AN EXPERIMENT ON THE CUES TO THE IDENTIFICATION OF FRICATIVES

HARTMUT TRAUNMÜLLER DIANA KRULL

Institutionen för lingvistik
Stockholms Universitet
S-106 91 Stockholm

ABSTRACT

Synthetic fricatives with two spectral peaks scanning a wide range of frequencies were put into three versions of the context <a_'ɛ:>, also generated synthetically, and imitating a male speaker (1), a child (2), and an aroused male speaker (3) with elevated F_0 and F_1. The stimuli were presented in two orders, with increasing or decreasing frequencies of the spectral peaks, to 16 speakers of Swedish who identified the fricatives as <f>, <s>, <ç>, <ʂ>, or <ɧ>. In a given context, the obtained phonetic boundaries followed mainly the spectral peak lowest in frequency, while the upper peak contributed only marginally even if it was at a distance less than the "critical distance" of about 3 Bark. In context (2), as compared with (1), the phonetic boundaries were shifted up, but less (in Bark) than the vowel formants.

INTRODUCTION

It is well known that the characteristic frequencies, i. e., the frequencies of the formants and the fundamental in speech sounds with a given phonetic quality vary with the overall dimensions of the speaker's vocal tract. If the characteristic frequencies of vowels are converted into a measure of tonotopical place, such as critical band rate (Bark), differences in speaker size can be seen to correspond to a tonotopic translation of the auditory pattern of excitation <11>.

Identifications of synthetic two-formant vowels revealed that a uniform tonotopic compression of the auditory pattern of excitation with a fixed point in the region of F3 also preserves phonetic quality <12>. Natural vowels are transformed in this way in shouting and in whispering <11>.

The present investigation is about the transformations the spectra of voiceless fricatives can be subjected to without affecting their phonetic quality. It is known that voiceless fricatives can be synthesized satisfactorily with two resonances and one antiresonance and that the cues to the phonetic identity of voiceless sibilants reside mainly in the stationary part of their spectrum, while the transitions are more important for non-sibilants <5, 7>. One-parameter sibilants can be synthesized using a resonance and an antiresonance one octave lower in frequency <5>. Such sibilants lack intrinsic cues to speaker size. In spectrogram reading, the Swedish voiceless sibilants can be distinguished by the frequency of spectral energy onset while there is more variation, even within the same speaker and context, in the detail above that frequency <6>. A second characteristic spectral peak can, however, often be discerned and one question we address here is whether this second peak is used to normalize for speaker size. We also investigate in how far a vocalic context can serve this purpose.

METHODS

Subjects

The experiments were conducted with a group of 20 native and 6 non-native speakers of Swedish, all employees or students at the Institute of Linguistics at Stockholm University. None of them reported auditory handicaps and all were familiar with the phonetics of Swedish, possessing /f/, /s/, /ɕ/, and /ʃ/. We report here the results of 16 native speakers with uniform behavior, mostly speakers of the local variety with the distributional allophones <ʂ> and <ɧ> for /ʃ/, but including three speakers of southern varieties, who had no <s> in their own speech.

Stimuli

The stimuli were synthetic VCV sequences. The vocalic segments had been obtained by synthetic imitation of a natural <a's:ɛ:>, produced by a male speaker of Swedish (Stockholm variety). A three parameter voice source <3> signal in accordance with that utterance was generated by the procedure described in <12>. The vocalic as well as the fricative segments were generated in serial synthesis by use of a block diagram simulating program (sampling at 16 kHz, 16 bit/sample). Eight vowel formants were used. Their bandwidths obeyed the standard relation $B_i = 0.05\ F_i + 50$ Hz.

The fricatives were generated by feeding white noise through a high-pass and a low-pass resonance filter, both of second order and with Q=10. The two resonance frequencies F_l and F_h were varied in steps of a factor $4^{1/9}$ (approx. 1.0 Bark). 42 combinations of F_l and F_h were used to scan the auditory space as shown in Figure 1. The fricatives had a duration of 0.20 s and the intensity onset and offset of the natural <s> was also imitated.

A second version of the vowel context was obtained by a uniform translation of all vowel formant frequencies by + 2.5 Bark. The voice source parameters were rescaled in such a way that the mean F_0, weighted according to amplitude, was also translated by + 2.5 Bark. This transformation

produces the characteristic frequencies in vowels of children four to five years of age from those of the same vowels pronounced by men <11>.

A third version of the vowel context was obtained by a uniform tonotopic compression of all formant frequencies and the weighted mean F_0. The compression is described by Equation <1>:

$$Z = Z_0 + 0.15\ (15.5 - Z_0) \qquad \text{<1>,}$$

where Z_0 is the critical band rate of a characteristic peak in the original version, and Z is the corresponding value in the compressed version. This transformation produces the characteristic frequencies of shouted vowels from those of the original <11>. Between these modes of speech, there are, however, additional differences which have not been imitated in our stimuli which provoked the impression of being produced by an aroused speaker rather than by a shouting one.

For conversion of the vowel formant frequencies f (in Hz) into critical band rate z (in Bark) Equation <2> that agrees to within ± 0.05 Bark with the empirical values <13> in the range of 0.2 to 6.7 kHz <10> was used and for reconversion Equation <3>. The formants, which were stationary, had the frequencies listed in Table 2 together with the weighted mean $\bar{F}_0$.

$$z = (26.81\ f\ /\ (1960 + f)) - 0.53 \qquad \text{<2>}$$

$$f = 1960\ (z + 0.53)\ /\ (26.28 - z) \qquad \text{<3>}$$

Table 2: The characteristic frequencies of the three versions of the same vowels (in Hz).

	Neutral male		Neutral child		Aroused male	
	<a>	<ɛ:>	<a>	<ɛ:>	<a>	<ɛ:>
F0	102	110	327	337	298	306
F1	751	442	1153	751	945	639
F2	1248	1799	1626	2617	1421	1932
F3	2501	2390	3702	3525	2558	2461
F4	3359	3413	5160	5258	3287	3332
F5	4311	4386	6977	7131	4052	4111

After D/A conversion the stimuli were recorded on tape in two different orders. First, F_1 and F_h started at their highest values, 24 and 25 log. units. F_1 subsequently decreased in steps of 2 u. and F_h in steps of 1 u. until the distance between the two peaks reached 7 u. In the following descending series of stimuli F_1 and F_h started 1 u. below the initial values, etc. In the second order F_1 and F_h started at their lowest values, 7 and 14 u., and ascended in reversal of the first order.

Each stimulus had a duration of .8 s and was presented twice in succession with an interval of 1.5 s. In the following, any sequence of this kind is considered as one "stimulus". Each stimulus was followed by a pause of 2.5 s for the subjects to respond. A pause of 5 s was inserted before each new series of stimuli. The stimuli were presented in six blocks, beginning with the neutral male version in the first (1) order, followed by child (2), aroused male (1), neutral male (2), child (1), and aroused male (2).

Procedure

The subjects were tested in a quiet, sound treated room and the stimuli were presented to them via Sennheiser HD414 headphones at a comfortable listening level. The subjects received answer sheets with a set of the five symbols "θ, s, tj, rs, sj" for each stimulus. After explaining the meaning of the symbols (<θ> or <f>, <s>, <ç>, <ʂ>, <ɧ>) and presenting a few stimuli for aquaintance, the subjects were asked to mark for each stimulus the symbol of the fricative they had heard. They were allowed to mark two different symbols in cases of doubt. Single-symbol responses were counted as two markings of the same symbol.

Two-dimensional histograms were obtained from the distribution of assigned labels as a function of the F_1 and F_h values. The histograms were locally normalized with respect to the total number of responses to each stimulus and smoothed by a spatial cosine filter. "Phonetic boundaries", say between <s> and <ç>, were obtained by considering only the <s> and <ç> labels and computing the 50 % level curve.

RESULTS AND DISCUSSION

Effects of presentation order

"θ"-labels were infrequent and mainly attached at the highest resonance frequencies and, occasionally, at the very lowest. The boundaries between the sibilants are shown in Figure 1. The effect of contrast can clearly be seen at the <ç> - <ʂ> boundary which is shifted by 0.9 Bark in F_1 between the two orders of presentation. Since contrast presupposes that at least one similar stimulus has been heard, there is no such effect at the beginning of each series (shown with thin lines in Figure 1). There, the responses are, instead, likely to be biased by expectation towards <s> or <ɧ> responses because the previous series of stimuli begun with these sounds. Outside this region, the <s> - <ç> boundary is shifted just as much as the <ç> - <ʂ> boundary. As for the boundary between <ɧ> and <ʂ>, the responses are likely to be biased towards <ʂ>, because this allophone would normally occur in an /aʃɛ:/ sequence as pronounced by most of our subjects. This would explain the deviant course of this boundary in the second order of presentation.

Effects of intrinsic properties

The perceptual role of the two spectral peaks in our stimuli can be understood by studying the slopes of the boundaries in Figure 1. The boundaries whose slope is not affected by order effects are well approximated by straight lines. Two of them (<ç> - <s> and <ʂ> - <ɧ>) have a course almost perpendicular to the F_1-axis, implying that the higher resonance F_h is practically irrelevant for these distinctions. Then, of course, the distance between the spectral peaks is also irrelevant. Thus, intrinsic properties of these stimuli were not used to normalize for speaker size.

Phonetic boundaries might possibly be given by a gross center of spectral gravity, like perceived "sharpness" <1>. Since F_h does affect the sharpness of our stimuli - as affirmed by informal listening - the results show that sharpness is not an invariant quantity in sibilants with a given phonetic quality.

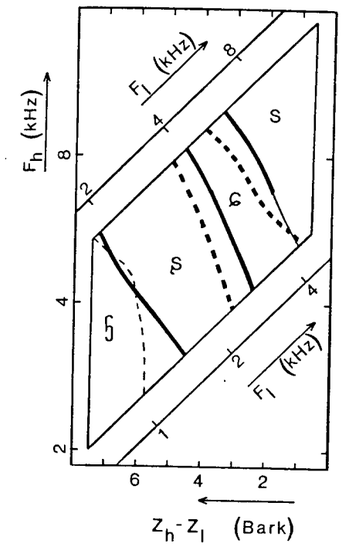

Figure 1: Phonetic boundaries between Swedish sibilants. First (continuous) and second (dashed) order of presentation. Pooled contexts.

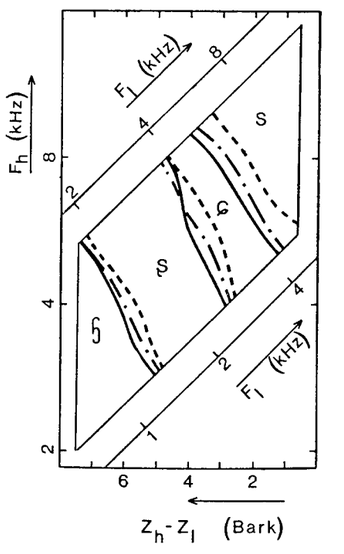

Figure 2: Phonetic boundaries between sibilants in contexts of a man's (cotinuous), a child's (dashed), and an aroused man's (dash-dotted) vowels. Pooled orders of presentation.

If the resonances are separated less than a critical distance of 3.5 Bark observed by Chistovich et al. <2> the phonetic boundaries might be expected to reflect an integrated spectral peak. The main part of our <ç> - <s> boundary runs through an area where $Z_h - Z_1 < 3.5$ Bark (see Figure 1). The slope of this line indicates, however, that this phonetic decision is only based on the pitch of the lower spectral peak or on the spectral onset of auditory excitation. Similar results have been obtained in non-phonetic pitch matching tasks <4, 9> for frequencies below 1 kHz.

The boundaries between <ç> and <ʂ> are, however, not completely independent of F_h. This may be due to the fact that <ç> and <ʂ> are the sibilants for which our synthetic stimuli were closest to the natural versions, as judged by comparison with measured spectra of Swedish sibilants <9, 8>. The other phonetic boundaries might have followed a similar course if the stimuli had been closer imitations of natural sibilants. The phonetic boundaries can be described by Equation <4>:

$$Z_1 + k_i\ Z_{hi} = I_i \qquad \text{<4>,}$$

where k_i is a factor expressing the perceptual weight of Z_{hi}, see Table 3, and I_i is a constant characteristic of boundary i. The factor k might reflect the goodness of fit between the auditory spectra of the synthetic stimuli and those of natural sibilants, but it might, alternatively, be a function of $(Z_h - Z_1)$. In that case the phonetic boundaries in Figures 1 and 2 should deviate slightly from linearity. Interestingly, k is most negative for $(Z_h - Z_1) \approx 3.5$ Bark. This reminds of the suggestion by Syrdal et al. <8> to regard this distance as specific of phoneme boundaries among sonorants. While our data do not immediately support this for sibilants - the observed boundaries are not perpendicular to the $(Z_h - Z_1)$-axis - they do show a tendency in this direction.

Table 3: Perceptual weight k of F_h in relation to that of F_1, cf. Equation <4>.

Phonetic boundary	<s>-<ç>	<ç>-<ʂ>	<ʂ>-<ç>	<ʂ>-<ɧ>
k	-0.05	-0.20	-0.27	-0.10

Effects of context

Since intrinsic normalization for speaker size is almost absent in our results, we would expect such a normalization, which theoretically would be appropriate, to be mediated by context. Figure 2 illustrates the effects of transforming the spectrum of the vowel context. We can see that the boundaries between sibilants are affected by the acoustic properties of the vowel context whose phonetic quality was close to invariant.

The extent of the boundary shift between the neutral male and the child version of the vowels (between +0.7 and +1.3 Bark) is, however, smaller than the translation of the vowel spectra (+2.5

Bark), especially at the <ʂ> - <ɧ> boundary. The boundaries in the aroused male version are shifted from those in the neutral version about halfway in the same direction as those in the child version. The <ʂ> - <ɧ> boundary (at 11.6 Bark = 1.6 kHz) is shifted by roughly +0.3 Bark, i. e., less than the vowel formants in the same frequency region (+0.6 Bark). Since, further, the upper vowel formants (above 15.5 Bark = 2.9 kHz) in the aroused male version are not shifted upwards but slightly downwards, the shift of the <s> - <ç> boundary (at Z_1 = 19 Bark) can not have been guided by the vowel formants in the same frequency region. Apparently, the sibilant boundaries are shifted about half as much as some weighted mean of the vowel formants, F2 given the highest weight. This would hold approximately for both of our context transformations, but the correlation of the extent of boundary shift with F_1 remains an open question.

ACKNOWLEDGEMENT

This research has been supported by a grant from HSFR, the Council for Research in the Humanities and Social Sciences.

REFERENCES

<1> G. v. Bismarck, Extraktion und Messung von Merkmalen der Klangfarbenwahrnehmung stationärer Schalle, München 1972.

<2> L. Chistovich and V. Lublinskaya, "The "center of gravity" effect in vowel spectra and the critical distance between formants", Hearing Res. 1, 1981, 185-195.

<3> G. Fant, "Glottal source and excitation analysis", STL-QPSR 1/1979, 85-107.

<4> R. Glave Untersuchungen zur Tonhöhenwahrnehmung stochastischer Schallsignale, Helmut Buske Verlag, Hamburg, 1973.

<5> J. M. Heinz and K. Stevens, "On the properties of voiceless fricative consonants", J. Acoust. Soc. Am. 33, 1961, 589-596.

<6> P. Lindblad, Svenskans sje- och tje-ljud i ett allmänfonetiskt perspektiv, CWK Gleerup, Lund 1980.

<7> J. Martony, "On the synthesis and perception of voiceless fricatives", STL-QPSR 1/1962, 17-22.

<8> A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition", J. Acuost. Soc. Am. 79, 1986, 1086-1110.

<9> H. Traunmüller, "Perception of timbre: ", in R. Carlson and B. Granström (eds.), The Representation of Speech in the Peripheral Auditory System, Elsevier Biomed., 1982, pp. 103-108.

<10> H. Traunmüller, "Analytical expressions for the tonotopical sensory scale", part of Ph. D. thesis, Stockholms Universitet, 1983.

<11> H. Traunmüller, "Some aspects of the sound of speech sounds", contr. to NATO-ARW on psychophysics of speech perception, Utrecht 1986.

<12> H. Traunmüller and F. Lacerda, "Perceptual relativity in identification of two-formant vowels", Speech Communication 5, 1987, ...

<13> E. Zwicker, "Zur Unterteilung des hörbaren Frequenzbereiches in Frequenzgruppen", Acustica 10, 1960, p. 185.