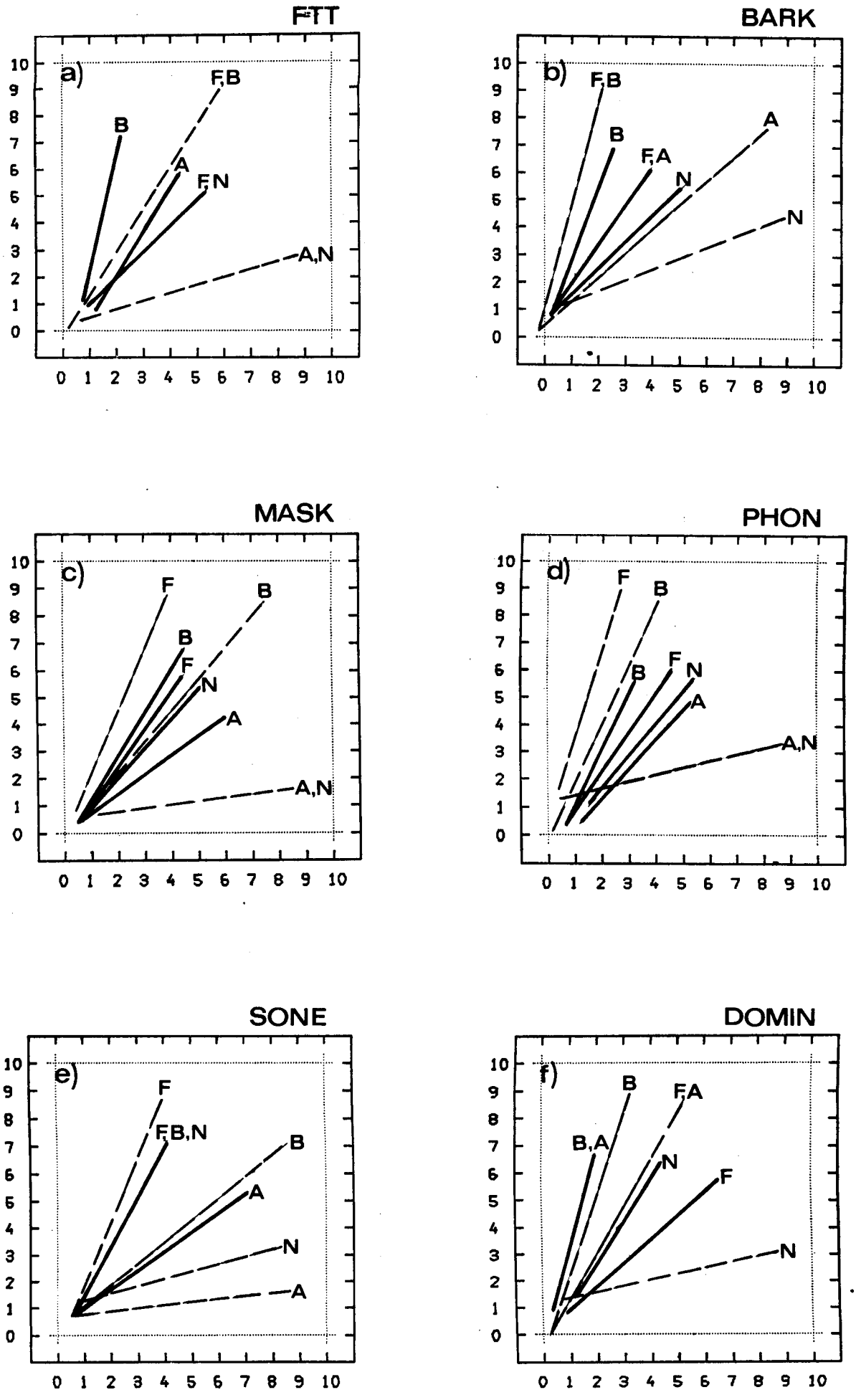


Fig. IV-C-4. Approximate regression lines for different subgroups (see the text).

dominant frequency model (Fig. IV-C-4f), the remaining problem being mainly that in this case the distance to filtered stimuli (N) are heavily overestimated by the model.

Concluding remarks

Our study has shown that the predictive value of the models we have tried, strongly depends on the subjective task and the set of stimuli used. We have failed to demonstrate a superiority for one of the more complex models. The sone/Bark representation proved to be of little use. Our stimuli were rather close in phonetic quality, and thinking of the subjective task as coming close to a discrimination task may aid in explaining why the sone/Bark model failed. From experiments on intensity perception (Rabinowitz et al (1975), we know that the sensitivity to intensity differences increases with level, but only mildly. The difference limen is more like a constant number of dB's or phons than a constant amount on the sone scale.

All models following the "Bark" model include filtering by an auditory filter given by Schroeder et al (1979). There is a possibility that this filter is too wide, considering the good results obtained by the FFT and Bark model and also alternative estimates of the "auditory filter". Experiments with a more narrow filter need to be performed.

The local dominant frequency model was the only one that scored reasonably well in both the psychoacoustical and phonetical tasks. Even if the recent work by Young and Sachs lends some plausibility to this kind of model, we had to draw the conclusion in an earlier study (Carlson, Fant, & Granström, 1975) that a similar model at least had to operate on a level where information from both ears is available. It is conceivable that the phonetic judgement includes more central processing. The set of stimuli mirroring a change in vocal tract length is a clear example of this.

Dealing with psychoacoustic experiments, it is always difficult to distinguish peripheral processes from more central decision

making. With this in mind, we still think it is worthwhile to pursue our experiments with the local dominant frequency model with a wider range of speech sounds, and also to study to what extent the present results are applicable to processing of human speech, which is notoriously dynamic.

Acknowledgments

We thank Dennis Klatt for valuable cooperation during our stay at MIT where this study was initiated.

During our work on auditory processing and vowel theory we have had fruitful discussions with Gunnar Fant, Jan Gauffin, Björn Lindblom, and Johan Sundberg.

Johan Liljencrants supplied the FFT package that forms the bases for the present models, which we gratefully acknowledge.

References

BLADON, R.A.W. & LINDBLOM, B. (1980): "Modeling the judgement of vowel quality differences", to be published.

CARLSON, R., FANT, G., & GRANSTRÖM, B. (1975): "Two-formant models, pitch and vowel perception", in Auditory Analysis and Perception of Speech (eds. G.Fant & M.A.A.Tatham), Academic Press, London, pp. 55-82.

CARLSON, R., GRANSTRÖM, B., & FANT, G. (1970): "Some studies concerning the perception of isolated vowels", STL-QPSR 2-3/1970, pp.19-35.

CARLSON, R., GRANSTRÖM, B., & KLATT, D.(1980): "Vowel perception: the relative perceptual salience of selected acoustic manipulations", pp. 36-44 in this issue of STL-QPSR.

CHISTOVICH, L.A., SHEIKIN, R.L., & LUBLINSKAJA, V.V.(1979): "'Centres of gravity' and spectral peaks as the determinants of vowel quality", in Frontiers of Speech Communication Research (eds. B.Lindblom & S.Öhman), Academic Press, London, pp. 143-158.

HOUTGAST, T.(1974): "Auditory analysis of vowel-like sounds", Acustica 31, pp. 320-324.

KAKUSHO, O., HIRATO, H., KATO, K., & KOBAYASHI, T.(1971): "Some experiments of vowel perception by harmonic synthesizer", Acustica 24, pp. 179-190.

KARNICKAJA, E.G., MUSHNIKOV, V.N., SLEPOKUROVA, N.A., & ZHUKOV, S.Ja. (1975): "Auditory processing of steady-state vowels", in Auditory Analysis and Perception of Speech (eds. G.Fant & M.A.A. Tatham), Academic Press, London, pp. 37-54.

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

KLATT, D.H. (1979b): "Perceptual comparisons among a set of vowels similar to /ae/: Some differences between psychophysical distance and phonetic distance", J.Acoust.Soc.Am. 66, pp. S86(A).

LINDBLOM, B., LUBKER, J., & PAULI, S. (1977): "An acoustic-perceptual method for the quantitative evaluation of hypernasality", J. Speech and Hearing Research 20, pp. 485-496.

PLOMP, R. (1970): "Timbre as a multidimensional attribute of complex tones", in Frequency Analysis and Periodicity Detection in Hearing (eds. R. Plomp & G.F. Smoorenburg), Sijthoff, Leiden, pp. 397-414.

POLS, L.C.W. (1970): "Perceptual space of vowel-like sounds and its correlation with frequency spectrum", in Frequency Analysis and Periodicity Detection in Hearing (eds. R. Plomp & G.F. Smoorenburg), Sijthoff, Leiden.

RABINOWITZ, W.M., LIM, J.S., BRAIDA, L.D., & DURLACH, N.I. (1978): "Intensity perception. VI. Summary of recent data on deviations from Weber's law for 1000-Hz tone pulses", J.Acoust.Soc.Am. 59, pp. 1506-1509.

SCHROEDER, M.R., ATAL, B.S., & HALL, J.L. (1979): "Objective measure of certain speech signal degradations based on masking", in Frontiers of Speech Communication Research (eds. B. Lindblom & S. Öhman), Academic Press, London, pp. 217-229.

VISWANATHAN, R., MAKHOUL, J., & RUSSEL, W. (1976): "Toward perceptually consistent measures of spectral distance", Proc. 1976 IEEE Int.Conf. on Acoustics, Speech and Signal Processing, pp. 485-488.

ZAGORUIKO, N.G. & LEBEDEV, V.G. (1974): "Models for speech signals analysis taking into account the effect of masking", Acustica 31, pp. 346-348.

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

ZWICKER, E. & SCHARF. B. (1965): "A model of loudness summation", Psychol. Rev. 73, pp. 3-26.

YOUNG, E.D. & SACHS, M.B. (1979): "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers", J.Acoust.Soc.Am. 66, pp. 1381-1403.