EVALUATION OF DISTANCE METRICS USING SWEDISH STOP CONSONANTS

Diana Krull

Institute of Linguistics University of Stockholm
Department of Phonetics S-106 91 Stockholm Sweden

ABSTRACT

In recognition algorithms and certain theories of speech perception the process of signal interpretation is modeled in terms of distance metrics comparing the signal with stored references. In order to evaluate such metrics, listening tests were performed. The stimuli were short (about 26ms) fragments derived from the consonantal release of Swedish $\grave{V}_1C:V_2$ "words". A stop (b,d,ḍ,g) appeared in a systematically varied context of phonologically short vowels (i,e,a,o,u). The test yielded confusions which appeared to make qualitative sense in terms of the acoustic properties of the stimuli.

The spectrum level of the stimuli was measured at two time points after the stop release. Euclidean distances were calculated using spectra derived by means of 1/4 octave filter analyses. Two kinds of distances were calculated: static, based on spectra sampled at the first time point, and dynamic, based on the differences in spectral change between the two samoling points. Linear regression analyses performed on symmetrized percent confusions versus stimulus-reference distance produced correlation coefficients of -.85 (static), -.83 (dynamic), and -.92 (static and dynamic combined.)

INTRODUCTION

This investigation is based on the conception of a perceptual space for speech sounds where the distance between different sounds reflects the degree of their perceptual similarity. The greater the similarity between two sounds, the smaller the distance between them. Similar sounds tend to be confused with each other, therefore the number of confusions between sounds can be used as a measure of their perceptual distance. A further assumption is that correct identification of a sound indicates minimal distance from a stored reference.

For both theoretical and practical reasons, it is often desirable to be able to predict perceptual similarity from acoustic data. Such predictions are important especially in automatic speech recognition. To implement such a model, it is necessary, on the one hand, to find a realistic transformation of the speech signal, e.g. in terms of a realistic auditory model, and, on the other hand, an empirically calibrated distance metric.

ELICITATION OF PERCEPTUAL CONFUSIONS

The aim of this study is the evaluation of such a prediction model for Swedish voiced stops. It has been shown for Swedish /1/ that there are considerable coarticulation effects for such stops in intervocalic position. To make use of these effects, stimuli of the form $\grave{V}_1C:\acute{V}_2$ were prepared, where the consonant was [b,d,ḍ,g] and the vowel [ɪ,ɨ,a,ɔ,ʊ]. The resulting one hundred nonsense words were read in random order by a male speaker of the Central Swedish dialect. The Swedish grave accent was used in order to give both syllables about equal prominence.

From these "words" shorter stimuli were prepared by cutting out ca 26ms long segments beginning at consonant release. For simplicity, these stimuli will henceforth be referred to as "Burst" although they can contain also the beginning of the vocalic transitions. Notwithstanding the fact that the duration of the noise burst varies with place of articulation, all stimuli were given the same length in order to avoid letting stimulus length constitute an extra place cue.

A tape was prepared where each "Burst" stimulus appeared three times. The order of the stimuli was randomized. 20 native speakers of the Central Swedish dialect listened to the tape, their task being to identify the consonant.

The results of the perception test are shown in 25 confusion matrices, one for each vowel context (Fig.1). In each row of matrices the preceding vowel changes from front to back while for each column of matrices it is the following vowel that changes in the same manner. Comparing the results by vowel contexts and consonants,

PERCENT ANSWERS

STIMULI

	I-I				ε-I				a-I				ɔ-I				ʊ-I			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	93	3	2	2	98		2		82	3	7	8	67	8	2	23	86	10	2	2
d		97	3		2	88	8	2	13	70	17		3	67	28	2	7	87	3	3
ɖ		58	40	2	2	33	63	2		32	68		3	33	62	2	·	42	58	
g		25	28	47	8	32	32	28	5	28	23	44	10	32	15	43	2	53	22	23

	I-ε				ε-ε				a-ε				ɔ-ε				ʊ-ε			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	98	2			90	2	5	3	90	2	8		91	2	5	2	89	3	5	3
d		87	10	3		78	20	2	5	57	28	8	2	76	20	2		87	13	
ɖ		32	68			20	80			28	70	2	2	28	65	5		25	73	2
g	8	38	22	32	17	35	13	35	2	27	18	53	7	23	27	43	10	20	15	55

	I-a				ε-a				a-a				ɔ-a				ʊ-a			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	97			3	97			3	98			2	91	2		7	97		3	
d	2	75	18	5	20	61	17	2	2	75	23		8	66	23	3	10	63	27	
ɖ		40	52	8	2	28	68	2		15	83	2	15	7	75	3		13	85	2
g	3	15	12	70	3	18	23	56	5	15	2	78	10	17	13	60	28	23	13	36

	I-ɔ				ε-ɔ				a-ɔ				ɔ-ɔ				ʊ-ɔ			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	90	2	5	3	-	-	-	-	92		5	3	91	2	5	2	85	2	5	8
d	3	48	37	12	3	55	35	7	5	40	50	5	2	55	40	3		48	35	17
ɖ	7	15	75	3		10	88	2	3	8	89		2	18	80		13	15	62	10
g	15			85	13		3	84	25		7	68	37	2	3	58	42		2	56

	I-ʊ				ε-ʊ				a-ʊ				ɔ-ʊ				ʊ-ʊ			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	92		3	5	92		5	3	98			2	96		2	2	85	5	5	5
d		63	32	5	8	57	17	18	5	76	17	2	10	42	25	23	3	37	30	30
ɖ	2	20	78			18	79	3	33	13	32	22	13	8	74	5	5	23	59	13
g	18	2	2	78	32			68	27		5	68	12		2	86	63			37

Fig.1. Confusion matrices for "Burst" stimuli in 25 vowel contexts.

it can be seen that the confusions form a regular pattern. For example, [g] in front vowel context was often confused with the dental and the retroflex, but seldom with the labial. In back vowel context, on the other hand, the velar was often confused with the labial but almost never with the dental or the retroflex. The consonants seem to have been easiest to identify in the context of /a/. The influence of the preceding vowel was less pronounced than that of the following one. (For more details see /2/). Perceptually, the distance between the velar and the dental is thus small in front vowel context and large in back vowel context, while the reverse is true for the pair labial-velar.

USING PHYSICAL DISTANCE MEASURES TO PREDICT THE PERCEPTUAL CONFUSIONS

A qualitative comparison of stimulus spectra showed that there are pronounced coarticulation effects and, also, that these can have influenced the direction and number of the confusions. With such effects in mind, three models were chosen for defining the acoustic distances to be correlated to the perceptual confusions. The first model was based on formant frequencies at the moment of consonant release, and the second on sone-Bark spectra /3/.

The third model was based on bandpass filtered spectra sampled at two points in time: t_1, integrated over the first 10ms after consonant release, and t_2, 10ms later. The measurements were carried out with 14 digital 1/4 octave filters, covering a frequency range from about .4kHz to about 4.5kHz. The measured sound pressure levels were plotted as a function of frequency. The resulting spectra showed similarities and differences not only according to the place of articulation of the consonant but also according to the following vowel, thus forming 12 groups: labial, dental, retroflex, and velar stops read in in the context of a following front vowel, /a/, and back vowel. Differences within groups being small, mean values were calculated for each group, both at t_1 and at t_2. The t_1 spectra were normalized with respect to their mean SPL in order to avoid including differences in overall intensity into the distance measure. Two examples of the resulting spectra are shown in Fig.2.

Distances between spectra were then calculated for t_1, the result was called "static" distance, using the Euclidean metric:

$$\mathrm{Dstat}_{i,j} = \sqrt{\sum_{n=1}^{14} \left| L_{i,n} - L_{j,n} \right|^2} \qquad \text{Eq(1)}$$

$\mathrm{Dstat}_{i,j}$ = the distance between stimuli i and j at time t_1
$L_{i,n}$= the level in band n

The changes in spectrum level that occur after stop release show characteristic differences with place of articulation. These dynamic differences have in recent years been investigated especially in connection with the question of acoustic invariance for stop consonants /4/.

Comparing the change in spectrum level of the twelve spectra, it could be seen that at low frequencies the spectrum level rises during the interval between t_1 and t_2 for all spectra, and is comparatively steady at 1.5kHz. It is at frequencies above 1.5kHz that the amount and direction of the change varies in a systematic way: before front vowels the level goes up for the labial, remains unchanged for the dental, and drops for the retroflex and the velar. Before /a/ the level also rises for the labial, but in contrast to the front vowel context, the level drops for both the dental and the retroflex but is stable for the velar. In back vowel context the level

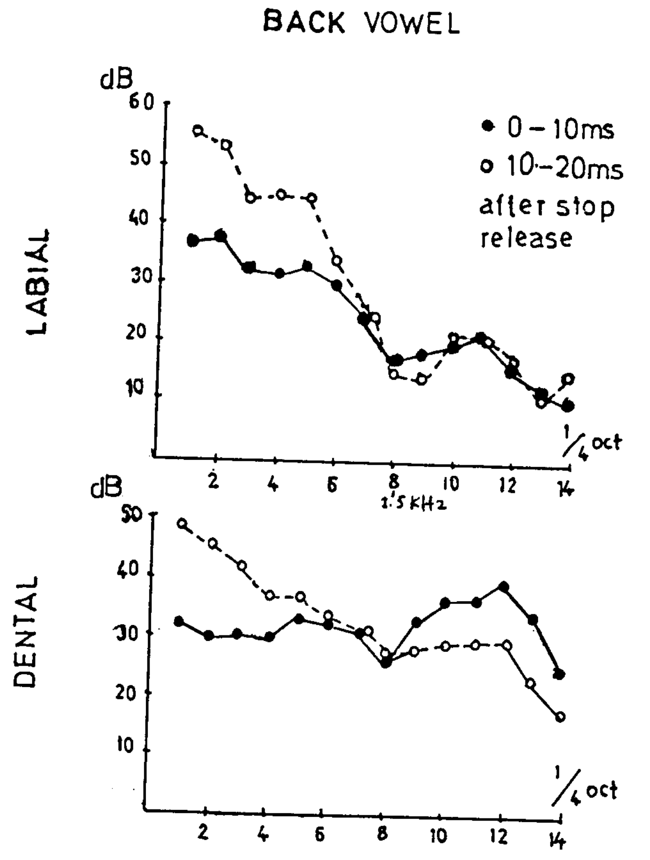

Fig.2. Examples of spectra sampled at two time points after stop release. 1/4 octave band-pass filters were used.

remains stable for the labial, while dropping with all other places of articulation, although the drop is comparatively small for the velar. It thus seems that although the change in level is dependent on place of articulation, the following vowel must be taken into account too. There tends to be less change if the spectra of the consonant and the following vowel are relatively similar as is the case for the dental and front vowels, for the velar and /a/ and for the labial and back vowels.

The dynamic distances were calculated in a similar way as the static ones with the help of the Euclidean metric, but on non-normalized spectra and only for the six filter bands above 1.5kHz, that is in the frequency range where there were systematic differences between groups:

$$\mathrm{Ddyn}_{i,j} = \sqrt{\sum_{n=1}^{6} \left| C_{i,n} - C_{j,n} \right|^2} \qquad \text{Eq(2)}$$

where $\mathrm{Ddyn}_{i,j}$= difference in level change in dB from t_1 to t_2 between stimuli i and j
$C_{i,n}$= level change for stimulus i, band n

Before performing regression analyses correlating acoustic distances and perceptual confusions, the results of the perception test were manipulated in two ways: first, the answers were divided into 12 groups in the same way as the spectra and mean values were calculated for each group; second, the answers were symmetrized according to a method described by Klein, Plomp, and Pols /5/. The regression analyses were then calculated between the symmetrized confusion data and three kinds of acoustic measures: (1) static, i.e. difference between spectra at t_1; (2) dynamic, i.e. difference in the amount and direction of change in two spectra; (3) static and dynamic distances combined according to the equation

$$D_{i,j} = \sqrt{(\mathrm{Dstat}_{i,j})^2 + (\mathrm{Ddyn}_{i,j})^2} \qquad \text{Eq(3)}$$

where $D_{i,j}$= combined static and dynamic distance between stimuli i and j

The resulting correlation coefficients are shown in the table below.

r(t1) -- static:			
	Front vowel	/a/	Back vowel
	-.78	-.93	-.55
r(t2-t1) -- dynamic			
	Front vowel	/a/	Back vowel
	-.78	-.94	-.14
static + dynamic			
	Front vowel	/a/	Back vowel
	-.80	-.98	-.58

It can be seen that good predictions can be made only for consonants before the vowel /a/. The results were especially negative for the back vowel context. What could be the reason for this? A possible answer could be that the listeners, if they could not recognize the following vowel, used a strategy somewhat different from that assumed here. Even if we are correct in assuming that a comparison of the stimulus with a stored reference does indeed take place in the listeners' processing, we might be wrong in supposing that the stored reference is the spectrum actually associated with the specific VCV word from which the stimulus had been derived. Conceivably, a given stimulus might lead the listener to postulate a "neutral" vowel reference spectrum from a context in cases where cues for V_2 were weak or absent. In order to obtain information on these questions, an additional test was carried out with the "Burst" stimuli using eight subjects, their task now being to identify the vowel. The results showed, first, that a back vowel could be identified only after a labial or velar consonant. After a dental or a retroflex listeners heard either a front or a neutral vowel. When the original vowel was a front vowel or /a/, listeners either made few errors or heard a neutral vowel.

With the above considerations and the preceding results in mind, acoustic distances for all stimuli (except labials and velars before back vowels) were calculated using consonants read before /a/ as references. The new correlation coefficients are shown below.

r(1) -- static

Front vowel	/a/	Back vowel	Contexts pooled
-.89	-.93	-.96	-.85

r(t2-t1)--dynamic

Front vowel	/a/	Back vowel	Contexts pooled
-.94	-.94	-.72	-.83

static+dynamic

Front vowel	/a/	Back vowel	Contexts pooled
-.96	-.98	-.89	-.92

References

/1/ Öhman, S. (1966): "Coarticulation in VCV utterances: Spectrographic measurements", JASA 39(1), 151-168

/2/ Krull, D. (1984): "The role of vowel context on the perception of place of articulation of stops", PERILUS, Report III, University of Stockholm

/3/ Krull, D. (1985) "On the relation between the acoustic properties of Swedish voiced stops and their perceptual processing", PERILUS, Report IV, University of Stockholm

/4/ Kewley-Port, D. (1983): "Invariant cues for place of articulation in stop consonants", JASA, 73(1), 322-335

/5/ Klein, W., Plomp, R., and Pols, L.W.C. (1970): "Vowel spectra, vowel spaces and vowel identification", JASA 48(8), 999-1009