INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Effect of formant and F0 discontinuity on perceived vowel duration: Impacts for concatenative speech synthesis

Tomáš Bořil, Pavel Šturm, Radek Skarnitzl, Jan Volín

Institute of Phonetics, Charles University, Prague, Czech Republic
{tomas.boril, pavel.sturm, radek.skarnitzl, jan.volin}@ff.cuni.cz

Abstract

Unit selection systems of speech synthesis offer good overall quality, but this may be countervailed by a sporadic and unpredictable occurrence of audible artifacts, such as discontinuities in F0 and the spectrum. Informal observations suggested that such breaks may have an effect on perceived vowel duration. This study therefore investigates the effect of F0 and formant discontinuities on the perceived duration of vowels in Czech synthetic speech. Ten manipulations of F0, F1 and F2 were performed on target vowels in short synthesized phrases creating abrupt breaks in the contours at the midpoint of the vowels. Listeners decided in a 2AFC task in which phrase the last syllable was longer. The results showed that despite identical duration of the compared stimuli, vowels which were manipulated in the second part towards centralized values (i.e., less peripheral) were systematically considered to be shorter by the listeners than stimuli without such discontinuities, and vice versa. However, the influence seems to be distinct from an overall formant change (without a discontinuity) since a control stimulus in which the manipulation was performed within the entire vowel was not perceived as significantly shorter or longer. No effect of F0 manipulations was observed.

Index Terms: concatenative synthesis, discontinuity perception, F0, formants, unit selection, vowel duration

1. Introduction

The popularity of HMM-based [1], hybrid [2] or DNN-based [3] speech synthesis has been growing over the past decade, but when it comes to research activities, unit selection concatenative synthesis systems continue to be used in many real-life applications [4] such as Amazon VoiceView [5]; the primary reason for this is their higher naturalness [6]. While the overall quality of the synthesized speech may be higher in unit selection systems, they are plagued by the sporadic, unpredictable occurrence of audible artifacts, even though speech segments are selected from the large source database to meet specific criteria – typically subsumed under 'target cost' and 'join cost'.

It is therefore obvious that the causes which underlie audible artifacts are not adequately captured by the criteria, and we are convinced that there is still space for improvement in detecting audible discontinuities in the speech signal. According to [7], audible artifacts in synthetic speech output may arise due to 1) errors in the database; 2) the imperfect correlation of the target and join costs with human perception; and 3) due to the preference of low global cost over low local cost, so that a unit which results in a local discontinuity may still be part of the globally "cheapest" cost.

Experiments have shown that the most disturbing artifacts in concatenative synthesis can be accounted for by discontinuities in voice fundamental frequency (F0) [8] and spectral information [9], [10]. This study focuses on the effect of these discontinuities on perceived vowel duration.

We are not interested in dynamic changes of formants and F0 within vowels and their impact on perceived vowel duration (see [11] for a review documenting that F0 movement in vowels has been shown to be associated with perceived lengthening). Instead, we focus on discontinuities in formants or in F0 at the concatenation point of two diphones in synthetic speech. This interest was stimulated by auditory analyses of expert phoneticians which revealed that discontinuities in vowel formants and possibly also in the F0 contour often affect the perception of the duration of the vowel harbouring the discontinuity. Our hypothesis – formulated based on our informal listening – is that the disruptive break in the target vowel could also render it "perceptually shorter" (or "longer") than its physical duration would suggest. Since Czech is a language in which vowel quantity is contrastive (*toto nemaž* ['nɛmaʃ] means *don't erase this* while *toto nemáš* ['nɛmaːʃ] means *you don't have this*), the shortening effect may even lead to a change in the meaning of the synthesized sentence. Such an effect would be especially deleterious if it were to appear in a prosodically salient word (i.e., one which is important from the information perspective). In other words, a discontinuity that affects perceived duration could affect not just naturalness, but also intelligibility.

The aim of this study is therefore to examine the effect of formant and F0 discontinuities in an experimental way by means of formant and F0 manipulations and assessing their perceptual impact on listeners. This will extend previous findings about the effects of F0 height [12], vowel height [13] and F0 movement [11] on perceived vowel duration.

2. Method

2.1. Material

We synthesized four target sentences in a male voice using the ARTIC synthesis system (Artificial Talker in Czech) [14], [15]. These were meaningful phrases comprising 7 syllables and 3 stress groups. The target syllable appeared in the final stress group: /ˈprovaːs/ *Tenhle dopis je pro vás. [This letter is for you.]*, /ˈprovas/ *Nejdřív rozmotej provaz. [First disentangle the rope.]*, /ˈprovos/ *Byl tam veliký provoz. [There was heavy traffic.]* and /ˈrɪbiːs/ *Zítra natrhej rybíz. [Tomorrow pick some currant.]*. The relevant context is the final vowel (/a aː o iː/), preceded by /v/ or /b/ and followed by /s/.

http://dx.doi.org/10.21437/Interspeech.2017-1161

First, we modified the duration of target vowels using the overlap-add (pitch synchronous) method in Praat [16] so that the duration was between typical values of short and long vowels given phrase-final lengthening [17] (see Table 1). These slightly modified 'original' stimuli were then resynthesized to maintain the same audio quality as the manipulated stimuli.

Target manipulations (see Table 2) were performed on the second half of the vowel, i.e., from the point where the diphones are concatenated. Additional duration or F0 shifts were carried out using the same overlap-add method. Formant value manipulations were arranged as follows. Mono 16-bit sounds were resampled to 16 kHz, the source sound signal and formants were estimated using the LPC Burg method with the prediction order of 15, window length of 25 ms, time step of 5 ms and pre-emphasis filter starting at 50 Hz. After the relevant formant shifts (using a FormantGrid object in Praat), the resulting sound signal was reconstructed by filtering the source sound signal with the FormantGrid object.

When the resynthesis process resulted in an unacceptable distortion of a consonant, we replaced the distorted segment with the sound from the original signal. However, this was necessary in only very few cases, particularly with some [s] and [n] consonants. This was a legitimate operation as these sounds were not of our testing interest, and the quality of the stimulus, without disturbing artifacts, was essential.

Table 1: *Final vowel of original stimuli. Duration of the vowel, initial and final F0 values of the second half of the vowel, F1 and F2 at the end of the first half.*

	/'prova:s/	/'provas/	/'provos/	/'rɪbi:s/
Duration	145 ms	132 ms	152 ms	159 ms
$F0_{initial}$	105 Hz	85 Hz	85 Hz	103 Hz
$F0_{final}$	92 Hz	76 Hz	80 Hz	94 Hz
F1	688 Hz	614 Hz	529 Hz	259 Hz
F2	1280 Hz	1159 Hz	967 Hz	2109 Hz

Table 2: *Performed manipulations.*

ID	Manipulation
1a / 1b	F0 shifted by +2 ST / −2 ST
2a / 2b	F1 shifted by +11.5% / −11.5%
3a / 3b	F2 shifted by +11.5% / −11.5%
4a	F1 shifted by −11.5%, F2 by +11.5%, excepting /'rɪbi:s/ F1 +11.5%, F2 −11.5%
4b	~ 4a, but in addition F0 shifted by −2 ST
5a / 5b	duration shifted by +30 ms / –30 ms

According to Table 2, the set of 10 new audio files was created for each sentence (see Figure 1 for an example of F1+ shift). Manipulations 1–4 involved no change in the duration of the segments, whereas Manipulations 5 involved only a change in duration. Manipulations 4 (4a without F0 shift, 4b with F0 shift) were intended for shifting both F1 and F2 towards centralized values (i.e., less peripheral), thus [i:] has opposite signs of shifts than the [a: a o] vowels. The ratio of 11.5% formant shift was determined by a preliminary test where the change from 1300 Hz to 1450 Hz led to an observable effect of the duration shift illusion without too much of a perceived vowel quality change.

For the sentence with /a:/, we created another manipulation based on 3b, namely a −11.5% shift in F2 but *in the entire portion of the vowel*. This might help us disambiguate whether a potential effect is due to a discontinuity in formant contours or to a general shift in vowel quality. In total, 45 audio files were created.

Additional items were synthesized and manipulated to be used as distractor items. There were 4 phrases (including /i: a: o ɛ/ in the relevant portion) and 5 manipulations (F0, F1, F2, F1+F2, duration). However, each manipulation was enhanced by a simultaneous difference in duration. This was done to ensure that the listeners had some easy items to process.

In addition, six new stimuli were prepared for the training session. This involved 3 sentences and up to 3 manipulations.

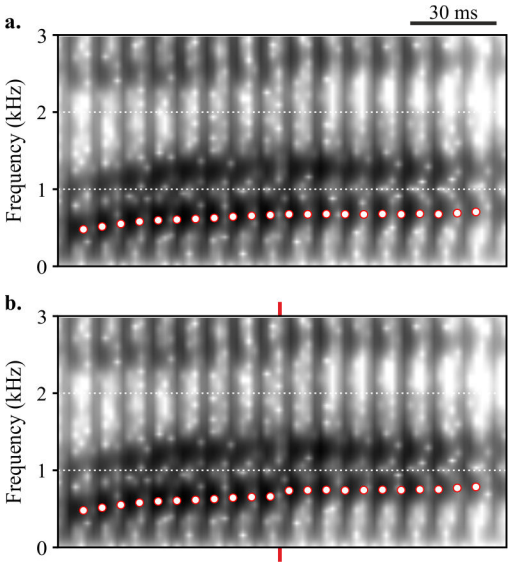

Figure 1: *Comparison of an original* [a:] *vowel* ***(a.)*** *and an F1+ shift* ***(b.)*** *in the* /'prova:s/ *stimuli.*

2.2. Participants

The listening test was administered to 25 respondents (18 females, 7 males, median age = 24). All were native speakers of Czech and studied phonetics at Charles University in Prague; none reported any hearing or speech disorders. The purpose of the experiment was not known to the participants except for 'improving the speech synthesis system'.

2.3. Test procedure

The testing was done in a quiet room using headphones. A 2AFC (two-alternative forced choice) experiment was created in the Praat multiple forced choice (MFC) environment [16] which played a sequence of two phrases (one of them manipulated) with a silent interstimulus interval (ISI) of 1000 ms. The participants decided in which of the two phrases they thought the last syllable was *longer* (they could replay the item three times). They indicated their choice by clicking on a button corresponding to the FIRST or SECOND phrase. Each target item and several distractor items appeared twice (once in 'orig > manip', once in 'manip > orig' order). The order of all items in the test session was randomized for each individual. The items were grouped into blocks, yielding 5 blocks of 20 items and one block of 10 items. Participants listened to 2 minutes of

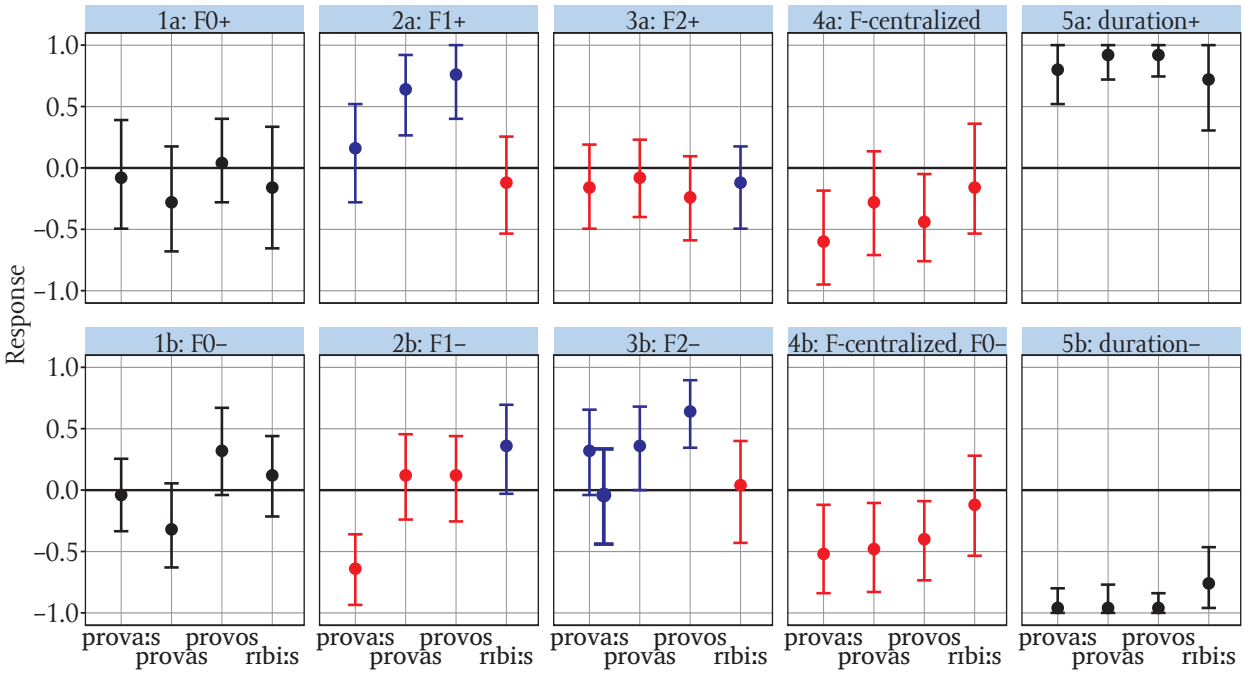

Figure 2: *Evaluations (mean values and conf. intervals) of percieved vowel duration in 4 target sentences (x-axis) and 10 different manipulations with respect to the original stimulus (+1 stands for longer; –1 for shorter perceived duration of the manipulated vowel). Blue = manipulation towards peripheral values, red = towards central values. In 3b, the bold item corresponds to a manipulation performed on the entire portion of the vowel (see text).*

music between the blocks for relaxation. The duration of the test was approximately 30 min.

2.4. Analysis

Since each target item appeared twice, the final score of the item by each listener was obtained as follows. If the *manipulated* item was marked as longer in both cases, the value is +1 ("longer"). If on the other hand the *original* stimulus was consistently rated as longer, the value is −1 ("shorter"). If the listener marked once the manipulated item and once the original stimulus as longer, the resulting value is 0 ("undecided").

This approach leads to values on a three-point scale for each target item. For each stimulus, we computed its mean value and confidence intervals from 25 listeners using a bootstrap method with the significance level $\alpha = 0.05$ and with the Bonferroni correction applied.

The response buttons were always in the opposite order in the second repetition of the same stimulus to eliminate listeners' preference for either the first or second button. Furthermore, there was intentionally no "same / not sure" button in the test. We assumed that even if listeners made the same choice only by random (and thus the resulting value was either +1 or −1), in the case of indistinct stimuli the mean value from all listeners would approach 0 and the variance of answers would increase the confidence interval so it would be clear that the stimulus is generally not evaluated as either longer or shorter.

3. Results

The control condition is represented by Manipulations 5a and 5b, which had longer and shorter objective duration (a 30 ms shift), respectively. The results are clearly distinct in Figure 2, with only /'rɪbi:s/ eliciting some degree of uncertainty in several listeners. Thus, we can assume that listeners were generally reliable and not responding in a random manner.

The F0 manipulation (1a/1b) did not yield a consistent shortening or lengthening effect. The mean value oscillated around 0 ("undecided"), suggesting a random pattern in the responses.

Raised F1 values (2a) were associated with a significant deviation from 0 in the stimuli /'provas/ and /'provos/, but the same words seemed to be immune in the opposite manipulation (2b). Moreover, /'prova:s/ was perceived as significantly shorter with a lower F1, and there was a trend for /'rɪbi:s/ to be considered longer.

Increasing F2 (3a) did not lead to a perceptually shorter or longer vowel, but lowering F2 (3b) yielded significantly longer perceptions with the exception of /'rɪbi:s/ (and with /'prova:s/ bordering on significance). As indicated by the bold black item in 3b, modification of the entire vowel did not lead to any significant effect in perceived duration.

A combination of F1 and F2 shifts (4a) seemed to be perceived as shorter, but a significant difference appeared only for /'prova:s/ and /'provos/. However, with an additional shift in F0 (4b) the manipulation was considered to be significantly shorter by the listeners, except for /'rɪbi:s/. The item /'rɪbi:s/ was thus unaffected by any of the eight target manipulations, which lends itself to interesting interpretations (see below).

4. Discussion

The aim of the experiment was to determine whether the presence of a discontinuity in the F0 or formant contour has a perceptual impact on listeners in terms of perceived vowel duration. The hypothesis stating that the discontinuity shortens or prolongs the vowel is based on our informal observations rather than on previous research because literature generally reports only changes in quality to *entire* vowels [12], [13], or it investigates the effects of dynamic movements [11].

The results of our experiment suggest that if the second half of the vowel is shifted abruptly to centralized, less peripheral formant values, listeners perceive the vowel as shorter, although the objective duration remains unchanged. This effect could be related to what the shift might signal to the listener: a relaxed articulatory setting in the vocalic space may be interpreted as the offset of the vowel, and the listener would then interpret the whole vowel as shorter than it really is. It would be interesting to check whether this effect manifests itself only in phrase-final positions (as in our material), or whether it would work in other places within a phrase. The opposite case, where the second part of the vowel was more peripheral (tense), was indeed interpreted by the listeners as perceptually longer. This could again be related to the speaker's perceived intention, with a more peripheral setting generally associated with longer vowels.

Although Manipulation 3b (F2–) to the stimulus /ˈprovaːs/ did not prove to be significant, it should be noted that the range of the confidence intervals is substantially broadened by applying the Bonferroni correction, and it can be expected that the trend towards lengthening would turn out to be significant with a larger sample of listeners. In contrast, perceived duration in the control condition (in which the same manipulation was performed on the entire portion of the vowel) was centred around 0. This indicates that the observed effect is associated with the discontinuity itself (an abrupt change in formants), and not merely with different duration perception in different vowel qualities.

It is also interesting to compare the Manipulations 2a (F1+) and 2b (F1–). Whereas raising F1 perceptually lengthened the sentences with /ˈprovas/ and /ˈprovos/ (but /ˈprovaːs/ remained unaltered), lowering F1 perceptually shortened only /ˈprovaːs/ (and the other two were without a change). The question is, then, whether on the one hand a phonologically long vowel (/aː/) is in such a position that a further amount of lengthening is perceived bad or unfit and, on the other hand, whether we are insensitive to further shortening of a phonologically short vowel. However, as Manipulation 4b proves, with a more salient change – shifting both formants and F0 – it is possible to achieve further perceptual shortening even of short vowels.

The vowel /iː/ merits special attention, since it is the only vowel which was resistant to the manipulation effects. A potential explanation may be linked to the fact that in Czech there is also a vowel quality difference between the short–long members of the opposition (i.e., [ɪ] × [iː]), and the durational ratio between long–short vowels is much lower than in the other pairs [18], [19]. As a result, Czech listeners exploit both the quantitative and the qualitative differences to identify the vowels, and duration as a cue is therefore less important. Given this interpretation, however, we would have expected formant manipulations in [iː] to yield more consistent shifts in perceived duration, since quality differences are more attended to than in other vowels. That was not the case; only Manipulation 2b (F1–) resulted, in line with the prediction, in the vowel being perceived as longer with marginal significance. (It should be noted here that we are interested in the manipulations from a relative perspective: an F1– change means centralization in [aː] but a more peripheral vowel in [iː], and *vice versa*. Manipulation 2b for [iː] is thus analogous to manipulation 2a for the other vowels.)

As for the manipulations of F2, the lower sensitivity of listeners to acoustic manipulations in [iː] might be explained by the *perceptual integration of formants* [20], [21]. F2 and F3, and possibly even F4, which lie close to each other in [iː], are perceptually merged and can be regarded as one 'effective formant'. Therefore, the manipulation of merely F2 may not be sufficient to change the position of the effective formant in [iː] and, in turn, the perception of its duration.

The effect of F0 discontinuities on the perceived duration of vowels was not confirmed. However, since some results did approach the 0.05 level of significance, it cannot be ruled out that it is possible to obtain such an effect, though it would probably still be weaker than that of formant manipulations. It is obvious that the size of the effect also depends on the size of the manipulation, and it is not clear whether a shift of 2 ST – the value around the just noticeable difference for F0 excursions in real speech [22] – is sufficient to cause a change in perceived duration.

With regard to future research, it is important to examine the behaviour of other vowels and possibly other consonantal contexts. It would also be interesting to test the effect of greater formant and F0 shifts and other (non-flat) trajectories. Only one degree of deviation was used in this study, because keeping the experiment at a reasonable duration (30 minutes) was believed to be more important than investigating all these possible combinations; presenting hundreds of items, even if divided into blocks, greatly reduces the ecological validity of the experiment.

5. Conclusions

The experiment showed an effect of discontinuities in formant trajectories on the perceived vowel duration in Czech. It would be of interest to examine whether this phenomenon occurs also in other than phrase-final positions, and whether languages without distinctive vowel quantity contrasts would display the same pattern.

The ultimate objective of this study was to contribute to improving the Czech unit selection speech synthesis. From our preliminary results, it would seem beneficial to assign stricter penalizations to spectral (formant) discontinuities when concatenating vowels, especially in prosodically salient positions where shortening or lengthening will, presumably, be most detrimental.

6. Acknowledgements

This research was supported by the Czech Science Foundation project No. 16-04420S.

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[2] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1278–1288, 2011.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks", in *Proc. ICASSP 2013*, pp. 7962–7966, 2013.

[4] T. Dutoit, “Corpus-based speech synthesis,” in J. Benesty, M. Sondhi, and Y. Huang (Eds.), *Springer Handbook of Speech Processing*, pp. 437–455. Dordrecht: Springer, 2008.
[5] Amazon, “Accessibility for Kindle,” retrieved on Feb 25, 2017 from https://www.amazon.com/b?node=14100715011, 2016.
[6] S. King, “Measuring a decade of progress in Text-to-Speech,” *Loquens*, vol. 1(1), 2014.
[7] J. Matoušek, D. Tihelka, and M. Legát, “Is unit selection aware of audible artifacts?” in *8th ISCA Speech Synthesis Workshop, August 31 – September 2, Barcelona, Spain, Proceed*ings, 2013, pp. 267–271.
[8] M. Legát and J. Matoušek, “Pitch contours as predictors of audible concatenation artifacts,” in *Proceedings of the World Congress on Engineering and Computer Science 2011, San Francisco, USA*, 2011, pp. 525–529.
[9] J. Wouters and M. W. Macon, “Control of spectral dynamics in concatenative speech synthesis,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 30–38, 2001.
[10] E. Klabber and R. Veldhuis, “Reducing audible spectral discontinuities,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.
[11] R. Cumming, “The effect of dynamic fundamental frequency on the perception of duration,” *Journal of Phonetics*, vol. 39, pp. 375–387, 2011.
[12] A. C. L. Yu, “Tonal effects on perceived vowel duration,” in C. Fougeron, B. Kühnert, M. D’Imperio and N. Vallée (Eds.), *Papers in Laboratory Phonology (Vol. 10)*, pp. 151–168. Berlin: Mouton de Gruyter, 2010.
[13] C. Gussenhoven, “Perceived vowel duration,” in H. Quené and V. van Heuven (Eds.), *On Speech and Language: Studies for Sieb G. Nooteboom*, pp. 65–71. Utrecht: LOT, 2004.
[14] J. Matoušek, D. Tihelka, and J. Romportl, “Current state of Czech text–to–speech system ARTIC,” in *Proc. of the 9th International Conference TSD 2006, Lecture Notes in Artificial Intelligence, vol. 4188*. Springer Berlin / Heidelberg, 2006, pp. 439–446.
[15] D. Tihelka, J. Kala, and J. Matoušek, “Enhancements of Viterbi Search for Fast Unit Selection Synthesis,” *Proc. Interspeech 2010*, pp. 174–177.
[16] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” version 6.0.25, retrieved 12 February 2017 from http://www.praat.org/.
[17] J. Volín and R. Skarnitzl, “Temporal downtrends in Czech read speech,” *Proc. Interspeech 2007*, pp. 442–445.
[18] V. J. Podlipský, R. Skarnitzl and J. Volín, “High front vowels in Czech: A contrast in quantity or quality?,” *Proc. Interspeech 2009*, pp. 132–135.
[19] R. Skarnitzl, “Dvojí *i* v české výslovnosti [Two kinds of *i* in the pronunciation of Czech]”, *Naše řeč*, vol. 95, pp. 141–153, 2012.
[20] J.-L. Schwartz and P. Escudier, “Does the human auditory system include large scale spectral integration?,” in M. E. H. Schouten, (Ed.), *The Psychophysics of Speech Perception*, pp. 284–292. Dordrecht: Martinus Nijhoff Publishers, 1987.
[21] L. A. Chistovich and V. V. Lublinskaya, “The ‘Center of gravity’ effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli.” *Hearing Research*, vol. 1, pp. 185–195, 1979.
[22] M. S. Harris and N. Umeda, “Difference limens for fundamental frequency contours in sentences,” *Journal of the Acoustical Society of America*, vol. 81, pp. 1139–1145, 1987.