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

Formant frequencies of Swedish vowels

Fant, G. and Henningsson, G. and Stålhammar, U.

journal: STL-QPSR
volume: 10
number: 4
year: 1969
pages: 026-031

KTH Computer Science and Communication

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

B. FORMANT FREQUENCIES OF SWEDISH VOWELS

G. Fant, G. Henningsson*, and U. Stålhammar*

Introduction

The purpose of our study has been to collect a reference material on formant frequencies of Swedish vowels. Our previous reference, Fant (1959), pertains to a study at the Ericsson Telephone Company in 1946-1948. In that early study vowels were sustained for a time of 4 seconds necessary to perform a sweep frequency analysis. The obvious draw-back was that the subjects could have difficulties to keep a steady phonation and that the results might be more representative of singing than of speech. It is accordingly due time to produce a more suitable reference. The present study is intended as a contribution to the collection of data for this purpose. Additional data on female and childrens' speech are needed. A limitation of the present study is that it includes long vowels only.

Data sampling

A group of 24 male students at the Royal Institute of Technology who took the course in Speech Communication 1967 read a list of isolated long vowels in Swedish orthography noted as /o//å//a//ä//e//i//y//u//ö/ with about 1.5 seconds intervals between the sounds. The IPA correspondence of these alphabetic spelling forms are [u:][o:][ɑ:][æ:][e:] [i:][y:][ʉ:][ø:]. Formant frequencies were measured from broad-filter spectrograms sampled at a time location of 1/4 of the vowel length from the onset. An exception from this rule was made for the [u:] and [ʉ:] which were sampled at 1/4 from the end of the vowel. The reason for making this exception was that the first part of a vowel [u:] may come very close to a vowel [o:] whereas the [u:] typically displays a downshift of both F_1 and F_2 or at least of F_1 during the course of the vowel as the result of a lipclosing gesture. The relation of the vowel [ʉ:] to [ø:] is similar. In the first part of the vowel they contrast less than at the end of the vowel where the lipclosing gesture of [ʉ:] accounts for a falling F_1 and F_2 transition. The vowels [i:] and [y:] have a tendency to be diphthongized with a closing gesture of the tongue towards the hard palate which causes

---

* Miss Henningsson was part time employee 1967-1968 and she carried out a main part of the data sampling. U. Stålhammar conducted the listening test for pronunciation control.

a rise in F_3 at constant F_2. The contrast between [y:] and [ʉ:] would accordingly have been accentuated if we had sampled [y:] at the end just as [ʉ:]. At the same time, however, the contrast between [y:] and [i:] would have become smaller.

This reasoning illustrated the dilemma encountered when choosing a sampling convention. An alternative would be to sample at time locations where the pattern appears to be maximally close to that of an articulatory extreme which would imply the end part and not the beginning of [i:] and [y:].

Although [i:][y:][ʉ:] and [u:], i.e. all the long Swedish vowels with extreme low F_1, generally are articulated with some degree of diphthongization as described above, see also Fant (1968, 1969a), there exist considerable individual variations with respect to the extent and course of the gesture. Judging from the spectrograms the typical pattern includes a terminating phase back to a less constricted articulation with a weak or breathy voice or merely an unvoiced friction. Some subjects have a more stationary pattern throughout the vowel. In these cases the observed patterns are fairly close to those considered to be typical for the particular vowels, according to the sampling conventions adopted here.

Results and discussion

Mean values of formant frequencies and durations are given in Table I-B-1. The criteria for duration was the visible appearance of F1 in the spectrogram. The general tendency of constant vowel duration noted in Fant (1969b) is apparent. Thus there is no significant trend of open vowels being longer than close vowels. On the contrary, there could be an opposite tendency. Thus the [y:] was on the average 3.5 % longer than other vowels. The small differences observed could have been influenced by the sequential order in the reading which was not randomized. The general tendency, however, appears to be that isolated vowels are timed by a constant neural phonatory control pattern independent of articulation and possible also independent of the relative degree of diphthongization of close vowels.*

* In a separate study U. Stålhammar has found a definite tendency of [ɑ:] being shorter than [i:] in context-free articulations. These measurements pertain to boundaries not limited to the baseline voicing, i.e. at any formant carrying visible energy. These data will be published in a forthcoming STL-QPSR.

TABLE I-B-1

Formant frequencies and durations

Vowel		Present study spoken vowels					Fant (1959) 4 sec duration			
IPA	STA	F_1 Hz	F_2 Hz	F_3 Hz	F_4 Hz	D ms	F_1 Hz	F_2 Hz	F_3 Hz	F_4 Hz
[u:]	o_1	290	595	2330	3260	390	310	730	2230	3300
[o:]	$å_1$	390	690	2415	3160	410	400	710	2460	3150
[ɑ:]	a_1	600	925	2540	3320	410	580	940	2480	3290
[æ:]	$ä_3$	625	1720	2500	3440	410	610	1550	2450	3400
[ɛ:]	$ä_1$	505	1935	2540	3370		440	1795	2385	3415
[e:]	e_1	345	2250	2850	3540	400	335	2050	2510	3400
[i:]	i_1	255	2190	3150	3730	410	255	2065	2960	3400
[y:]	y_1	260	2060	2675	3310	425	260	1930	2420	3300
[ʉ:]	u_1	285	1640	2250	3250	410	285	1635	2140	3310
[ø:]	$ö_1$	380	1730	2290	3325	410	365	1690	2200	3390

Before the final averages of formant frequencies were articulated we discarded vowel samples which according to a control listening appeared to depart from a standard Swedish pronunciation. A comparison with old data from Fant (1959) included in Table I-B-1 shows that the formant data of [u:][o:][ɑ:] as well as [ø:] and [ʉ:] do not differ much. In the new data F_2 and F_3 of [i:][e:][æ:] and [y:] are 100-350 Hz higher than in the old data. Since overall vocal tract length provides the scale factor we might be faced with a difference in the mean vocal tract size of the two groups. This explanation, however, is not very likely. A calculation of the individual scale factor for each speaker according to a formula* for an average F_3:

$$F_{3av} = \frac{F_{3ø} + F_{3e} + F_{3æ} + F_{3i} + F_{2i}}{5} \qquad (1)$$

* In this connection it was noted that a simple average of F_2 and F_3 of the vowel [e:] can serve as an approximate measure of F_{3av} which is 2 % below the value calculated from Eq. (1) or close to 2550 Hz.

gave a mean for all 24 speakers of F_{3a} = 2590 Hz compared with F_{3av} = 2440 Hz for the Fant (1959) data. The earlier study included seven male subjects. In the new study the standard deviation of F_{3av} among speakers was 5.2 % or 135 Hz and 30 Hz only with respect to the mean of the group which accordingly should be rather representative.

A more probable explanation can be found by considering the sustained articulation in the early work as more typical of singing than of speaking. It is hard to sustain a vowel for 4 seconds without adapting a singing mode. Also most of the subjects were amateur singers. In the study of Sundberg (1968) the subjects tested in a singing mode showed 100-250 Hz lower F_2 and F_3 of [i:][e:][æ:] and [y:] than in speaking which is of the same order of magnitude as the differences between the old and the new vowel data. These observations are a strong evidence for the assumption that the observed differences pertain to a speaking versus a singing mode rather than to differences in vocal tract average size. A lowering of the larynx is one means of increasing the overall length of the vocal tract which is a characteristic of the singing mode, Sundberg (1968, 1969).

We shall next discuss the formant data as a basis of phonetic contrasts. Given the vowels as points in a multidimensional space we are concerned with the absolute location of these points and the multivalued vectors from each point to its neighbor's. With the physical dimensions limited to F_1 and F_2 a plotting of all the subjects displays an overlap of vowel regions which is much reduced if F_3 is included.

In our data we had a few cases of ambiguities such as that of one subject's [y:] having approximately the same F_1 F_2 and F_3 and even F_4 as a second speaker's [e:]. In these cases an inspection of the spectrograms most often revealed an erroneous formant measurement or a sampling at a less representative point within the vowel. Remaining ambiguities could always be removed by reference to the speaker's average F_3, i.e. to his scale factor. Any subject thus preserves an invariance within his own vowel space. This implies that within his ensemble of vowels there is no confusion but that two vowels phonemically different and uttered by different subjects still could have been heard as phonetically the same when listened to out of contexts. This was not the case in our study but remains an interesting possibility to be investigated.

It is clear, however, that there is more information within an isolated vowel than merely the F-pattern, e.g. F_0 and formant bandwidths. An important cue often present in the maximally close vowels [i:][y:][ʉ:] and [u:] is as already noted a diphthongization. The Swedish [e:] is rather close to [i:] and [y:] in terms of F_1 but there should always be a lower F_1 in the maximally close phase of [i:] and [y:] than in [e:]. It also seems possible that a correction for F_0 could remove an F_1 ambiguity in the sense of a positive correlation between F_0 and F_1 in a distribution with respect to speakers of different size factors, Fant (1959). In the present data several exceptions were found from this average trend, i.e. speakers with high F_0 and low F_{3av} or low F_0 and high F_{3av}.

To what extent are distances between formants more obvious cues than their absolute frequencies? According to Table I-B-1 one apparent pattern aspect of the front vowels [e:][ɛ:][y:][ʉ:] and [ø:] is that F_3-F_2 is close to 600 Hz but is almost 1000 Hz for [i:]. Also [i:] has a higher F_4 than any other vowel. This F_4 feature, however, was not observed in the old Fant (1959) data. The low F_3 of [ʉ:] and [ø:] close to F_2 is associated with a high F_4-F_3 of the order of 1000 Hz for these vowels whereas F_4-F_3 of [y:] averages 650 Hz.

Thus, typically [e:] and [y:] are characterized by equidistant approximately 650 Hz spacings of F_2 F_3 and F_4 with higher bandwidths for [e:] than for [y:] and 200 Hz higher mean value of the F_2 F_3 F_4 group in [e:] than in [y:]. The vowel [i:] has the largest F_4-F_2 span and the largest F_3-F_2 of all vowels and F_3 is typically closer to F_4 than to F_2. In [ɛ:] and [æ:] F_3 is somewhat closer to F_2 than to F_4 and in [ʉ:] and [ø:] F_3 is much closer to F_2 than to F_4.

The perceptual importance of these pattern aspects is not known. It appears that in female and children's voices the relative role of individual formants may differ whilst an overall pattern aspect, yet to be defined, could be similar. Within the group of male voices there also exist variations especially in the dynamic pattern of [y:]. Some speakers produced the [y:] without diphthongization and among those some had an F_3 closer to F_2 than to F_4. The transitional y-patterns all showed a rising F_3 whilst F_4 could be level or slightly falling contrary to the [i:] transitions where F_4 would be level or rising.

It is known that the phonetic interpretation of a vowel within certain limits is rather insensitive to variations in the amplitude of a single formant and to shifts in the spectrum level of F_2 and higher formants considered as a group. However, a shift up in frequency of F_4 may have the effect of concentrating the perceptual importance of the upper part of the spectrum to F_3 and F_2 providing F4 does not have an excessively high amplitude. Similarly F2 of [i:] appears to have rather small perceptual importance.

In this connection one could reinterpret a finding of Fujimura (1967) as follows. In his 3-formant synthesis he found that an increase in the separation of F_3 and F_2 retaining their geometric mean would cause a shift of the response from [y:] to [ʉ:]. However, a small distance between F_3 and F_2 is more typical of [ʉ:] than of [y:] according to our data. It seems probable that F_3 of the Fujimura stimuli carried the role of F_4 in natural speech whilst his F_2 might substitute F2 and F3 of human speech.

There remains much to be learned concerning how various spectrum pattern aspects influence phonetic identification. One possible model of perception of F2 and higher formants of front vowels would be to lay the primary importance on some kind of mean frequency, Fant (1969a). In the second approximation we would have to consider the frequency width of this higher group and perhaps also relative formant distances as that of F_3 with respect to F_2 and F_4. Two-formant synthesis now carried out by Carlsson and Granström may throw some light on these problems.

References:

FANT, G. (1959): "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", Ericsson Technics No. 15, pp. 3-108.

FANT, G. (1968): "Analysis and Synthesis of Speech Processes", in Manual of Phonetics, ed. by B. Malmberg (Amsterdam), pp. 173-277.

FANT, G. (1969a): "Distinctive Features and Phonetic Dimensions", STL-QPSR 2-3/1969, pp. 1-18.

FANT, G. (1969b): "Stops in CV-Syllables", STL-QPSR 4/1969, pp. 1-25 (this issue).

FUJIMURA, O. (1967): "On the Second Spectral Peak of Front Vowels: A Perceptual Study of the Role of the Second and Third Formants", Language and Speech 10, pp. 181-193.

SUNDBERG, J. (1968): "Formant Frequencies of Bass Singers", STL-QPSR 1/1968, pp. 1-6.

SUNDBERG, J. (1969): "Articulatory Differences Between Spoken and Sung Vowels in Singers", STL-QPSR 1/1969, pp. 33-46.