A FORMANT-BASED MODEL FOR PREDICTING PERCEPTUAL DISTANCES BETWEEN SWEDISH STOPS

Diana Krull
Institute of Linguistics
University of Stockholm

This paper constitutes a part of a larger investigation (Krull, 1988) whose aim is to evaluate the capacity of different perception models to predict listeners' responses. Such models are needed especially in automatic speech recognition.

The investigation is based on the notion of a perceptual space where distances are inversely related to perceptual similarity, that is, the shorter the distance, the greater the similarity. The perceptual distances should be predictable, given acoustic data and a good perception model.

To obtain a quantitative measure of the perceptual (dis)similarity, listening tests were performed using fragments of nonsense words of the form $V_1C:V_2$ where V=[I,ɛ,a,ɔ,U] and C=[b,d,ɖ,g] The tapes with the stimuli were presented to 20 listeners whose task was to try to identify the consonant. The greatest number of confusions were obtained using stimulus fragments of 26ms beginning at the moment of consonant release. The results were accounted for in 25 confusion matrices, one for each vowel context. Three examples of the matrices are shown in Fig.1, for a complete set of matrices see Krull (1987, 1988). The confusions displayed a regular pattern which was clearly dependent on the vowel context, especially the following vowel. For example, /g/ before a front vowel was often confused with a dental or a retroflex but seldom with a labial, when followed by a back vowel, however, /g/ was often confused with a labial but almost never with a dental or retroflex. The perceptual distance is thus short between velar and dental in front vowel context and long between velar and labial. In back vowel context the reverse is true.

I-I

	b	d	ɖ	g
b	93	3	2	2
d		97	3	
ɖ		58	40	2
g		25	28	47

a-a

	b	d	ɖ	g
b	98			2
d	2	75	23	
ɖ		15	83	2
g	5	15	2	78

ʊ-ʊ

	b	d	ɖ	g
b	85	5	5	5
d	3	37	30	30
ɖ	5	23	59	13
g	63			37

Fig.1 Three examples of confusion matrices

The listeners' confusions were now plotted as a function of the calculated distances. The resulting plot contained much noise in the form of over and underestimations of perceptual confusions. Part of these were due to asymmetries in the listeners' responses - the same calculated distance would thus be related to different percent confusions, as for example in the case of b→g and g→b in the /ugu/ stimulus (see Fig.1). An analysis of the asymmetries showed that they were dependent on the acute-grave dimension of the following vowel in a regular way and consequently predictable. The listeners' answers were therefore temporarily symmetrized by calculating the mean values of the confusions in the two directions. Secondly, there were differences in the formant-based distances depending on the following vowel: all distances were relatively short before front vowels, and longer before back vowels. Also, the calculated regression lines for percent confusions as a function of acoustic distance had a steeper slope with stimuli with front V_2 than with back V_2. Thirdly, in calculating distances, reference values with the same V_2 as the stimulus were used. Listeners may not have recognized the vowel and may have used references with another - for example neutral - vowel context.

Regression analyses were performed taking all these aspects into account. The calculations were performed separately for each of the three groups: stimuli with a front vowel, /a/ and back vowel as V_2. An additional listening test was performed, showing that V_2 could not always be recognized, especially, back vowels after a dental or retroflex stop were perceived as front or neutral vowels. Experimenting with different reference values showed that those with V_2=/a/ gave the best correlation to perceptual confusions except for labial and velar stimuli with back V_2 where the vowel could be clearly recognized. In these cases references with the same V_2 as the stimulus were used.

The results now showed that for stimuli with V_2=/i,e/ or /a/ the confusions diminished with distance up to about 2 Bark and remained unaffected by further increasing distance. With back V_2 too, large distances gave few confusions. However, there were several labial-velar pairs which were relatively seldom confused in spite of a short distance between them - listeners must have used some additional cue(s). A possible cue in this case was the length of the noise segment after consonant release: it is known to be long for velars stops and short for labials. The noise burst was therefore measured for all stimuli and the differences in its length between stimulus and reference were calculated. Thereafter, these differences were included in the distance measure using the equation

$$Dm_{i,j}=\sqrt[p]{(w_1*Df_{i,j})^p+(w_2*Db_{i,j})^p} \qquad \text{Eq.(3)}$$

where Dm is the modified distance, Df the formant based distance and Db the difference in burst length, i and j different stimuli, w_1 and w_2 different weighting factors and p a variable. In this case, $w_1=1.0$, $w_2=.1$ and p=2 gave the best results. (That means that 10ms was given about the same perceptual weight as 1 Bark.)

This new distance measure gave markedly better correlations between acoustic distance and percent confusions for stimuli with back V2, and a slight improvement for the other stimuli (Fig.3).

What acoustic properties correspond to these perceptual distances? I found a possible example of such properties in Fant (1973) where F_2 and F_3 at CV-boundaries were plotted against each other, C in this case was /b,d,g/ followed by nine different vowels, the syllables were read by a male Swedish speaker. The F_2-F_3 points for /g/, for example, were near those of dentals and retroflexes in front vowel context, and near labials in back vowel context. A corresponding plot with the material of the present investigation showed similar acoustic distances (Fig.2). However, the overlap between F_2-F_3 points for dental and retroflex consonants appeared to be too great in relation to the confusions between these two places of articulation, therefore F_4 was added - the frequency of this formant constitutes an important difference between the dental and the retroflex place of articulation. The formant frequencies had been measured in Hertz. In order to get distances better corresponding to what is received by the human ear, the frequencies were first converted to Bark, using the equation from Traunmüller(1983)

$$z=(26.81*f/(1960+f))-.53 \qquad \text{Eq.(1)}$$

where f is the frequency in Hz and z the value in Bark corresponding to that frequency. Thereafter, the Euclidean distances in the three-dimensional formant space were calculated according to the equation

$$D=\sqrt{(\Delta F_2)^2+(\Delta F_3)^2+(\Delta F_4)2+)^2} \qquad \text{Eq.(2)}$$

where F_n is the critical band rate in Bark of the n-th formant.

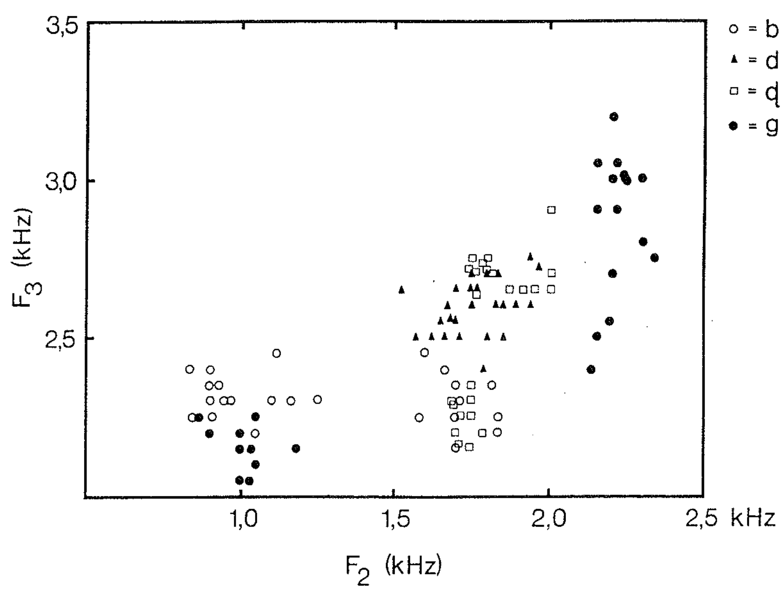

Fig.2 F_2 and F_3 at the CV-boundary.

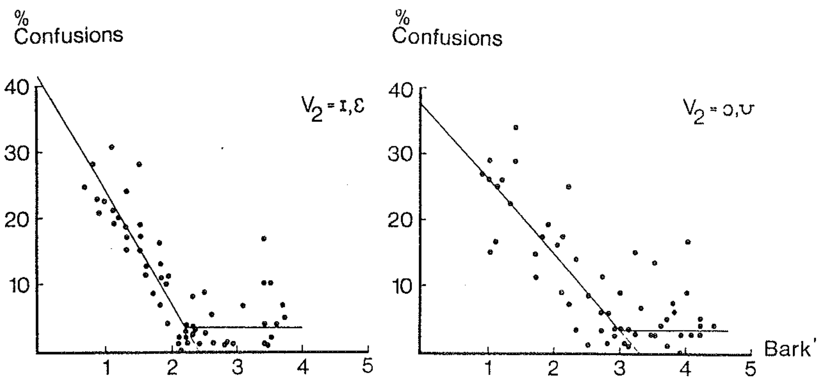

Fig.3 Percent confusions as a function of calculated distances based on formant frequencies in combination with differences in burst length (Bark′).

Finally, the asymmetries in the listeners′ confusions were calculated using formant values (Bark) at the CV-boundary and in the middle of V_2 as predictors (see Krull, 1988 p.108 for details), and percent confusions for each stimulus was predicted. The correlation coefficient between the predicted and observed percent confusions was r=.85. The predictions were calculated separately for the V2 contexts front, /a/ and back. Normalizing the V_2-dependent differences in the formant-based distances (Krull,1988, p.92f) made it possible to calculate percent confusions also for all vowel contexts together, which resulted in about the same correlation between the predicted and observed values, r=.86.

The formant-based model is an improvement compared to the spectrum-based one described in Krull (1987). Now percent confusions can be predicted not only for mean values for each V_2-context but also for single stimuli. In particular, the asymmetries in the listeners′ answers can be predicted. It is also more satisfactory to be able to perform the calculations for stimuli with all vowel contexts together.

REFERENCES

Fant, G.(1973), "Stops in CV-syllables", in G. Fant: Speech Sounds and Features, MIT Press

Krull, D.(1987), "Spectrum and dynamics in the perception of stop consonants", Papers from the Swedish Phonetics Conference, RUUL 17, Department of Linguistics, Uppsala University

Krull, D. (1988), Acoustic Properties as predictors of perceptual responses, PERILUS VII, Institute of Linguistics, University of Stockholm.

Traunmüller, H. (1983), "Analytical expressions of a tonotopical sensory scale", (Part of a doctoral dissertation), Inst. of Linguistics, University of Stockholm