INTERSPEECH 2009 BRIGHTON

On-line Formant Shifting as a Function of F0

Kateřina Chládková [1], *Paul Boersma* [1], *Václav Jonáš Podlipský* [2]

[1] Amsterdam Center for Language and Communication, University of Amsterdam, The Netherlands
[2] Department of English and American Studies, Palacký University Olomouc, Czech Republic
katerina.chladkova@seznam.cz, paul.boersma@uva.nl, podlip@ffnw.upol.cz

Abstract

We investigate whether there is a within-speaker effect of a higher F0 on the values of the first and the second formant. When asked to speak at a high F0, speakers turn out to raise their formants as well. In the F1 dimension this effect is greater for women than for men. We conclude that while a general formant raising effect might be due to the physiology of a high F0 (i.e. raised larynx and shorter vocal tract), a plausible explanation for the gender-dependent size of the effect can only be found in the *undersampling hypothesis*.
Index Terms: speech production, formants, F0

1. Introduction

The formants of vowels produced by female speakers have been shown to differ consistently from vowels produced by male speakers of the same language. Not only do we find that women have higher formants than men [1], which can be explained by the longer vocal tracts of men (e.g. [2]), but also the ratio of the formants of high and low vowels, and the ratio of the formants of front and back vowels, i.e. the size of the F1 and the F2 vowel space, seems to be larger for women than for men [3].

Previous research attempted to explain this difference in female and male vowel space sizes. Some studies proposed that it arises from differences between the female and male vocal tract anatomy, e.g. [2] [4]. Others have suggested that socio-phonetic factors are involved because women aim at speaking more clearly, and both women and men want to sound differently from the other sex [5] [6].

Yet another direction of research has related the vowel space size difference between men and women to the fact that women generally have higher fundamental frequencies. Diehl *et al.* [7], for instance, tested the perception of vowels synthesized with typical female and male formant values at various levels of fundamental frequency (F0). Since listeners' performance declined as F0 increased, the authors concluded that the poor vowel identifiability was due to the sparser distribution of harmonics in the spectra of the vowels synthesized at higher F0s. Diehl *et al.* thus ascribe their results to the *undersampling hypothesis* (formulated earlier by Goldstein [6] and Ryalls and Lieberman [8]), which claims that the greater between-category dispersion that is found in female vowels can function "as a means of offsetting the deleterious effect on vowel identifiability of (typically) higher F0s" [7].

The present study was designed to test whether speaking at a higher-than-normal or lower-than-normal pitch results in vowel formants shifting as a function of the varying F0. The specific questions we aim to answer are (1) whether the height of a speaker's pitch has an effect on the formant frequencies of her vowels in general, and more importantly, (2) whether speakers produce more dispersed vowel inventories when speaking at a higher-than-normal pitch than when speaking at a normal or lower-than-normal pitch.

2. Method

2.1. Data collection

2.1.1. Participants

Nine female and nine male Czech native speakers volunteered as participants in the present experiment. They were students at Palacký University. The female participants were aged 19–24 (mean 21.6, standard deviation 1.7), the male participants were aged 19–27 (mean 23.3, standard deviation 2.9). None of the 18 subjects reported to have had any speech, hearing, or other language-related impairments. Prior to testing, the participants were not familiar with the aim of the experiment.

2.1.2. Recordings

Six of the speakers (3 women and 3 men) were recorded in a quiet room using a Marantz PDM671 recorder (at a sampling rate of 44.1 kHz, and 16 bits quantization) and a headmounted microphone Samson QV (hypercardioid; with a Samson PM6 phantom power adapter). The remaining twelve speakers were recorded in a sound-treated booth using a Røde Broadcaster microphone (cardioid), a Mackie 1642-VLZ3 mixer, and an M-audio Delta 66 computer sound card (sampling rate 44.1 kHz, 32 bits quantization).

The experimental task was a phrase-list reading. The phrases were of the template: *Ve slově CVC máme V* (meaning 'In the word CVC we have a V'), where the V was always one of the ten Czech monophthongs /iː, ɪ, ɛː, ɛ, aː, a, oː, o, uː, u / (orthographically *í/ý, i/y, é, e, á, a, ó, o, ú/ů, u*). Each vowel was embedded into seven different consonantal contexts (preferably voiceless), yielding five existing and two non-existing words per vowel. An example sentence with the vowel /a/ is: *Ve slově sak máme a* (meaning 'In the word *sack* we have an *a*').

Each subject recorded the list of the 70 phrases in three intended-pitch conditions: Normal, High, and Low. Subjects were asked to read aloud the phrases as naturally as possible. In addition, for the High and the Low intended-pitch condition, they were asked to read the phrases at a *slightly* raised or slightly lowered pitch, respectively; before the recording, the participants practised the pitch modification with the experimenter to ensure a natural production (see the findings of [9] [10], which suggest that if people change their F0 as a result of impersonating a cartoon figure for instance, they may thereby also adopt that cartoon figure's formants). If during the recording the experimenter judged the production to be unnatural, unclear, or misread, or the pitch modification to be either exaggerated or absent, the participant was asked to repeat the whole phrase. Therefore, more than 3780 CVC tokens were recorded (i.e. at least 210 tokens per speaker). Participants were allowed to take a short break at any time during the recording.

10.21437/Interspeech.2009-164

2.2. Data analysis

The data analysis aimed at retrieving reliable measurements of the F0, F1 and F2 of the recorded vowel tokens. In each recorded CVC word, the start point and the end point of the vowel were identified manually in the digitized waveform. The highest peaks of the first and last periods that resembled the central periods of the vowel and still had considerable amplitude were taken as the start and end points, respectively. About a dozen tokens were excluded from further analyses because they were either creaky-voiced, noisy, or they did not sound natural. In the end, we thus had 3818 CVC tokens to be measured.

2.2.1. Fundamental frequency

Fundamental frequency was measured in Praat [11] by its accurate auto-correlation method [12], in time steps of 1 millisecond, with the searchable pitch range set to 65–500 Hz for females and 45–480 Hz for males. The median measured F0 value of the middle 40% part of each vowel token was taken as that vowel token's F0. Vowel tokens for which Praat failed to report an F0 value were measured manually in the waveform (this happened for 16 of the 3818 tokens).

As the participants could be expected to make relative rather than absolute F0 changes, we stored in our data tables not the original F0 values (which had been measured in Hz), but logarithmic transformations of them, so that we could apply linear statistical models.

2.2.2. Formants

The first and the second formant were determined by the Burg algorithm built into Praat [11]. In time steps of 1 millisecond, a 50-ms long Gaussian window was applied to the sound, and Praat was made to search for formants within the range 0–5500 Hz for females and 0–5000 Hz for males. The number of formants searched for within that frequency range was five for the non-back vowels /iː, ɪ, ɛː, ɛ, aː, a / and six for the back vowels /oː, oː, u, uː/ (see [13] for the influence of vowel backness on formant measurements). The median measured F1 and F2 values over the middle 40% portion of each vowel token were taken as that vowel token's F1 and F2.

As vowel articulations can be expected to result in relative rather than absolute F1 and F2 changes (roughly independent of the speaker's overall vocal tract size), we log-transformed the original F1 and F2 values (which had been measured in Hz) so that we could apply linear statistical models to these data.

2.2.3. Statistical analysis

As indicated above, statistical analyses were carried out with linear models. However, such models require that the data be normally distributed, and this will often not be the case, because the speakers and/or the analysis software may make mistakes in producing or measuring valid F0 or formant values. To mitigate the influence of outliers, then, we took the median value over the seven consonantal contexts as representative of each vowel of each speaker. Thus, for each of the 18 speakers we ended up with 30 values of F0, F1 and F2 (3 intended pitch conditions × 10 vowel categories).

The 540 values of e.g. F1 could then be submitted to a linear model. The design of the experiment dictates a repeated-measures analysis of variance with gender as the between-subjects factor and vowel category and intended pitch as the within-subject factors. Since the data typically fail to pass Mauchly's test of sphericity, our *F*-tests were standardly performed (in SPSS [14]) with Huynh-Feldt's correction, which multiplies the numbers of degrees of freedom by a factor between 0 and 1.

3. Results

Although all statistical analyses (analyses of variance and computations of means and confidence intervals) were performed on log-transformed values, readability considerations demand that averages and confidence intervals are reported as values in Hertz, as are the axes of the figures. For reporting, we therefore transform all results back to the Hertz domain, and differences in the log domain are reported as ratios in the Hertz domain.

3.1. F0 as a function of intended pitch

The repeated-measures analysis of variance on the 540 F0 values reveals a main effect of intended pitch ($F[2{\cdot}0.757, 32{\cdot}0.757] = 42.496$; $p = 7.2{\cdot}10^{-8}$). Figure 1, where each point represents the mean F0 over 9 speakers and 10 vowels, indicates that Czech speakers, as expected, raise their F0 when asked to speak at a High pitch: the ratio by which they multiply their F0 between the Normal and High conditions is 1.29 (the 97.5% confidence interval [i.e. Bonferroni-corrected for two comparisons] is 1.17..1.42). Speakers also seem to lower their F0 when asked to speak at a Low pitch: the 97.5% confidence interval (c.i.) of the ratio by which Czech speakers divide their F0 between the Normal and Low conditions is 1.005..1.12. As the two confidence intervals do not overlap, we conclude that Czech speakers respond more successfully to the High- than to the Low-pitch task.

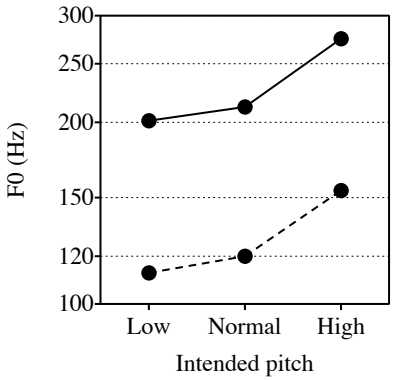

Figure 1: *F0 as a function of intended pitch. Solid lines: women; dashed lines: men.*

We can summarize the task effect in one number: the ratio of the F0 values between the High and Low conditions is 1.37 (95% c.i. = 1.24..1.50). This result is an important preliminary to the analyses of F1 and F2 below: the participants are apparently able to follow the task they are given.

The tests show no interaction of gender with intended pitch or with vowel category, and no triple interaction either [all three $F < 1$]. There is an interaction of vowel category and intended pitch ($F[18{\cdot}0.569, 288{\cdot}0.569] = 5.209$; $p = 1.0{\cdot}10^{-6}$). The cause of this interaction seems to be that speakers avoid F0 differences between long and short vowels when changing their intended pitch: in the Normal condition, short vowels have a higher F0 than long vowels, by a ratio of 1.067 (95% c.i. = 1.048..1.085); in the High condition, the short-long F0 ratio drops to 1.032 (95% c.i. = 1.018..1.045), which is reliably smaller than in the Normal condition ($t[17] = 4.286$; $p = 2.5{\cdot}10^{-4}$); and in the Low condition the ratio drops to 1.034 (95% c.i. = 1.022..1.045), i.e. also reliably smaller than in the Normal condition ($t[17] = 4.049$; $p = 4.2{\cdot}10^{-4}$).

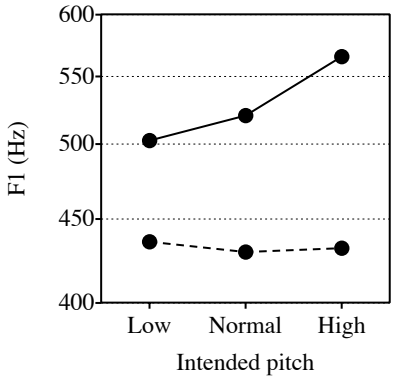

Figure 2: *F1 as a function of intended pitch. Solid lines: women; dashed lines: men.*

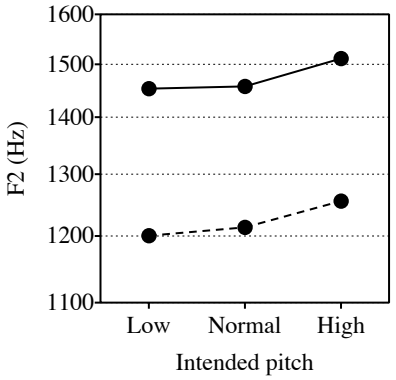

Figure 3: *F2 as a function of intended pitch. Solid lines: women; dashed lines: men.*

3.2. F1 as a function of intended pitch

The repeated-measures analysis of variance on the 540 F1 values reveals a main effect of intended pitch (F[2·0.974, 32·0.974] = 9.656; $p = 6.0 \cdot 10^{-4}$). This indicates that Czech speakers vary their F1 with the intended pitch. Importantly, we find a significant interaction of intended pitch and gender (F[2·0.974, 32·0.974] = 11.709; $p = 1.8 \cdot 10^{-4}$): as illustrated in Figure 2, Czech women raise their F1 between the Low- and High-pitch conditions by a large factor of 1.125 (t[8] = 5.315; 95% c.i. = 1.07..1.18), whereas Czech men raise their F1 slightly or not at all (t[8] = –0.491; 95% c.i. = 0.95..1.03).

Figures 4 and 5 illustrate these results forcefully: for all five short vowels and all five long vowels, the average F1 of the 9 Czech female participants is greater in the High- than in the Low-pitch condition.

3.3. F1 range as a function of intended pitch

While some compensation for undersampling is already achieved by raising the F1 value of *every* vowel, even more compensation can be achieved by raising the F1 values of open vowels more than the F1 values of closed vowels. Figures 4 and 5 suggest that such a stretching of the F1 range indeed takes place, both for the short and for the long vowels: the vertical shift of /a/ looks larger than the vertical shifts of /ɪ/ and /u/, and the vertical shift of /aː/ looks larger than the vertical shifts of /iː/ and /uː/.

To test this accurately, we computed for each of the 9 women her *F1 range*, which we define as the geometric average of the F1 values of her /a/ and /aː/ divided by the geometric average of the F1 values of her /ɪ/, /u/, /iː/ and /uː/. We thus obtain 27 F1 range values: 9 speakers × 3 intended pitch conditions. A paired-samples t-test shows that for the population of Czech women the F1 range may indeed be greater in the High- than in the Low-pitch task, namely by a factor of 1.10, although this result is not very reliable (t[8] = 1.729; 90% c.i. = 0.99..1.23; one-tailed p from 1 is 0.061).

3.4. F2 and F2 range as functions of intended pitch

The repeated-measures analysis of variance on the 540 F2 values reveals a main effect of intended pitch (F[2·0.696, 32·0.696] = 14.131; $p = 3.9 \cdot 10^{-4}$). This time, the analysis reveals no interaction between intended pitch and gender. Both findings are illustrated by Figure 3. The *F2 range*, defined as the geometric average F2 of /ɪ/ and /iː/ divided by the geometric average F2 of /u/ and /uː/, is, for Czech women, greater in the High- than in the Low-pitch task, namely by a factor of 1.06 (t[8] = 3.093; 90% c.i. = 1.03..1.11).

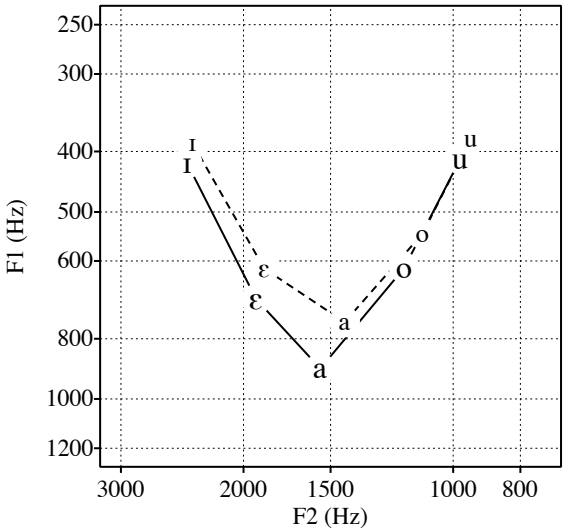

Figure 4: *Average Czech female short vowels (large font, solid lines: High intended pitch; small font, dashed lines: Low intended pitch).*

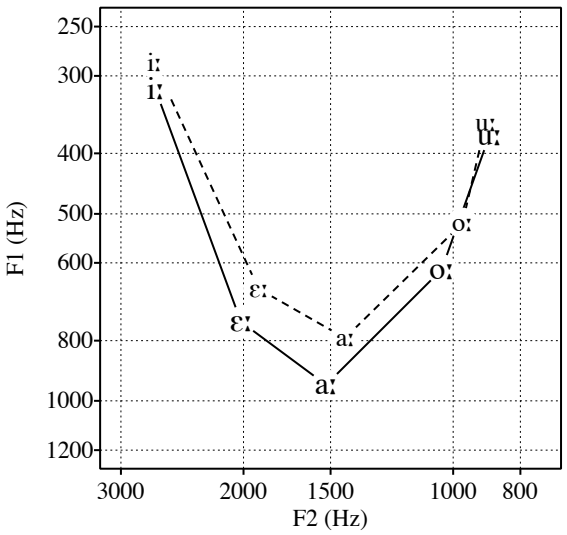

Figure 5: *Average Czech female long vowels (large font, solid lines: High intended pitch; small font, dashed lines: Low intended pitch).*

4. Discussion

The results of our study show that when speakers are asked to raise or lower their pitch, they tend to do so: the average ratio of produced F0 values between the High- and Low-pitch tasks is 1.37. This is unsurprising. The more interesting observation is that as *female* speakers raise their F0, they raise their F1 values as well (by an average ratio of 1.125), something that male speakers do hardly if at all. We will now discuss which of the hypotheses mentioned in the Introduction is supported by these findings.

First, the rise of F1 with F0 could have a physiological cause: the articulatory implementation of F0 raising tends to involve raising the larynx (e.g. [15]), something that shortens the vocal tract and is therefore likely to raise the formants. Although we can more or less see the effect of this in Figures 2 and 3, this physiological explanation would predict similar effects for both sexes, which is not what we observed in §3.2, so the gender dependence that we see in Figure 2 cannot be ascribed to the physiology.

Second, the hypothesis that the reason why women have larger vowel spaces than men is because they care more about speaking clearly than men do [6] cannot in itself explain the fact that with rising F0 women raise their F1 values: after all, why would women want to speak even *more* clearly if the F0 happens to be high?

The only remaining explanation for the rise of F1 with F0 is the *undersampling hypothesis* [6] [7] [8]: the higher the F0 is, the fewer harmonics of F0 fit inside the vowel space; such "undersampling" causes a loss of clarity, and a speaker can compensate for this by increasing the size of his or her vowel space. Importantly, we observed that women but not men raise their F1 when they speak at a higher pitch. A plausible explanation is that spectral undersampling happens especially whenever F0 is very high. A male raising his F0 from 120 to 180 Hz, for instance, will then feel less need to increase his vowel space than a female who raises her F0 from 200 to 300 Hz. After all, a spectral spacing of 300 Hz is much worse perceptually (i.e. will deteriorate vowel identifiability more) than a spectral spacing of 180 Hz (see [16] for a comparable effect of F0 on formant values and vowel dispersion in tenor versus bass singers). The undersampling hypothesis then predicts that women who raise their F0 want to raise their formants to a larger extent than men who raise their F0, and this is exactly what we found in our experiment. The findings of the present study therefore support the undersampling hypothesis: it seems totally possible that both men and women like to speak clearly, but that for women it is more difficult to do so.

The speakers' degree of compensation for undersampling can be numerically calculated. When F0 is raised by a factor of 1.37 (as Czech speakers do), the number of harmonics that fit in the F1 space is reduced by a factor of 1.37; when then the F1 of every vowel is raised by a factor of 1.125 (as Czech women do), the number of harmonics that fit in the F1 space is increased by a factor of 1.125, which partly compensates for the loss caused by F0 raising. The size of the compensation can be estimated as $0.125/0.37 = 0.34$, i.e. the female raising of the F1 recovers 34 percent of the harmonics that are lost by raising the F0. If the increase in the F1 range is indeed the (unreliably) estimated factor of 1.10, the fraction of recovered information will be $(1.10 \cdot 1.125 - 1)/0.37 = 64$ percent.

5. Conclusion

We have seen that when Czech women are asked to raise their pitch, they not only raise their F0, but their first and second formants as well. Moreover, we have seen that when speaking at a higher pitch, they tend to disperse their vowels more, even when measured along logarithmic scales. We have argued that the gender dependence of this effect cannot be due to a physiological or socio-phonetic cause. We have therefore concluded that the effect of F0 on the formants, and, importantly, the fact that this effect is present only in female speakers but not in male speakers, can be accounted for by the spectral undersampling hypothesis [6] [7] [8]. Our estimate is that by raising their F1 values, especially those of the open vowels, the F1 space of the female speakers recovers 64 percent of the information that is lost by raising the F0. However, to be more confident about this number, investigating more speakers will be necessary.

6. Acknowledgements

Thanks go to Louis Pols for stimulating comments.

7. References

[1] Peterson. G. E. and Barney, H. L., "Control methods used in a study of the vowels", Journal of the Acoustical Society of America, 24(2):175–184, 1952.

[2] Fitch, W. T. and Giedd, J., "Morphology and development of the human vocal tract: A study using magnetic resonance imaging", Journal of the Acoustical Society of America, 106(3): 1511-1522, 1999.

[3] Fant, G., "Non-uniform vowel normalization", STL-QPSR, 16(2-3):1–19, 1975.

[4] Whiteside, S. P., "Sex-specific fundamental and formant frequency patterns in a cross-sectional study", Journal of the Acoustical Society of America, 110(1):464–478, 2001.

[5] Lieberman, P., "Some aspects of dimorphism and human speech", Human Evolution, 1(1):67–75, 1986.

[6] Goldstein, U., "An articulatory model for the vocal tracts of growing children", D.Sc. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980.

[7] Diehl, R. L., Lindblom, B., Hoemeke, K. A. and Fahey, R. P., "On explaining certain male-female differences in the phonetic realization of vowel categories", Journal of Phonetics, 24:187–208, 1996.

[8] Ryalls, J. H. and Lieberman, P., "Fundamental frequency and vowel perception", Journal of the Acoustical Society of America, 72(5):1631–1634, 1982.

[9] Endres, W., Bambach, W. and Flösser, G. "Voice spectrograms as a function of age, voice disguise, and voice imitation", Journal of the Acoustical Society of America, 49(6B):1842–1848, 1971.

[10] Zetterholm, E., "Same speaker – different voices. A study of one impersonator and some of his different imitations", in P. Warren & C. I. Watson [Eds], Proceedings of the 11th Australian International Conference on Speech Science & Technology, 70–75, 2006.

[11] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer (Version 5.1.02) [Computer program]", retrieved March 9, 2009, from http://www.praat.org, 2009.

[12] Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", IFA Proceedings 17: 97-110, 1993.

[13] Escudero, P., Boersma, P., Schurt-Rauber, A. and Bion, R., "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese", to appear in Journal of the Acoustical Society of America, 2009.

[14] SPSS for Macintosh, Rel. 16.0.1. Chicago: SPSS Inc., 2007.

[15] Sundberg, J., "Data on maximum speed of pitch changes", STL-QPSR, 14(4): 39-47, 1973.

[16] Cleveland, T. F., "Acoustic properties of voice timbre types and their influence on voice classification", Journal of the Acoustical Society of America, 61(6):1622–1629, 1977.