THE CONTEXT SENSITIVITY OF THE PERCEPTUAL INTERACTION BETWEEN F0 AND F1

Hartmut Traunmüller

Institutionen för lingvistik, Stockholms universitet,
S - 106 91 Stockholm, Sweden.

ABSTRACT

According to a known hypothesis, the perceived degree of openness in vowels is given by the CB-rate difference (tonotopic distance) between F1 and F0. Synthetic vowels and diphthongs with non-stationary F_0 and/or F_1 were used to find out whether it is the instantaneous F_0, its average, or the prosodic baseline, that is relevant here. Most subjects behaved in accordance with the basic hypothesis, but some attached a smaller weight to F_0. The results support the relevance of the prosodic baseline as well as that of the instantaneous value of F_0. Between speaker differences in behaviour were prominent.

1. INTRODUCTION

It is well known that the phonetic quality of phonated vowels, in particular their perceived degree of openness, or vowel "height", depends not only on the frequencies of their formants but also on their F_0. According to one hypothesis, the perceived openness is given by the tonotopic distance (CB-rate difference) between F1 and F0 [6]. Data on F_0 and the formants of vowels produced at different degrees of vocal effort and by speakers with differently sized vocal tracts are largely compatible with such an hypothesis [5, 7]. It is, however, still in question whether it is the instantaneous F_0, its average, or some other kind of context dependent reference value that is relevant here.

The tonotopic distance hypothesis was first proposed to explain the results of perceptual experiments with synthetic vowels [6]. Its quantitative validity has been questioned on the basis of results obtained in another perceptual experiment, in which the influence of F_0 turned out to be smaller [4]. The discrepancy can be explained if it is assumed that listeners relate F_1 to the prosodic baseline rather than to an instantaneous or average value of F_0 [8]. Such a baseline is obtained by interpolation between successive minima in the F_0-contour of the breath-group in question.

Data on F_0 in different styles of speech show that an invariant minimal value of F_0 is characteristic of each speaker [9]. That value of F_0 is normally reached close to the end of statements. It appears to be stable in various types of paralinguistic variations, such as the degree of involvement [1] and in different styles of speech [2, 3], at least as long as these do not involve an overall change in vocal effort. More precisely, the invariant value of F_0 is slightly above its minimum, and it might represent an average of the baseline.

If this is to be reflected in speech perception, listeners should, *in effect*, relate F_1 to an estimate of the speaker's prosodic baseline in judging vowel openness. According to slightly different hypotheses, the minimum F_0 in the whole breath-group or in a smaller unit of speech might be relevant instead. In order to test the various hypotheses, an experiment was performed with synthetic vowels and diphthongs in which either F_1 or F_0 varied or both varied in unison.

2. METHOD

2.1 Stimuli

The stimuli were synthesized digitally by means of a terminal analog of the vocal tract, using a three-parameter voice source and 8 formant filters in cascade. The excitation signal used imitated that observed, by inverse filtering, in a vowel produced by a woman. Thus, F_0 followed a natural intonation contour. The nominal F_0-values referred to in the following are amplitude weighted mean values. These were 161, 250, 347, 453, 569, and 697 Hz, representing steps of 1 Bark. The stationary positions of F_1 were 250, 347, 453, 569, 697, and 838 Hz. The formants above F1 were in all stimuli invariably at the following positions in Hz: 2 220, 3 406, 4 434, 5 050, 5 741, 6 785, 7 829.

The stimuli had a duration of 470 ms. Prospective diphthongs were obtained by frequency modulation of F1 and/or F0 with part of a sinusoid with a period of 360 ms, phased such that the nominal target values of F_1 and F_0 were reached 30 ms after the beginning and 80 ms before the end of the stimuli. The asymmetry was motivated by a final decrease in excitation amplitude.

The nominal F_0-targets for the diphthongs were 250 and 453 Hz (stimulus series 3a and 3b), 161 and 569 Hz (4a and 4b), and 250 and 347 Hz (5a and 5b). The targets of F_1 were in each sereis 1 Bark above those of F_0.

2.2 Subjects

The stimuli were listened to and transcribed phonetically by 20 subjects, recruited among the personnel and students of the institute. Their first languages were Swedish (12), German (2), Finnish, Estonian, Russian, Bulgarian, English, and Portuguese (1 each). The subjects reported no hearing disorders and they claimed good vocal proficiency in 4.7 languages, on average.

2.3 Procedure

The stimuli were presented binaurally through headphones in 8 series with 6 (first two series only) or with 9 stimuli each, as follows: (1) nominal F_0 = 161 Hz, F_1 rising in steps of 1 Bark. (2) Both F_0 and F_1 rising in steps of 1 Bark. The remaining series (3a to 5b) contained stimuli in which both F_0 and F_1 varied between the chosen target values. Each of these six series included also one sample of each combination of stationary target values: F_0 low, F_1 low; F_0 low, F_1 high; and F_0 high, F_1 high. Series a and b differed only in the order of presentation.

3. RESULTS

The stimuli were predominantly heard as front unrounded vowels with or without diphthongization. In some cases subjects heard front rounded vowels. The responses were computed according to the associated degree of openness as follows: [i y]: 1, [e ø]: 2, [ə]: 2.5, [ɛ œ]: 3, [æ]: 4. For diacritical marks "more (less) open" 0.5 was added (subtracted). In order to accommodate various diphthongs, the responses were quantified using four subsequent values according to the following model: [e]: 2222, [ej], [eˈ]: 2221, [ei]: 2211, [ᵉi]: 2111.

Fig. 1 shows the average perceived degree of openness in the vowel series with subsequently rising F_1 with and without rising F_0 (series 1 and 2). The last one of the four values assigned to each response was ignored. The vowels with the same F_0 and the same higher formants, but with subsequently rising F_1 were unanimously perceived as subsequently more open, from [i] to [æ] (upper line). The spread in perceived

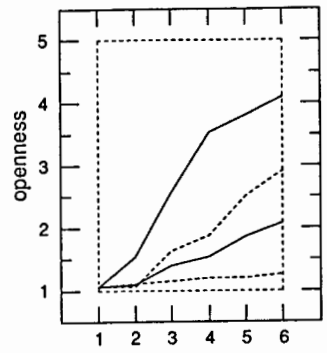

Fig. 1

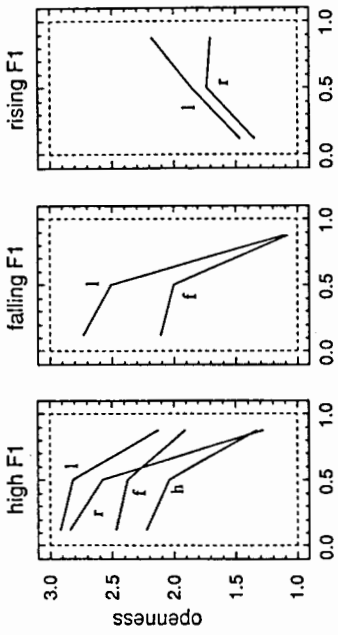

Fig. 3 (3a 3b 3c)

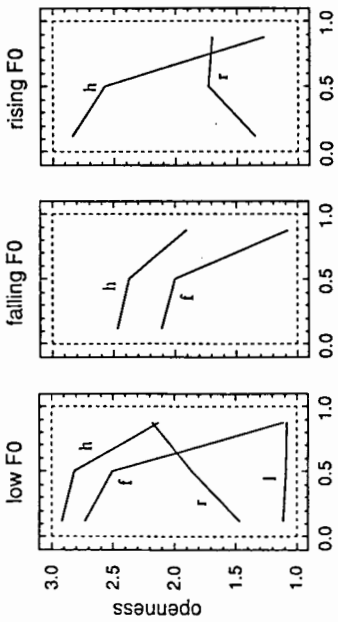

Fig. 2 (2a 2b 2c)

openness was small. As for the stimuli in which both F_1 and F_0 increased (lower line), ten subjects perceived essentially no change in openness, hearing all as [i] (lower dashed line), while the other ten were less uniform in behaviour. For them, only about 40 to 70 % of a shift in F_1 was compensated by an F_0-shift equal in Bark (upper dashed line). The subjects behaved in a unanimous fashion only up to F_0 = 250 Hz.

Fig. 2 shows the effect of variations in F_1 on the perceived degree of openness as a function of time from the beginning to the end of each stimulus in the series 3 to 5. The two non-terminal openness values have been averaged for these figures. The figure shows the results pooled over all subjects and over all three choices of extreme values for F_1 and F_0. There was no noticeable difference between the two orders of presentation. Fig. 2a includes the four cases in which F_0 was low, while F_1 was low (l), rising (r), high (h), and falling (f). In Fig. 2b, F_0 is falling, while F_1 is either high or falling. In Fig. 2c, F_0 is rising, while F_1 is either high or rising.

Fig. 3 is analogous to Fig. 2, but it shows the effect of variations in F_0 when F_1 is given. Fig. 3a includes the four cases in which F_1 was high, while F_0 was low (l), rising (r), high (h), or falling (f). In Figs. 3b and 3c, F_1 is rising and falling, respectively, while F_0 is either low or rising and falling with F_1.

The stimuli in which F_1 and F_0 were "stationary" were often heard as finally diphthongized. This tendency is exaggerated in the results, since even a slight degree of closing diphtongization in open vowels was often transcribed as [V^i] or [Vj].

4. DISCUSSION

The results of the first experiment show that the typical listener behaves quite precisely in accordance with the tonotopic distance hypothesis. The results of the large group of listeners who appear to attach a smaller weight to F0 are troublesome. Considering the quite high degree of naturalness of the stimuli, these results tell us that there will be large between speaker discre-

pancies in perceived phonetic quality even in natural speech produced at high vocal effort, in particular by children, and in soprano singing. As for the age-conditioned variation *per se*, which is also reflected in an approximately uniform shift in F_0 and F_1, there is a cue to vocal tract size in the formants above F2, which is likely to reduce between listener variation for that case.

Fig. 3 demonstrates clearly that the instantaneous F_0 (or a short time average) is of some importance. If the subjects were only sensitive to F_0 averaged over the whole stimulus, the contours in each panel would run in parallel, with a vertical displacement. If they were only sensitive to the F_0-minimum within each stimulus, the contours in each panel, except h in 3a, would coincide. If they were only sensitive to the baseline, all contours would coincide within each panel. On average, the data show a combination of baseline and instantaneous effects, the relative weight of the latter increasing from 0.36 to 0.68 during the course of the stimuli, but this does not hold for each subject.

The responses of individual subjects to the stimuli of Figs. 2 and 3 are not generally predictable from their responses to those of Fig. 1. This is shown in Fig. 4, in which the F_0-sensitivity (in % compensation) in the two types of context is shown for each subject. The comparison includes only the stimuli with "stationary" F_0. The correlation between the two sets of data is low (0.44). There were some subjects who relied entirely on the instantaneous F_0, while others relied on the baseline. The former appear along the diagonal ($y = x$), the latter where $y = 0$. Thus, between speaker differences turned out to be very prominent.

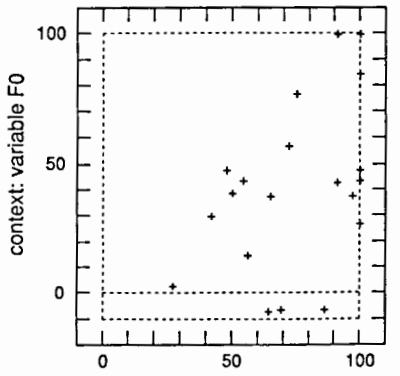

Fig. 4

In a previous experiment, in which the perception of F2' was in focus, it was also observed that some subjects behaved consistently in agreement with the tonotopic distance hypothesis, while others showed a reduced influence of F_0 and often a less consistent behavior [10]. The proportion of the latter was lower among speakers of Swedish than among speakers of Turkish. Apparently, it had been still lower in speakers of Austrian German [6]. This might then be correlated with functional load: The *minimum* number of openness distinctions which are necessary to describe the phonological distinctions in the vowel systems is two for Turkish, three for Swedish, and four for Austrian German. As for the balance between instantaneous F_0 and its baseline, the functional load of tone might be of importance, but there are no data to substantiate such a hypothesis.

5. REFERENCES

[1] Bruce, G. *Working Papers 23* (1982) 51–116, Dep. linguist., Lund univ.

[2] Graddol, D. and

[3] Johns-Lewis, C. in *Intonation in Discourse*, C. Johns-Lewis (ed.), Croom Helm, London & Sidney, 1986, pp. 221–237 and 199–219.

[4] Nearey, T. M. *JASA 85* (1989) 2088–2113.

[5] Syrdal, A. K.; Gopal, H. S. *JASA 79* (1986) 1086–1100.

[6] Traunmüller, H. *JASA 69* (1981) 1465–1475.

[7] " *Phonetica 45* (1988) 1–29.

[8] " *JASA 88* (1990) 2015–2019.

[9] Traunmüller, H.; Branderud, P.; Bigestans, A. *PERILUS X* (1989) 47–64. Inst. linguist. Stockholm univ.

[10] Traunmüller, H.; Lacerda, F. *Speech Comm. 5* (1987) 143–157.