AKUSTICKÉ LISTY

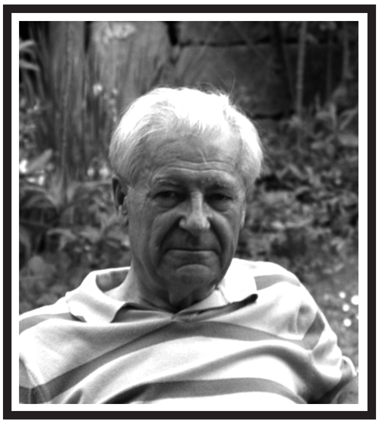
České akustické společnosti
www.czakustika.cz

ročník 25, číslo 1–2 červen 2019

Obsah

ČESKÁ AKUSTICKÁ SPOLEČNOST

Zemřel prof. Dr. Ing. Jiří Tichý, CSc.



S lítostí vám musíme oznámit, že nás navždy opustil profesor Jiří Tichý. Narodil se 4. července 1927 v Bratislavě a hned po válce v roce 1945 začal studovat na Vysoké škole strojního a elektrotechnického inženýrství v Praze. Absolvoval v roce 1950, ale již od roku 1947 byl pomocnou vědeckou silou ve fyzikálním ústavu a později na katedře fyziky u profesora Josefa B. Slavíka. V této době získal doktorát technických věd (1953), titul kandidáta věd a v roce 1965 se habilitoval jako docent. Již za svého pobytu na katedře publikoval řadu odborných statí a učebních textů, oblíbené byly i jeho přednášky z fyziky. Věnoval se zejména otázkám zvukové pohltivosti a problémům prostorové akustiky.

Po roce 1968, kdy odešel do Spojených států amerických, se nejprve stal profesorem na katedře architektury Pensylvánské státní univerzity, na které působil do roku 1973. Od roku 1973 až do léta 1997 vedl Graduate Program in Acoustics, který zahrnoval několik desítek profesorů nabízejících více než sto kurzů z celé oblasti akustiky a souvisejících oborů. Počet absolvujících studentů (Ph.D, M.S. a M. Eng.) přesáhl za jeho působení sto. Byl aktivní v řadě prestižních vědeckých společností nejen ve Spojených státech, kromě České akustické společnosti (původně Československé) byl například členem Japonské akustické společnosti nebo New York Academy of Sciences. V letech 1986–1987 byl předsedou společnosti Institut of Noise Control Engineering. Pracoval v oblasti mezinárodní normalizace: ISO/TC43 – Akustika, IEC/TC 29 – Elektroakustika a v Americké společnosti pro normalizaci (ANSI). Americkou akustickou společností byl v roce 1991 zvolen místopředsedou společnosti a v roce 1993 předsedou společnosti. Tuto náročnou funkci zastával až do roku 1995. Předsedal mnoha národním a mezinárodním konferencím a seminářům po celém světě a také je organizoval (Švédsko, Francie, Japonsko, Singapur, Brazílie a samozřejmě USA), například v roce 1999 uspořádal první společný kongres Americké akustické společnosti a Evropské akustické asociace v Berlíně, což bylo největší setkání akustiků v minulém století. V roce 1998 byl jmenován čestným členem České akustické společnosti.

Ve své vědecké práci ve Spojených státech navázal na své výsledky z Československa a na katedře architektury Pensylvánské státní univerzity se věnoval stavební a prostorové akustice. Po roce 1975 se pak zaměřil na otázky snižování hluku strojů. Pracoval na vývoji a použití akustické intenzity, později akustické holografie. Více než patnáct let bylo jeho hlavním odborným zájmem aktivní snižování hluku a další aplikace aktivních metod v akustice. Výsledky svých prací prezentoval ve více než 30 vyzvaných přednáškách a 60 odborných referátech na národních a mezinárodních konferencích. I další publikační činnost profesora Tichého je velmi rozsáhlá. Je autorem více než 80 zásadních článků, spoluautorem sedmi knih. V poslední z nich s názvem *Acoustics of Small Rooms* se vrátil ke své oblíbené prostorové akustice. Více než deset výzkumných zpráv souvisí s jeho konzultační činností pro takové firmy, jako jsou Magnavox, Ford Motor, IBM, Applied Acoustics Research Corporation apod.

Po roce 1989 se profesor Tichý pravidelně vracel do Prahy, kde pomáhal s rozvojem akustiky zejména na své alma mater. Třikrát zde přednášel semestrální kurz Architekturní akustiky a kurz Aktivního snižování hluku. Umožnil také některým kolegům návštěvu Pensylvánské státní univerzity. Významně pomáhal při pořádání kongresu Inter-noise 2004, který se konal v Praze. V Praze také 27. března 2019 zemřel. Čest jeho památce.

Discontinuities in fundamental frequency: When do they really matter in synthetic speech?

Nespojitosti základní frekvence: kdy mají v syntetické řeči vliv?

Tomáš Bořil and Radek Skarnitzl

Charles University, Faculty of Arts – Institute of Phonetics, náměstí Jana Palacha 2, 116 38 Praha 1

Attempts at improving the naturalness of synthetic speech have typically led to penalizing unsuitable candidates or large differences in acoustic parameters around the concatenation point. This paper reports a perceptual experiment which aimed at the opposite: relaxing the criteria for concatenation cost in the domain of fundamental frequency (f_0), specifically when concatenating diphones pertaining to voiced consonants. A listening test which involved several types of artificial f_0 discontinuities was administered to 21 respondents. The results suggest that f_0 discontinuities only matter in sonorant consonants (nasals and approximants) and only when they exceed 1 semitone. Most importantly, the direction of f_0 change should be taken into account, and not only the values around the concatenation point.

1. Introduction

Concatenative speech synthesis systems based on dynamic unit selection continue to dominate real-life applications, although research endeavours have, to a large extent, moved away from this relatively costly approach to generating artificial speech. It is the still superior naturalness of concatenative speech synthesis which lies behind this continued preference [1]. However, the output of concatenative synthesis may suffer from the sporadic occurrence of audible discontinuities. These artefacts, which may have an intrusive effect on the listener, may have several causes, as summarized by [2]. First, the database from which units (typically diphones) are selected for synthesis may feature some errors, either random or systematic (see [3] and also [4] for a proposal to eliminate some of the latter ones from the Czech synthesis system ARTIC [5]); this is the case especially in languages with a more or less straightforward relationship between spelling and pronunciation like Czech. Second, the *target cost* and *concatenation cost*, two functions governing the selection of units from the database, may not correlate perfectly with human perception and may thus fail to capture some audible discontinuities. Finally, because selection algorithms typically prefer a low global cost over a low local cost, the globally "cheapest" set of selections may feature a local artefact at a specific concatenation point.

A number of experiments have addressed the question of artefacts in concatenative synthesis. The most intrusive effect on the listener seems to be exerted by "jumps" in the fundamental frequency (f_0) of the voice [6], [7] and by discontinuities in the spectral domain [8], [9]. Many past attempts at improving the specification of the *target* and *concatenation cost* have focused on stipulating penalties concerning, for instance, the permissible difference in the acoustic parameters of neighbouring diphones or the context in which the source and target diphones could appear.

Naturally, the more rules there are and the more potential diphone candidates are penalized, the fewer units remain for selection. That is why, in our most recent attempts at improving the ARTIC synthesis system, we have adopted an opposite perspective: we are applying phonetic experimentation to investigate in which specific contexts a given acoustic difference does need to be taken into account in calculating the concatenation cost, and when a difference of, stated objectively, the same or even greater magnitude may be ignored because the acoustic discontinuity is not perceptually detectable. This study addresses fundamental frequency which, according to our informal observations, continues to be one of the most frequent sources of intrusive artefacts in the ARTIC synthesis system. In the current implementation of ARTIC [5], the transition of f_0 between neighbouring diphones is part of the *concatenation cost* calculation in all voiced segments. The aim of this study is to verify whether this is necessary, or whether acoustic (objective) discontinuities may be ignored in some contexts because they are not perceptible.

2. Fundamental frequency vs. perceived pitch

First of all, it must be emphasized that the f_0 contour (an output of a f_0 extractor) does not correspond to the pitch contour (the subjective percept of pitch movements); in other words, listeners do not perceive pitch objectively. There are several components of the discrepancy between an f_0 contour and its corresponding pitch contour. Researchers often talk about *pitch contour stylization*, which refers to such an approximation of the extracted f_0 contour so that it is perceptually indistinguishable (or at least so that it perceptually resembles) from the original [10], [11].

Přijato 14. října 2018, akceptováno 21. ledna 2019.

The first step in bringing f_0 and pitch closer to each other consists in expressing differences in a psychoacoustic unit rather than in the physical unit Hertz; it was found that semitones (ST) best correspond to the perceptual impression of pitch [12]. The next important component that has to be accounted for concerns the so-called microprosodic variations [13], [14], where f_0 is affected by the voicing status of the surrounding consonants; these small perturbations are not perceptible and have to be eliminated. We can state in general that f_0 changes of short durations and small magnitudes are not perceptible [10], [11]; that is why the f_0 contours should always be smoothed (i.e., lowpass-filtered).

Another important aspect of pitch perception is the alignment of perceived pitch to the segmental chain. As summarized by [11] or [15], evidence suggests that we perceive pitch mostly in syllabic nuclei (i.e., typically vowels, sometimes sonorant consonants), most likely in their central portion. Most frequently, every syllable is perceived as having one tone; it is only in final syllables of prosodic phrases, which carry the nuclear tone and where syllabic nuclei are sufficiently lengthened, where we perceive melodic changes [11].

If we consider these findings from the opposite perspective, it is clear that f_0 changes in consonants should not contribute to the perceived pitch contour. That does not automatically mean, however, that larger f_0 jumps occurring within consonants may not be audible. The main research question of the current study therefore is whether discontinuities in fundamental frequency, when concatenating diphones pertaining to a consonant, will have an intrusive effect on listeners. More specifically, we want to examine whether there is a threshold beyond which the f_0 jump is already perceptible, whether a larger context of the f_0 contour may play a role in the perceptual judgements, and whether this effect applies to all consonant classes. Since this is an exploratory study, we only formulate a general hypothesis: it is predicted that listeners will not be equally sensitive to all types of f_0 discontinuities.

3. Method

To investigate the effect of f_0 discontinuities, it was essential to use very short sound stimuli and manipulate them in a strictly controlled manner. As source material, we used recordings of [aCa] disyllables, where the voiced intervocalic consonant (C) included two plosives [b, d], two fricatives [z, ʒ], two nasals [m, n], two liquids [l, r], and also [v], a voiced fricative which, however, retains some properties of sonorant sounds [13]. These source disyllables were recorded by 4 female and 4 male native speakers of Czech; an EGG signal using the VoceVista system [14] was recorded alongside the audio to ensure completely reliable f_0 values. Attention was paid during the recording that the intervocalic consonant was pronounced with full voicing (obstruents frequently lose some voicing in intervocalic positions [15]).

The subsequent manipulations of f_0 were performed by means of PSOLA [16] in Praat [20] on these source disyllabic recordings, using a Praat script. The time points used for the manipulations, stipulated based on [21], are shown in Fig. 1.

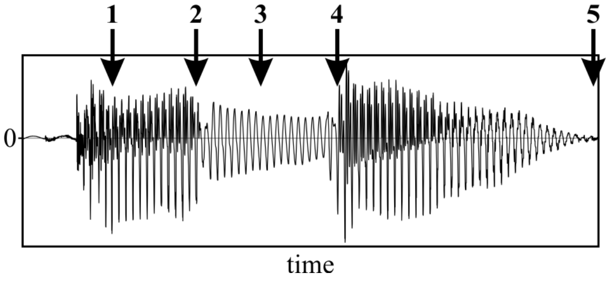


Figure 1: Time points of disyllables indicated in the waveform of [aba]. 1: onset of the periodic part of vowel 1; 2: consonant onset; 3: midpoint of consonant (or of its closure phase); 4: consonant offset; 5: offset of vowel 2

We simulated f_0 jumps of 1 and 5 semitones (ST) around time point 3; these intervals were selected because they seem to correspond to the range of f_0 discontinuities encountered during an analysis of the ARTIC synthesis outputs. There were two types of experimental manipulations, as illustrated in the left panel of Figure 2. The two types differ in how f_0 changes between time points 2 and 4, i.e., during the target consonant (or, in case of plosive sounds, during their closure phase). In the first type, f_0 remained stationary before and after the jump itself; this type, which is based on the Heaviside step function, will be henceforth referred to as type H (see the f_0 contours in H1 and H5 in Fig. 2). In the second type, f_0 was manipulated so that it changes during the consonant beyond the 1- or 5-ST jump itself; importantly, the change is in the opposite direction with respect to the target jump, resembling a sawtooth. Specifically, f_0 remained stationary in the vowel, then dropped by 0.5 or 2.5 ST respectively during the first half of the consonant (or, in the case of plosives, of the closure phase, between time points 2 and 3), jumped up abruptly by 1 or 5 ST respectively (this is the target f_0 jump), and dropped again by 0.5 or 2.5 ST during the second half of the consonant, between time points 3 and 4. This sawtooth-like change is henceforward referred to as type S. The target f_0 jump always occurred within 2 milliseconds, which is comparable to jumps occurring in synthetic speech. In total, this yielded four types of modified stimuli.

As shown in the right panel of Fig. 2, a “default” version was created as a control to each of the experimental manipulations, which involved either flat f_0 or a “natural” jump (i.e., one which may occur in ordinary speech), around time point 2 (i.e., at the onset of the intervocalic consonant). The objective was to generate pairs of disyllables which – if the performed manipulation were not per-

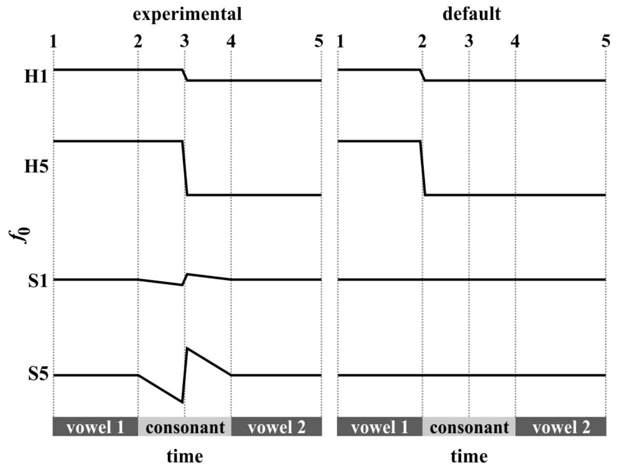


Figure 2: Four types of experimental manipulations on the left – Heaviside (H) and sawtooth (S) jumps by 1 and 5 ST – with their corresponding default versions on the right (see text)

ceptible – would have identical perceptual effect (i.e., their perceived intonation would be the same).

In total, the study is based on 8 speakers, 4 female and 4 male. Since we did not want listeners to have to perceptually "switch" between male and female voices, as the acoustic differences between the stimuli (caused by the manipulations) are very small, we created two tests, and the listeners were randomly divided into two groups, listening only to male or only to female stimuli. The manipulated and default variants were used to create a listening test. Each test item consisted of a pair of stimuli, one default and one manipulated. In total, the listening test contained 144 items (4 speakers × 9 consonants × 8 variants). No test items were repeated.

The listening test was administered to 21 respondents via ARTIC-Tests 3.0, a web-based environment created by the West Bohemian University in Pilsen (11 respondents evaluated the female stimuli, 10 evaluated the male stimuli); all were students at Charles University, Faculty of Arts. The respondents' task was to listen to random-order sorted items consisting of two sounds (one always being the manipulated, the other the default version, in random order) and to decide whether one of them sounded intrusive or whether both sounded the same. They indicated their choice by clicking one of three radio buttons: the first sound is worse, the second sound is worse, they are both of equal quality. They were allowed to repeat each sound at will. The listeners were instructed to use closed headphones. Since the sound stimuli were very short, the entire listening test, with the 144 items, did not last longer than 15 minutes.

Statistical analyses were carried out using R [22], and graphical outputs were created using the R package *ggplot2* [23].

4. Results

The listeners' responses were associated with values as follows: 1 = the manipulated stimulus sounds worse; 0 = both sounds are of equal quality; and −1 = the default stimulus sounds worse. Figure 3 shows these results split into groups by combining the consonant in the disyllable, manipulation type (H and S), and size of the manipulation interval (1 and 5 ST). For each group we calculated the mean value and estimated confidence intervals using the bootstrap method with a significance level of 0.05 (Bonferroni-corrected for multiple testing). This means that a null hypothesis of no noticeable hearing difference between the manipulated and default version of a stimulus cannot be rejected if the confidence interval includes the value of 0.

It is immediately apparent that the listeners perceived no clear difference in the quality of the sound when Heavi-

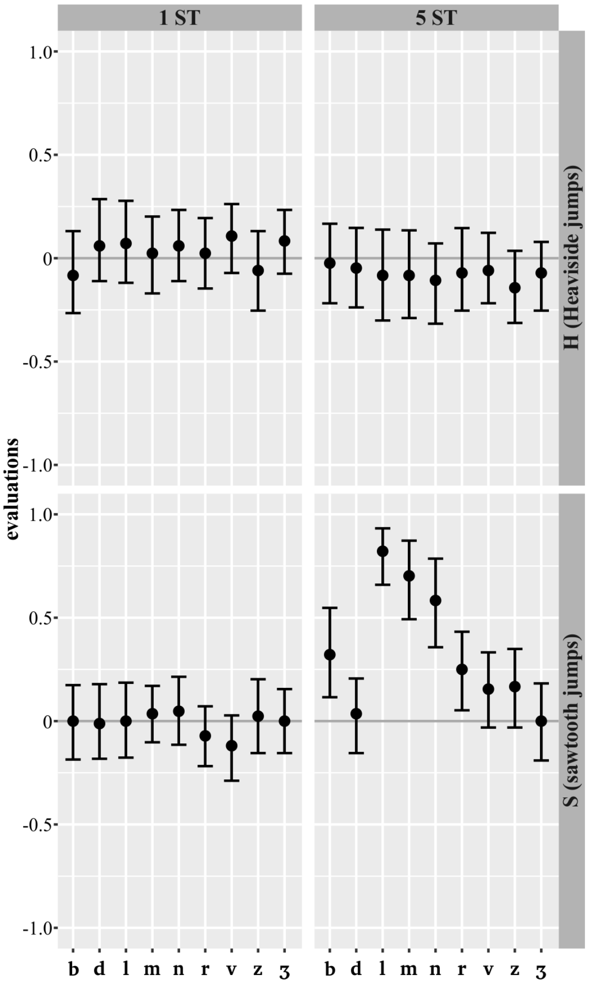


Figure 3: Responses for individual consonants to the Heaviside (H, top) and sawtooth (S, bottom) discontinuities of 1 semitone (left) and 5 semitones (right); see section 3 for more details. The evaluation of 1.0 corresponds to the manipulated stimulus sounding worse, 0 to no difference in evaluation (i.e., chance level), and the evaluation of −1.0 to the default stimulus sounding worse

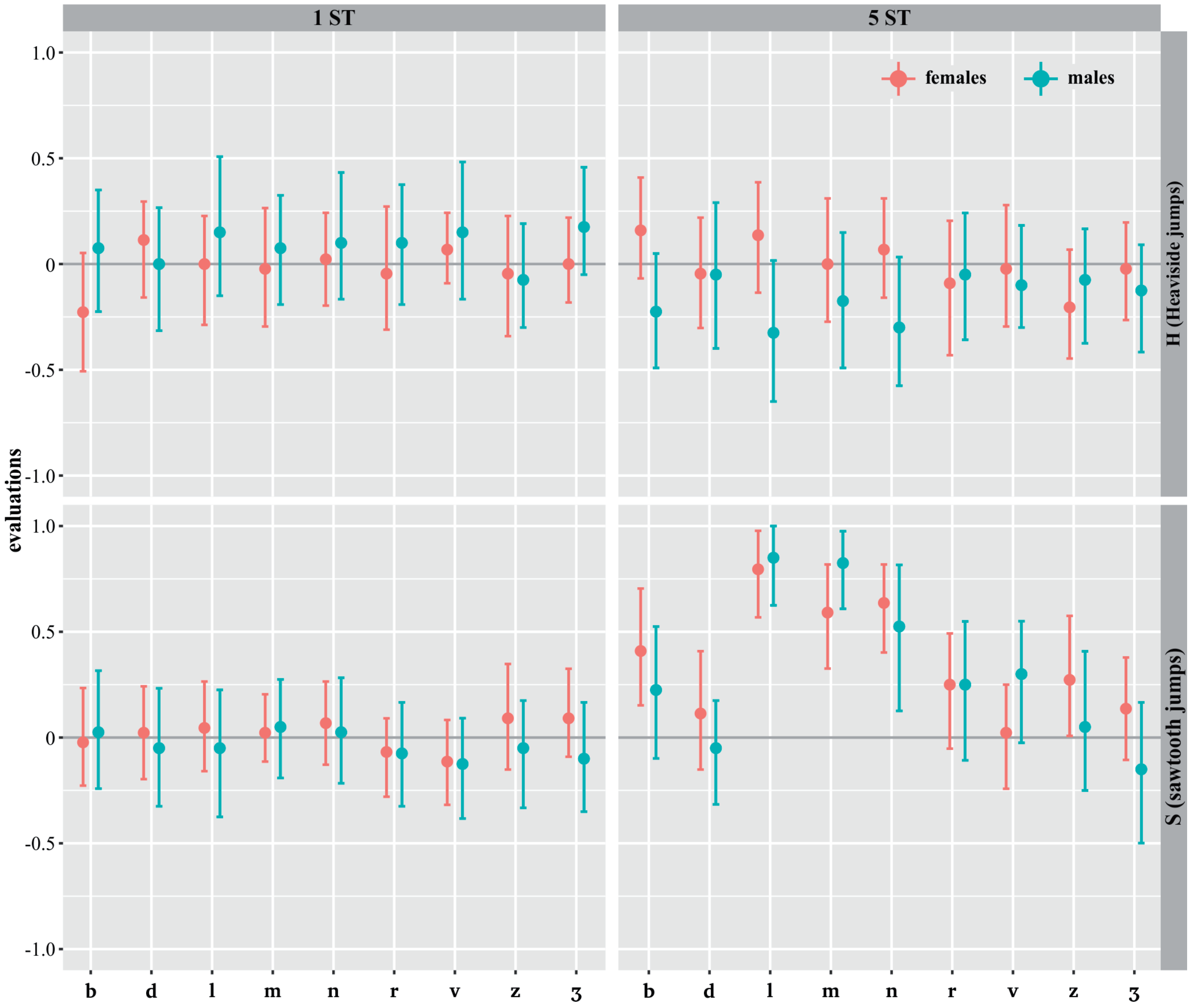


Figure 4: Responses for individual consonants to the Heaviside (H, top) and sawtooth (S, bottom) discontinuities of 1 semitone (left) and 5 semitones (right), separately for male and female speakers; see section 3 for more details. The evaluation of 1.0 corresponds to the manipulated stimulus sounding worse, 0 to no difference in evaluation, and the evaluation of −1.0 to the default stimulus sounding worse

side jumps (marked H) were concerned: confidence intervals for all consonants intersect the value of 0, irrespective of the f_0 manipulation interval. While the same applies for all the consonants in which a sawtooth (S) discontinuity of 1 semitone was introduced (see the bottom left panel of Figure 3), some sawtooth f_0 discontinuities in the order of 5 semitones clearly do matter. It can be seen that it is especially the sonorant consonants (i.e., [l, m, n, r]) and also the plosive [b] where the listeners could hear a difference in the quality of the sound. Specifically, the manipulated stimuli were perceived as significantly inferior in comparison with the default versions.

In Figure 4 the results of the listening test are shown separately for the female and male speakers. Each female-speaker group consists of 44 values (4 speakers × 11 respondents) and each male-speaker group consists of 40 values (4 speakers × 10 respondents). It was not the purpose of this study to examine the effect of speaker sex; for that our data would not be sufficient. The figure merely shows that there may be some small differences in the results, which may be caused by the specificity of the individual voices.

Naturally, these separated results are comparable to the pooled data presented in Figure 3. First, the Heaviside discontinuities in f_0, of either 1 ST or 5 ST, do not seem to be perceptually salient in any of the examined consonants, as indicated in the top panel of Figure 4. Again, the same applies for the sawtooth discontinuities of 1 ST. In addition to these similarities, however, there are some differences in the bottom right quadrant of the figure which are worth pointing out.

Most importantly, it can be seen that the manipulated stimuli of the sonorants [l, m, n] were evaluated as significantly worse in quality than their corresponding default stimuli. While the pooled evaluation for the trill [r] did reach statistical significance, the evaluation is not significant when the stimuli from male and female speakers are considered separately. The figure also suggests that the significant effect in the assessment of the plosive [b] was pulled by the responses to the female speakers' stimuli.

5. Discussion and conclusions

The objective of this exploratory study was to investigate in greater detail the perceptual aspects of concatenating diphones, where the concatenation involves various kinds of discontinuities in the fundamental frequency of the voice (f_0). Although the listening test itself was not excessively long and the web-based environment allowed the respondents to interrupt the experiment and resume it later, informal post-hoc queries from some of the respondents indicated that the listening was tedious. More specifically, what may have been slightly frustrating for the listeners was the inevitable tendency that, in line with our predictions, many stimuli pairs would sound the same in terms of their quality. We therefore believe that the fact that positive results were obtained – i.e., that the listeners diligently compared the stimuli throughout the 144 items – is worth emphasizing.

The results of the presented experiment are positive in several aspects. First, they confirm previous findings related to the perception of pitch (see [11] or [15]), but make them more detailed. The most important implications are related to our ultimate aim, which was to simplify the selection of diphones for concatenative speech synthesis using dynamic unit selection. Our results show that acoustic discontinuities at the point of concatenation within a consonant which are smaller than 1 semitone do not seem to be perceptually relevant. Based on this finding, f_0 jumps smaller than (at least) 1 ST can be ignored when concatenating diphones pertaining to any voiced consonant.

More interesting are our findings regarding the nature of the introduced discontinuity. The upper right panels of Figures 3 and 4 suggest that even discontinuities of 5 semitones do not lead to an intrusive perceptual effect, if they are "smooth" in the sense that the f_0 contour in the vicinity of the jump does not involve movement contrary to the jump (these changes were labelled H, as they resemble the Heaviside function). It is only 5-ST jumps which involve a more salient change of direction of the f_0 contour – these were labelled S for sawtooth – that have resulted in a significant perceptual effect. The conclusion that can be drawn from this result is that it would be highly beneficial to calculate f_0 not only in the frames closest to the concatenation point, but to also incorporate the direction of f_0.

To provide a more specific example, extrapolating on our results, we may hypothesize that the discontinuity in the f_0 track marked as A in Figure 5 will not be perceptually salient, while that marked as B – which involves exactly the same jump around the concatenation point in terms of its magnitude but one that is sawtooth-like – most likely will.

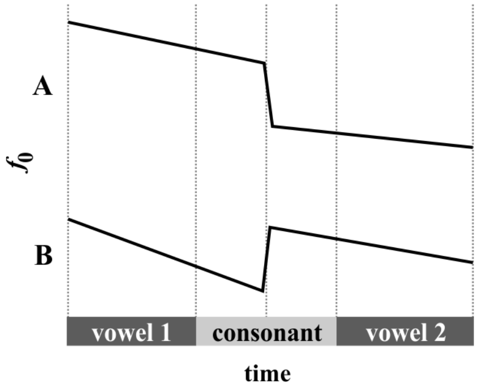


Figure 5: Schematic illustration of two "identical" f_0 discontinuities around the concatenation point (see text)

Finally, let us turn to the finding which concerns the individual consonants (or more precisely, consonant types). As mentioned in the Introduction, the current implementation of the ARTIC system [5] considers f_0 in all voiced segments to determine the *concatenation cost*. The results of this study prove that this not necessary. Perceptually salient artefacts have been conclusively obtained only for the sonorant sounds, specifically the nasals (in Czech, this would be [m n ɲ ŋ]) and the lateral approximant [l]; it may be assumed that the same would apply to the palatal glide [j]. The manipulated stimuli of the trill [r] – also classified as a sonorant sound – were also evaluated as worse in the pooled data. On the other hand, the significantly worse evaluation of manipulated [b] items is not straightforward.

To conclude, this experiment aimed at simplifying unit selection when it comes to incorporating f_0 in the *concatenation cost* when concatenating diphones pertaining to voiced consonants. The results show that the direction of f_0 change needs to be taken into account, that only sonorant sounds should be considered, but only when the discontinuity exceeds 1 semitone. It may be worthwhile to conduct a mode detailed experiment which would determine with greater precision where between 1 and 5 ST the boundary of perceptual intrusiveness lies.

While this study was motivated by audible artefacts in the Czech speech synthesis ARTIC [5], it is to be expected that our results may be applicable in any speech synthesis algorithm which makes use of the f_0 criterion in the computation of the *concatenation cost*. Although savings in terms of computation time or in terms of the number of diphones which would have been previously eliminated from selection and retained after the inclusion of the proposed relaxed criteria have not been examined, we assume that especially the latter aspect – having more diphones available for concatenation – is an important result of this experiment.

Acknowledgements

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

References

[1] Dutoit, T.: Corpus-based speech synthesis, in Benesty, J., Sondhi, M., Huang Y. (Eds.), *Springer Handbook of Speech Processing*, p. 437–455. Springer, Dordrecht, 2008.

[2] Matoušek, J., Tihelka, D., Legát, M.: Is unit selection aware of audible artifacts? *Proc. 8th ISCA Speech Synthesis Workshop 2013*, p. 267–271, 2013.

[3] Matoušek, J., Tihelka, D.: Anomaly-based annotation error detection in speech-synthesis corpora, *Computer Speech & Language*, 46, p. 1–35, 2017.

[4] Skarnitzl, R.: Alofonická variabilita v češtině z pohledu řečové syntézy, *Akustické listy*, 24, p. 15–20, 2018.

[5] Tihelka, D., Hanzlíček, Z., Jůzová, M., Vít, J., Matoušek, J., Grůber, M.: Current state of text-to--speech system ARTIC: A decade of research on the field of speech technologies, in Sojka, P., Horák, A., Kopeček, I., Pala, K. (Eds.), *Text, Speech, and Dialogue, TSD 2018.* Lecture Notes in Computer Science, vol. 11107. Springer, Cham, 2018.

[6] Legát, M., Matoušek, J.: Pitch contours as predictors of audible concatenation artifacts, *Proc. World Congress on Engineering and Computer Science 2011*, p. 525–529, 2011.

[7] Dutoit, T.: Corpus-based speech synthesis in Benes ty, J., Sondhi, M. M. Huang, Y. (Eds.), *Springer Handbook of Speech Processing*, p. 437–455. Springer, Berlin, 2008.

[8] Klabber, E., Veldhuis, R.: Reducing audible spectral discontinuities, *IEEE Transactions on Speech and Audio Processing*, **9**(1), p. 39–51, 2001.

[9] Bořil, T., Šturm, P., Skarnitzl, R., Volín, J.: Effect of formant and F0 discontinuity on perceived vowel duration: Impacts for concatenative speech synthesis, *Proc. Interspeech 2017*, p. 2998–3002, 2017.

[10] Hart, J., Collier, R., Cohen, A.: *A perceptual study of intonation: An experimental-phonetic approach to speech melody*, Cambridge University Press, Cambridge, 1990.

[11] Hermes, D. J.: Stylization of pitch contours, in Sudhoff, S., Lenertová, D. Meyer, R., Pappert, S., Augurzky P, Mleinek, I., Richter, N., Schließer, J. (Eds.), *Methods in Empirical Prosody Research*, p. 29–62. De Gruyter, Berlin, 2006.

[12] Nolan, F.: Intonational equivalence: An experimental evaluation of pitch scales, in *Proc. 15th ICPhS*, Vol. 1, p. 771–774, 2003.

[13] Lehiste, I., Peterson, G. E.: Some basic considerations in the analysis of intonation, *Journal of the Acoustical Society of America*, 33, p. 419–425, 1961.

[14] Hanson, H. M.: Effects of obstruent consonants on fundamental frequency at vowel onset in English, *Journal of the Acoustical Society of America*, 125, p. 425–441, 2009.

[15] Volín, J.: Extrakce základní hlasové frekvence a intonační gravitace v češtině, *Naše řeč*, 92, p. 227–239, 2009.

[16] Skarnitzl, R., Volín, J.: Czech voiced labiodental continuant discrimination from basic acoustic data, in *Proc. Interspeech 2005*, 2921–2924, 2005.

[17] Miller, D. G., Nair, G., Schutte, H., Horne, R.: *VoceVista* version 3.2, 2017.

[18] Skarnitzl, R.: *Znělostní kontrast nejen v češtině*, Nakladatelství Epocha, Praha, 2011.

[19] Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, **9**(5–6), p. 453–467, 1990.

[20] Boersma, P., Weenink, D.: *Praat: doing phonetics by computer*, version 6.0.25, retrieved 12 February 2017 from `http://www.praat.org/`.

[21] Machač, P., Skarnitzl, R.: *Principles of Phonetic Segmentation*, Praha, Epocha, 2009.

[22] R Core Team: *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved 1 September 2018 from `https://www.R-project.org/`.

[23] Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*, New York, Springer-Verlag, 2016.

The mapping of voice parameters in connected speech of healthy Common Czech male speakers

Mapování hlasových parametrů v souvislé řeči zdravých mužských mluvčích obecné češtiny

Lea Tylečková and Radek Skarnitzl

Charles University, Faculty of Arts – Institute of Phonetics, náměstí Jana Palacha 2, 116 38 Praha 1

This study examines a set of voice parameters to map objective ranges of voice-source characteristics of healthy male speakers of Common Czech. Objective assessment of voice quality is conducted mainly in speakers with voice pathologies, typically using sustained vowels as basis for measurements. In our study, we focused on non-pathological voices and performed acoustic measruments of the voice parameters which are believed to reflect glottal characteritics. The analyses were based on the open vowels [a aː] extracted from fifty healthy male speakers who performed a reading task. Voice parameter estimation included f_0 perturbation measures (jitter and shimmer), harmonicity (HNR), Cepstral Peak Prominence (CPP), and harmonic amplitude measures which reflect short-term spectral slope (e.g., H1−H2, H2−H4, or H1−A3). The obtained data relate to connected speech and are compared to the measurements on sustained vowels.

1. Introduction

The role of voice in everyday social interactions could hardly be underestimated; it is an important part of our communication and it also represents a rich source of information about the speakers reflecting their physical, psychological and social characteristics [1]. Voice quality can be treated in a broad perspective, when it comprises specific settings at both the laryngeal and supra-laryngeal (articulatory) level [1]. In a narrower perspective, voice quality only refers to phonatory modifications (i.e., changes in the manner of vocal fold vibra-tions). In this paper, we are interested in the laryngeal level only, and voice quality will thus pertain only to phonation.

Differences in voice quality may arise due to anatomical and physiological factors; apart from these biological aspects, however, socio-cultural aspects also play a considerable role [2, 3]. Voice quality as a significant idiosyncratic aspect of an individual's speech pattern is also examined within the field of forensic phonetics. Acoustic analyses focus on measuring voice parameters enabling to capture inter-speaker variability. In the Czech context, this research area is addressed, for instance, by Weingartová et al. [4].

Generally, when assessing voice quality, speech scientists may make use of methods deriving from three viewpoints: articulatory, where we describe the phonatory behaviour *per se*, perceptual and acoustic. Perceptual ratings of voice quality reflect subjective assessment but the overall impression of the voice can be decomposed into a few dimensions that are perceptually distinct and correspond to various terms, such as breathiness, roughness etc. Assessing voice quality using perceptual rating scales [1, 5, 6, 7] should remain constant across different listeners and voices, so that all the listeners use the measurement tools in the same way, and ratings across different voices can be compared in a meaningful manner. Voice quality is thus assumed to be constant across listeners, so that it can be dealt with as an attribute of the voice signal itself rather than a listener's perception product [8: 73–74]. In most cases, valid and reliable judgments of voice quality require trained judges, especially when it comes to the auditory-perceptual assessment of voice disorders [6].

Measuring acoustic parameters of voice quality is of great interest to scientists dealing with various voice pathologies. Their findings enable clinicians to diagnose voice disorders and are used in voice re-education aiming at acquiring appropriate phonation habits in patients suffering from vocal disorders [3, 6].

Acoustic analyses are used to provide measurements and quantification of various voice parameters, examining voice quality and phonation types in an objective way. The most common acoustic measures reflecting variability in the voice signal are *jitter, shimmer* and *HNR* (harmonics-to-noise ratio). These parameters are commonly used in clinical practice when evaluating voice disorders and voice quality disruptions such as breathiness, roughness and hoarseness, because they are relatively low-cost and non-invasive [6].

Jitter corresponds to variations in frequency between successive vibratory cycles [9, 10]. Jitter measurements can be conducted in two different ways – by peak-picking or waveform matching [9]. The latter tries to identify the time distance at which two consecutive waveshapes look most similar, while the peak-picking technique strives to find time locations where waveform amplitude is at its maximum. It is frequently a lack of precise control of vocal fold vibration that mainly affects jitter; patients with

Přijato 8. října 2018, akceptováno 21. prosince 2018.

voice pathologies often have a higher percentage of jitter. A typical percentage range indicating frequency variation from cycle to cycle for sustained phonation in young healthy adults stated by most researchers is 0.5–1.0 % [10]. Values above 1.04 % are considered pathological [9, 10].

Shimmer provides measurements of variations in amplitude between successive vibratory cycles. The methods used to measure shimmer are identical to jitter, but while jitter takes into account the duration of periods, shimmer considers the peak amplitude of the signal [10]. The amplitude variation of the sound wave is expressed in percentage or decibels. The value 3.81 % is stated as limit for detecting pathological voices [10].

HNR enables researchers to quantify the ratio between periodic and aperiodic components in the signal. HNR estimation can be carried out in two ways: on a time-domain basis (using autocorrelation) and on a frequency-domain basis. In the former case, HNR is computed directly from the acoustic signal, while in the latter case, HNR measurements are conducted from a transformed representation of a waveform [11]. The higher an HNR value is, the more sonorant and harmonic a voice is. HNR values below 7 dB are considered pathological [10, 12].

In time-domain analyses, jitter and shimmer estimations rely on the identification of cycles of vocal fold vibration in speech signals (so-called *pitch marks*), which might have some limitations. For instance, in case of severely dysphonic or aperiodic vowel samples, the degree of disturbance or perturbation may be so high that an accurate location of cycle boundaries is difficult and, in turn, fundamental frequency (f_0) detection is impossible. Another potential problem may arise when using continuous speech samples containing variations in pitch and loudness as well as rapid consonant–vowel and vowel–consonant transitions [6, 13]; as mentioned above, jitter and shimmer are typically measured in sustained vowels.

Cepstral-based techniques represent an alternative approach towards extracting f_0 and towards estimating the relative amplitude of harmonic versus noise components; importantly, these techniques eliminate the need for identifying cycle boundaries [6]. Cepstrum, a Fourier transform of the power spectrum of the speech signal, is a spectral-based method comprising prominent peaks – rahmonics (anagram of harmonics). A cepstrum of an acoustic signal displaying a well-defined harmonic structure shows a prominent peak; this *cepstral peak prominence* (CPP) is a measure of the amplitude of that cepstral peak which corresponds to f_0, normalized for overall signal amplitude. The amplitude of CPP thus reflects both harmonic organization and the overall amplitude of the signal [14]. It has been used by a number of investigators to evaluate voice quality, as it provides valid and reliable measurements not only in sustained vowel samples, but also in continuous speech [6, 13, 15].

Apart from *jitter, shimmer, HNR* and *CPP*, *harmonic amplitude measures* are commonly used when examining glottal characteristics, representing short-term acoustic manifestations of voice quality. These parameters are sensitive to varying degrees of vocal fold adduction in normal speakers. Based on theoretical models, they are related to the existence and size of glottal chink [16]. Differences in amplitudes of the first and second harmonics (H1−H2) and the harmonic amplitudes located closest to the first, second and third formant frequencies (H1−A1, H1−A2, H1−A3) of the voice spectrum have been found useful when quantifying the degree of glottal adduction in different voices [16, 17]. The amplitude of the first harmonic relative to that of the second (H1−H2) is used as an indication of the open quotient, i.e., the proportion of a glottal cycle in which the glottis is open. As the OQ relates to the overall glottal stricture, the H1−H2 measure is used to characterize the differences along the glottal constriction continuum [16, 17, 18] The amplitude of the second harmonic relative to the fourth (H2−H4) has also been found to be an important acoustic measure for distinguishing modal from nonmodal phonation [19], especially in cases when H1−H2 does not seem to work [18].

The amplitude of H1 relative to a higher frequency component can quantify the strength of higher frequencies in the spectrum relating to the closing velocity of the vocal folds, and perhaps to muscle tension. Thus, H1−A1, H1−A2 and H1−A3 are measured. These parameters can also distinguish modal and breathy phonation in some languages [18, 20] where H1−H2 does not seem to be useful. The amplitude of the first harmonic relative to that of the first formant prominence in the spectral domain (A1) reflects the bandwidth of F1, and may also be affected by source spectral tilt. H1−A1 is an indication of the presence of a posterior glottal chink, i.e., the degree to which the glottis fails to close completely during the closing phase [16, 17]. The amplitude of the first harmonic relative to that of the strongest harmonic in the second formant (H1−A2) is used as an indicator of the source spectral tilt (i.e., energy decrease with increasing frequency) at the mid formant frequencies [16]. Finally, H1−A3 reflects the spectral tilt at the higher formant frequencies, [16, 17, 21].

Harmonic amplitude measures can be compared across different speakers and vowels only if the measures are corrected for the effect of F1, F2 and F3 vocal tract resonances (frequencies and bandwidths) on harmonic amplitudes; uncorrected values reflect both the voice source and the supra-glottal filter. The corrected harmonic amplitude values are denoted with an asterisk, e.g. H1*−H2*, H1*−A1* etc. [17, 18, 22].

The objective of this study is to provide the aforementioned voice measure estimation in young healthy Czech male speakers of common Czech. A number of studies dealing with voice parameters published so far concentrate mainly on speakers with voice disorders or on patients with neurodegenerative diseases whose impact on voice quality has been scientifically proved [23]. This study seeks to establish quantitative ranges against which it would be possible to gauge the production of non-pathological voice.

A sample of fifty male speakers will be used to map acoustic parameter value ranges of voice-source characteristics based on a read speech task.

2. Method

2.1. Material

Recordings of fifty male speakers aged between 19 and 43 years (mean age: 24.7 years, SD: 6.1 years) were selected from the Database of Common Czech, a reference database for forensic purposes [24]. The speakers, who reported no voice or hearing problems, were recorded while reading a phonetically rich text of 150 words including all the Czech phonemes and their context-dependent variants in their natural voice; the length of the recording was approximately 60 seconds. Based on reported findings ([25: ch. 4] for a review), no age-related vocal changes were assumed in the speakers.

The recordings were acquired in a quiet environment using a portable recorder Edirol R09 and its in-built microphone, at a sampling rate of 48 kHz.

2.2. Parameter extraction and analyses

For each speaker, we extracted the voice quality parameters from 30 manually segmented /a a:/ vowels (16 phonologically short and 14 long vowels). Only phrase-internal vowels were chosen for analysis, so as not to confound the measurements by phrase-final phenomena such as creak or breathiness. However, vowels in all segmental contexts (incl. nasal) were included. Boundaries of the target vowels were determined based on the phonetically motivated recommendations for manual segmentation of the speech signal [26]. Briefly, the boundaries were located at the onset or the offset of full vowel formant structure. In case of the transition phase, the boundaries were placed in the temporal midpoint of this area. The total number of 1,500 target vowel sounds (30 vowels × 50 speakers) had to be reduced to 1,492, as the visual and auditory inspection revealed that 8 target items were of different vowel quality, due to an error in the speakers' reading.

Jitter, shimmer and HNR measurements were extracted using a Praat script [27] with the default settings for each parameter. As for jitter, values of local jitter (the most common measurement) were extracted using waveform matching (see section 1). The measure represents the average absolute difference between consecutive periods divided by the average period, and is expressed as a percentage [9, 10]. Shimmer measurements were performed using local shimmer parameter expressing the average absolute difference between the amplitudes of consecutive periods divided by the average amplitude. Similarly to local jitter, it is expressed as a percentage. HNR extraction, representing the degree of acoustic periodicity expressed in dB, was conducted by means of the cross-correlation method, as recommended for voice analysis in Praat [27].

The spectral magnitudes of H1*−H2*, H2*−H4*, H1*−A1*, H1*−A2* and H1*−A3* as well as CPP values were automatically extracted using Voice Sauce, a free stand-alone software [28], using the labelled Praat TextGrids. In order to estimate the location of harmonics, f_0 measurements needed to be carried out. We used the Voice Sauce default algorithm STRAIGHT [29] detecting f_0 at 1ms intervals and computing the harmonic magnitudes pitch-synchronously over a three-cycle window. This method eliminates much of the variability obtained in spectra computed over a fixed time window, and is equivalent to using a very long FFT window, providing more accurate measurements without relying on large FFT calculations [22].

CPP calculations in Voice Sauce are based on the algorithm [14] using a variable window length which is equal to five pitch periods by default. The obtained data are then multiplied with a Hamming window and transformed into the real cepstral domain. The CPP is estimated by conducting a maximum search around the quefrency of the pitch period. The peak is normalized to the linear regression line calculated between 1 ms and the maximum quefrency [22, 30].

The raw voice parameters data were processed in R [31] and visualised using the package *ggplot2* [32]. The statistical (mean, standard deviation, as well as the median in the final summarizing table) are computed for all analyzed vowels.

3. Results and discussion

The estimated values of the respective voice parameters will be presented in the following subsections. A table summarising the extracted mean values is presented in section 4 (Table 1). In section 3.5, we will focus on the relationship among the acoustic measures, and finally, we will comment on some speakers' results.

3.1. F_0 perturbation measures: jitter and shimmer

Fig. 1a shows the value ranges of the jitter measure for each speaker. The mean value is 1.83 % (SD: 1.97 %; 95% confidence interval: 1.73–1.94 %) and is above the threshold value of 1.04 % for pathological voices [10]. The mean value for the shimmer measure is 13.02 % (SD: 6.75 %; 95% confidence interval: 12.66–13.38 %) and is also higher than the pathological threshold of 3.81 % [10]. The shimmer value ranges for individual speakers are displayed in Figure 1b.

Both the estimated jitter and shimmer values are above the stated limits for detecting voice pathologies. However, as already mentioned above, the stated threshold values refer to the measurements performed on sustained vowels [9, 10], while our voice parameter extraction is based on continuous speech, which causes fast changes in pitch and formants [6, 13, 33]. The jitter and shimmer measures are thus necessarily higher than the pathological threshold,

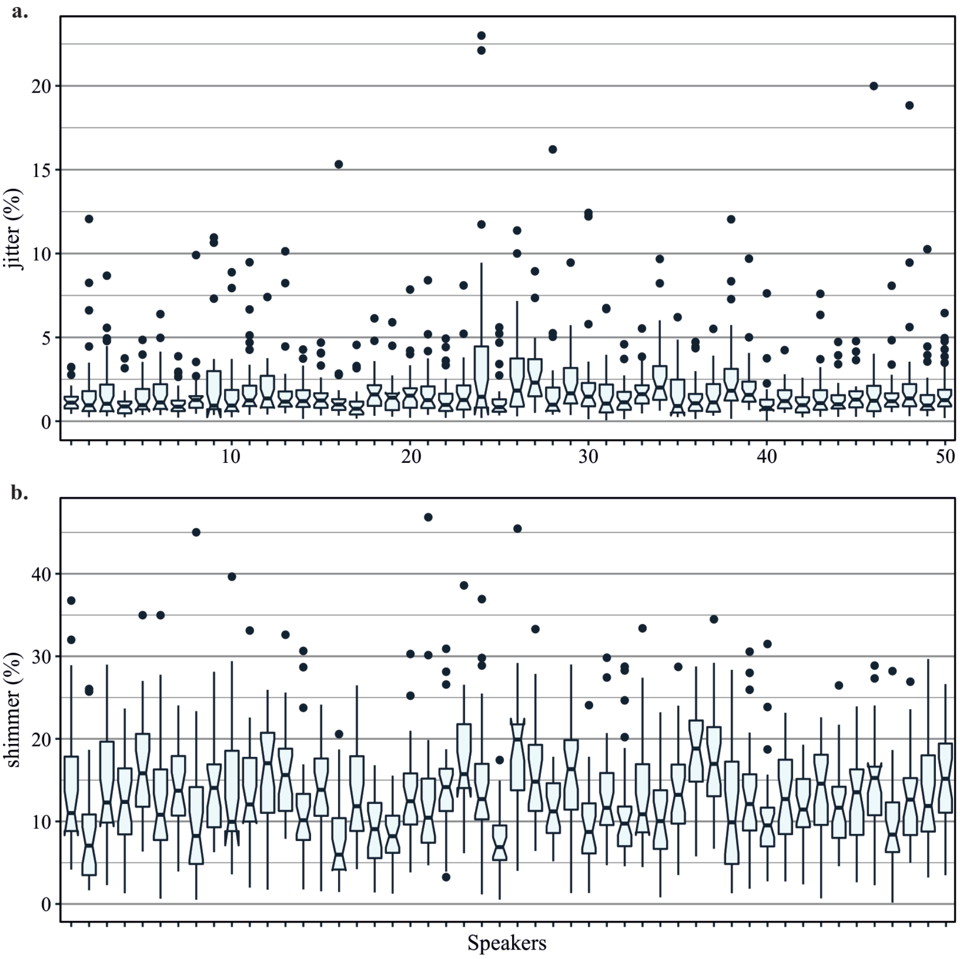


Figure 1: **a.** jitter value ranges, x-axis displays 50 speakers, y-axis shows jitter values (%). **b.** shimmer value ranges, x-axis displays 50 speakers, y-axis shows shimmer values (%)

and it is clear that they do not reflect any voice pathology (Boersma, 2017, personal communication); in fact, proposing perturbation values and ranges corresponding to healthy voices in connected speech is the main objective of this study.

3.2. Harmonicity (Harmonics-to-noise ratio, HNR)

Figure 2a shows the value ranges for each speaker. The mean value is 9.41 dB (SD: 4.05 dB; 95% conf. int.: 9.20–9.62 dB), which is well above the threshold value of 7 dB for voice pathologies and is in line with previous findings [10, 12, 34]. It will be useful to compare our data with previous studies. For example, Yumoto and Gould [34] examined the HNR parameter in relation to the degree of hoarseness in both healthy speakers and speakers with laryngeal disorders pre- and post-operatively using a sustained /ɑ:/ vowel. The estimated HNR for the healthy speaker group ranged between 7 and 17 dB with the mean of 11.9 dB (12.2 dB for males and 11.5 dB for women) compared to the estimated value range between −15.2 and 9.6 dB with the mean of 1.6 dB in preoperative speakers. It can be seen in Figure 2a that only two speakers' values in our sample fall below 7 dB.

3.3. Cepstral Peak Prominence (CPP)

Figure 2b displays the value ranges for the CPP measure extracted automatically using Voice Sauce. The mean value is 20.28 dB (SD: 3.69; 95% conf. int.: 20.26–20.30 dB). There exists a negative correlation between the CPP and the levels of aperiodicity of the glottal source – the higher the CPP, the lower the degree of aperiodicity in the voice signal [13, 15, 18]. As an acoustic measure of voice quality, some researchers evaluated the effectiveness of CPP in predicting breathiness ratings, and our results will thus be compared with theirs. Hillenbrand et al. [14] tested the parameter in healthy native English speakers who were asked to produce sustained vowels in nonbreathy, moderately breathy and very breathy phonation. The results confirmed that periodicity measures, namely CPP, provide the most accurate predictions of perceived breathiness [15, 18]. These findings were also confirmed for dysphonic voices and continuous speech [15]. In their study, Hillenbrand and Houde [15] provide exam-

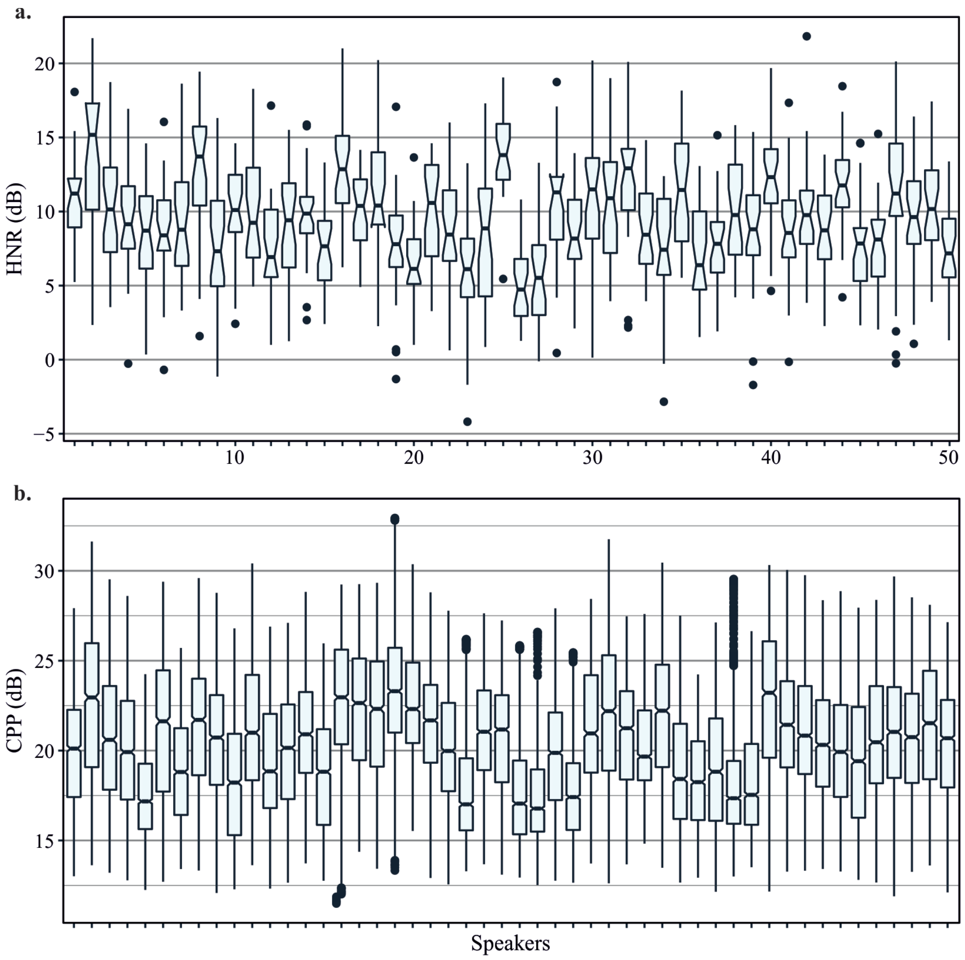


Figure 2: **a.** HNR value ranges, x-axis displays 50 speakers, y-axis presents HNR values (dB). **b.** CPP value ranges, x-axis displays 50 speakers, y-axis presents CPP values (dB)

ples of the CPP measures for signals perceived as nonbreathy and moderately breathy: 21.6 dB and 13.1. dB, respectively. Garellek and Keating [36] reported the CPP mean value of 22.5 dB for modal phonation extracted from /a, æ, o/ uttered in words by male speakers. The mean values for both creaky and breathy phonations were lower than the value of 20 dB. The CPP mean value we obtained should therefore reflect a nonbreathy/modal phonation.

3.4. Harmonic amplitude measures

The value ranges for H1*−H2*, as automatically extracted in Voice Sauce, are captured in Figure 3a. The mean is 1.83 dB (SD: 6.04; 95% conf. interval: 1.79–1.86 dB). As a correlate of the Open Quotient, lower values indicate a greater glottal constriction [18]. Cross-linguistically, H1*−H2* also represents one of the most successful measure of phonation type [35] and is often cited as an acoustic correlate of breathiness (e.g. [21]). Nevertheless, it seems to be a more reliable predictor of breathiness ratings for sustained vowels than for sentences or continuous speech [15]. H1*−H2* values for nonbreathy and breathy phonation were reported in [15]: 1.7 dB and 19.3 dB, respectively. Hanson and Chuang [17] obtained the following mean values using sustained vowel production in healthy speakers: men: 0.0 (SD: 1.8) dB and women: 3.1 (2.0) dB. Narra et al. [16] also used sustained vowels for their measurements in healthy speakers and present the following mean values for H1*−H2* (sustained /a/): 7.18 (SD: 3.7) dB for male and 11.49 (2.73) dB for female speakers.

H2*−H4* parameter estimation yielded the mean of 9.37 dB (SD: 6.09; 95% conf. int.: 9.33–9.4 dB). Figure 3b displays the value ranges for all our speakers. Similarly to H1*−H2*, H2*−H4* is also mentioned as a significant acoustic correlate of the perception of the contrastive breathiness in some languages [19]. Garellek et al. [20] measured H2*−H4* and H1*−H2* of the samples of sustained /a/ which were inverse-filtered and copy-synthesized to find out how they correlate with the perceived breathiness. The obtained mean values in dB for H2*−H4* were 8.93 (SD: 3.74) for men and 11.57 (4.99) for women, and for and H1*−H2* 6.13 (4.11) for men and 8.93 (4.55) for women, respectively.

Finally, let us look at the value estimations of the amplitude of the first harmonic relative to that of the

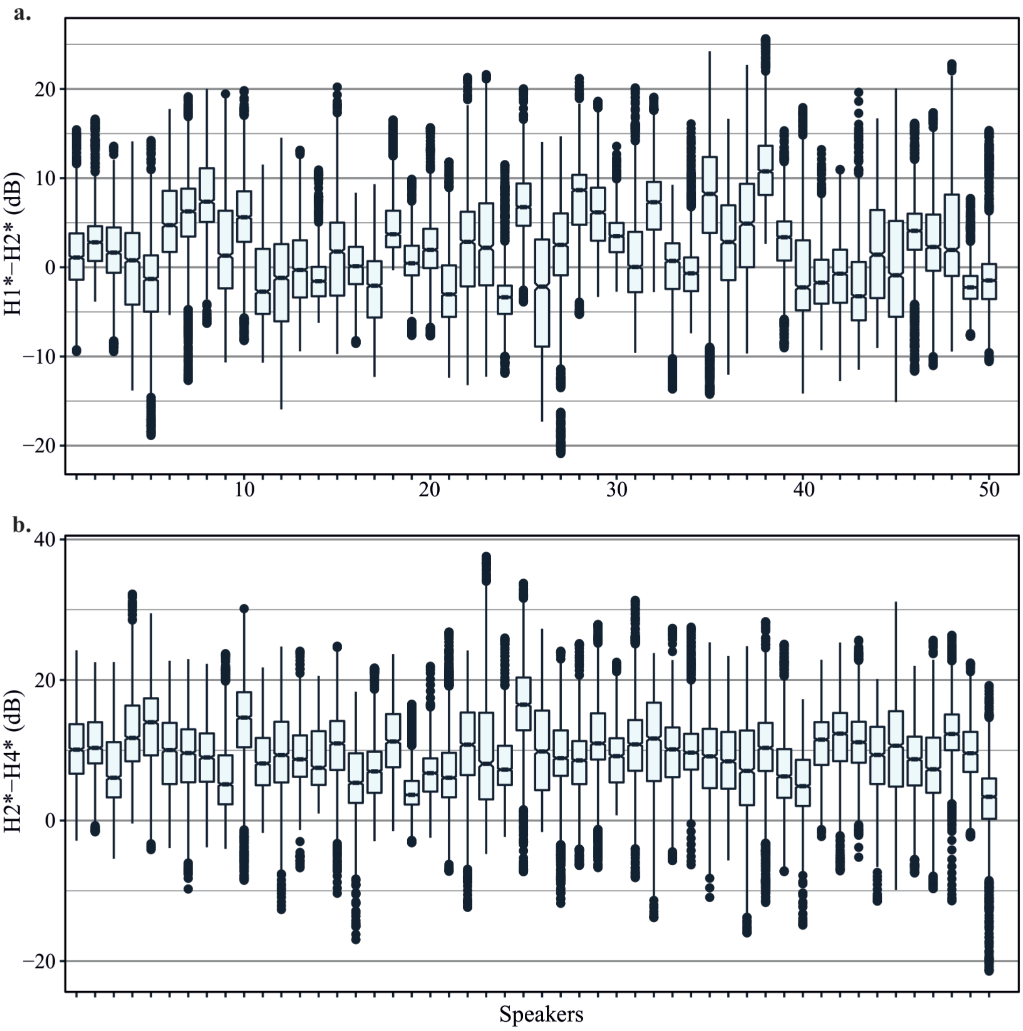


Figure 3: **a.** H1*−H2* value ranges, x-axis displays 50 speakers, y-axis presents H1*−H2* values (dB). **b.** H2*−H4* value ranges, x-axis displays 50 speakers, y-axis presents H2*−H4* values (dB)

F1, F2 and F3 prominence. Greater differences between H1*−A1*, H1*−A2* and H1*−A3* indicate less strong higher frequencies and more noise components in the spectrum [35].The mean values are: 21.43 dB (SD: 8.4) for H1*−A1*, 24.89 dB (SD: 8.42) for H1*−A2* and 18.87 dB (SD: 10.4) for H1*−A3 (see Table 1 in section 4). In [16], the following average and standard deviation values for sustained /a/ are reported: H1*−A1* in healthy men: 6.7 (2.53) dB and women 11.17 (4.54) dB, H1*−A2* 9.64 (4.79) dB in men and 12.73 (3.0) dB in women, and H1*−A3* 24.53 (6.06) dB in men and 28.79 (5.41) dB in women.

3.5. Acoustic measure relationships

Let us now have a look at the relationships among the extracted parameters. Figure 4 captures the correlations between the extracted mean values. In each case, we plotted a particular acoustic measure against CPP, as this parameter has been found to provide valid and reliable measurements in continuous speech [6, 13, 14, 15]. Spearman's rank correlation coefficient ρ was computed due to the presence of outlier values.

The plots suggest only mild or weak correlations, which confirms the relative independence of the different measures. Specifically, there is a positive correlation between CPP and HNR ($\rho = 0.4$, $p < 0.005$), and CPP and some of the harmonic amplitude measures: H1*−H2* ($\rho = -0.26$, $p < 0.1$), H2*−H4* ($\rho = -0.27$, $p < 0.1$). The negative correlation between CPP and the jitter did not even reach significance ($\rho = -0.14$, $p > 0.1$).

Correlations were stronger when we examined the interdependence of the harmonic amplitude measures. They are all positive and significant correlations: H1*−H2* *vs.* H1*−A1* ($\rho = 0.78$, $p < 0.001$); H1*−H2* *vs.* H1*−A2* ($\rho = 0.58$, $p < 0.001$), and H1*−H2* *vs.* H1*−A3* ($\rho = 0.61$, $p < 0.005$). Only the correlation between H1*−H2* and H2*−H4* was not significant ($\rho = 0.66$, $p > 0.5$), which indicates that they reflect different properties of the voice.

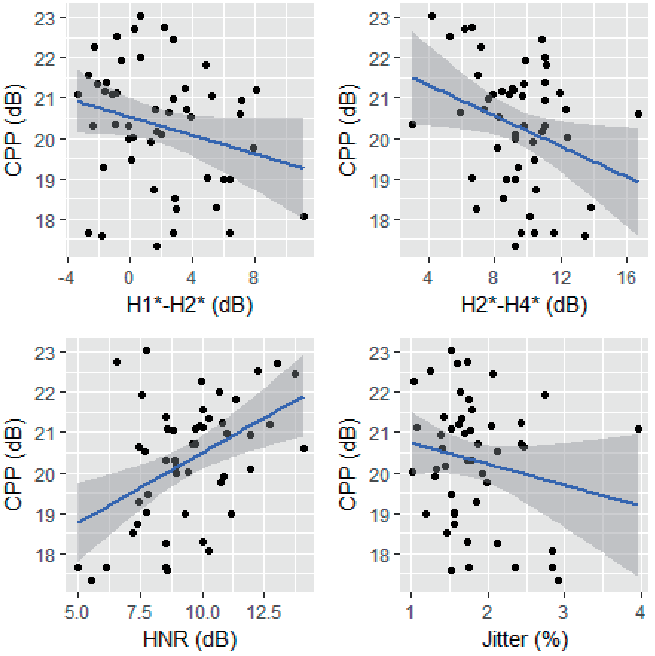


Figure 4: Scatterplots (extracted mean values) with trendlines (and 95% confidence bands). From top left: H1*−H2*, H2*−H4*, HNR and jitter plotted against CPP

3.6. Comments on particular speakers' values

Taking into account the relationships among our acoustic measures presented in the previous subsection, we will examine some speakers' mean values, taking values of the cepstral peak prominence close to the extremes as starting points.

The second highest CPP mean value was measured in Speaker 40 (S40): 22.52 dB (the overall mean across all speakers being 20.28 dB, and the mean value being higher for only one speaker, 23.06 dB). S40's HNR mean value is 12.24 dB (the overall mean: 9.41; the maximum mean value: 14.06), the H1*−H2* mean value of −0.88 dB is well below the overall mean of 1.83 (the minimum mean value: −3.43), and so is the H2*−H4* mean: 5.25 dB (the overall mean: 9.37; the minimum mean value: 3.01), and finally, S40's jitter mean value of 1.25 % is also below the overall mean of 1.83 % (the minimum value: 1.02).

The results reported in the previous paragraph imply a certain consistency across all the parameters. However, that is not always the case in all the speakers. For instance, in S2, we estimated the highest CPP mean value (23.06 dB), but S2's H1*−H2* and H2*−H4* means (2.78 dB and 10.9, respectively) are above the overall mean values (1.83 and 9.37, respectively).

Let us turn to Speaker 27 from the other end of the scale. S27 has the lowest mean value of CPP (17.32 dB) and his HNR mean of 5.57 dB is well below the overall mean of 9.41 dB. Also, this speaker's jitter mean of 2.9 % is above the overall mean. However, S27's H1*−H2* and H2*−H4* means (1.77 and 9.21, respectively) are below the overall mean values, which should not be expected considering the indicated relationships among the respective acoustic measures.

4. Conclusions

The aim of this study was to establish quantitative ranges of voice quality parameters in healthy Czech male speakers of common Czech in an objective way, based on a continuous speech reading task. The key values of all the parameters are summarized in Table 1.

Parameter	Mean (SD)	Median	Q1–Q3
Jitter	1.83 % (1.97)	1.18 %	0.72–2.12 %
Shimmer	13.02 % (6.75)	11.9 %	8.33–16.81%
HNR	9.4 dB (4.05)	9.4 dB	6.58–12.22 dB
CPP	20.3 dB (3.69)	20.2 dB	17.33–23.02 dB
H1*−H2*	1.8 dB (6.04)	1.6 dB	2.36–5.75 dB
H2*−H4*	9.4 dB (6.09)	9.2 dB	5.17–37.59 dB
H1*−A1*	21.4 dB (8.4)	20.9 dB	15.6–26.6 dB
H1*−A2*	24.9 dB (8.42)	24.3 dB	19.13–30.13 dB
H1*−A3*	18.9 dB (10.4)	18.9 dB	11.84–68.72 dB

Table 1: The estimated mean and median values and the values of the first and third quartile (Q1–Q3)

Although sustained vowel productions are commonly used to assess voice quality when conducting acoustic measurements, we decided to use a continuous speech sample based on a reading task. As human voice represents a dynamic time-varying source of vocal tract excitation, it is connected speech (characterized by rapid successions of different articulatory controls) that should provide relevant, ecologically valid data in terms of what makes speech production normal, and should enable researchers and clinicians to understand and assess the abnormality of speech production in different speech styles.

Our estimated jitter and shimmer values are above the commonly stated threshold limits for voice pathologies, especially in the case of shimmer. Needless to say, continuous speech contains variations in pitch, formants and loudness as well as rapid consonant-vowel and vowel-consonant transitions; our data thus cannot be compared with those obtained from speakers sustaining vowels for several seconds, but may provide reference for similar endeavours in the future.

The HNR measurements were conducted in a similar way as jitter and shimmer estimation, i.e. using a temporal- based method. Although the obtained mean value is quite above the stated threshold value for pathological voices, considering we used continuous speech. It would be useful to compare our data with HNR estimation using a spectral- (or more precisely, cepstral-) based technique.

Harmonic amplitudes measuring yielded somewhat higher values in most parameters compared to other studies. As in the case of the acoustic parameters mentioned above, harmonic amplitude measurements are commonly performed on sustained vowels. Finally, based on findings available in literature, the estimated CPP values seem to reflect modal phonation in most of our speakers.

While mapping voice parameters in our study, we also tried to examine the suitability/usefulness of the parameter estimations when using connected speech material. Future research might further examine the parameter extraction techniques relating to connected speech and conduct further measurements across different groups of speakers.

Acknowledgements

This research was supported by the project “Interdisciplinary approach to the linguistic theory issues”, subproject “The mapping of voice parameters in young healthy male speakers of Common Czech” solved at Charles University from the Specific university research in 2017. The second author was supported from European Regional Development Fund-Project “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

[1] Laver, J.: *The Phonetic Description of Voice Quality*, CUP, Cambridge, 1980.

[2] Arnold, A.: Le rôle de la fréquence fondamentale et des fréquences de résonance dans la perception du genre. *Travaux interdisciplinaires sur la parole et la langue,* 28, p. 2–14, 2012.

[3] Mendoza, E., Valecia, N., Muňoz, J., Truillo, H.: Differences in Voice Quality Between Men and Women: Use of the Long-Term Average Spectrum (LTAS), *Journal of Voice*, 10(1), p. 59–65, 1996.

[4] Weingartová, L., Bořil, T., Vaňková, J.: Spektrální sklon. In: Skarnitzl, R. (Ed.), *Fonetická identifikace mluvčího*, FF UK, Praha, 2014.

[5] Bhuta, T, Patrick, L., Garnett, J. D.: Perceptual evaluation of voice quality and its correlation with acoustic measurements, *J. of Voice*, 18, p. 299–304, 2004.

[6] Awan, S. N., Solomon, N. P., Helou, L. B., Stojadinovic, A.: Spectral-Cepstral Estimation of Dysphonia Severity: External Validation, *Annals of Otology, Rhinology & Laryngology*, 122(1), p. 40–48, 2013.

[7] Kreiman, J., Gerratt, B. R.: Jitter, Shimmer, and Noise in Pathological Voice Quality Perception, *VOQIAL'03*, Geneva, p. 57–61, 2003.

[8] Kent, R. D., Ball, M. J.: *Voice Quality Measurement*, Singular Publishing Group, San Diego, 2000.

[9] Boersma, P.: Should Jitter Be Measured by Peak Picking or by Waveform Matching?, *Folia Phoniatr. Logop.*, 61, p. 305–308, 2009.

[10] Teixeira, J. P., Oliveira, C., Lopes, C.: Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters, *Procedia Technology,* 9, p. 1112–1122, 2013.

[11] Qi, Y., Hillman, R. E.: Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals, *Journal of Acoustical Society of America*, 102(1), p. 537–543, 1997.

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

[13] Murphy, P. J.: Periodicity estimation in synthesized phonation signals using cepstral rahmonic peaks, *Speech Communication,* 48, p. 1704–1713, 2006.

[14] Hillenbrand, J., Cleveland, R., Erickson, R.: Acoustic correlates of breathy vocal quality, *J. Sp. Hear. Res.*, 37, p. 769–778, 1994.

[15] Hillenbrand, J., Houde, R. A..: Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech, *Journal of Speech and Hearing Research*, 39, p. 311–321, 1996.

[16] Narra, M., Anu, T. D., Varghese, S. M., Dattatreya, T.: Harmonic Amplitude Measures to Note Gender Differences, *Advances in Life and Technology*, 31, p. 17–23, 2015.

[17] Hanson, H. M., Chuang, E. S.: Glottal characteristics of male speakers: Acoustic correlates and comparison with female data, *J. Acoust. Soc. Am.,* 106(2), p. 1064–1077, 1999.

[18] Keating, P. A., Esposito, C.: Linguistic Voice Quality, *UCLA Working Papers in Phonetics*, 105, p. 85–91, 2007.

[19] Garellek, M., Keating, P., Esposito, C. M., Kreiman, J.: Voice quality and tone identification in White Hmong, *J. Aoucst. Soc. Am.*, 133(2), p. 1078–1089, 2013.

[20] Garellek, M., Samlan, R. A., Kreiman, J., Gerratt, B.: Perceptual sensitivity to a model of the source spectrum, *Proceedings of Meetings on Acoustics*, 19, p. 1–5, 2013.

[21] Wayland, R., Jongman, A.: Acoustic correlates of breathy and clear vowels: the case of Khmer, *Journal of Phonetics*, 31, p. 181–201, 2003.

[22] Shue, Y., Keating, P., Vicenik, C., Yu, K.: VoiceSauce: A program for voice analysis, *Proc 17th ICPhS*, Hong Kong, p. 1846–1849, 2011.

[23] Tykalová, T., Rusz, J., Čmejla, R., Růžičková, H., Růžička, E.: Acoustic investigation of stress patterns in Parkinson's disease, *Journal of Voice*, 28(1), 129.e1–129.e8, 2014.

[24] Skarnitzl, R., Vaňková, J.: Fundamental frequency statistics for male speakers of Common Czech, *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, p. 7–17, 2017.

[25] Kreiman, J., Sidtis, D.: *Foundations of Voice Studies*, Blackwell Publishing, Oxford, 2011.

[26] Machač, P., Skarnitzl, R.: *Principles of Phonetic Segmentation*, Epocha, Praha, 2009.

[27] Boersma, P., Weenink, D.: *Praat: doing phonetics by computer* (Version 5.4.08), Retrieved: 5. 5. 2015, `http://www.praat.org`.

[28] Shue, Y.: *VoiceSauce: A program for voice analysis* (Version 1.31), Retrieved: 31. 5. 2017, `http://www.phonetics.ucla.edu/voicesauce/`.

[29] Kawahara, H., Masuda-Katsuse, I., de Chevigne, A.: Restructuring speech representation using a pitch adaptive time-frequency smoothing and an instantaneous frequency based F0 extraction, *Speech Communication*, 27, p. 187–207, 1999.

[30] Vicenik, C., Lin, S., Keating,P., Shue, Y.: Online documentation for VoiceSauce. Available at: `http://www.phonetics.ucla.edu/voicesauce/documentation/index.html`, accessed: 31. 5. 2017.

[31] R Core Team: *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. Retrieved: `http://www.R-project.org`., 2016.

[32] Wickham, H.: *ggplot2: Elegant graphics for data analysis (use R!)*, Springer, New York, 2009.

[33] Fourcin, A.: Aspects of Voice Irregularity Measurements in Connected Speech, *Folia Phoniatrica et Logopaedica*, p. 126–136, 2009.

[34] Yumoto, E., Gould, W.J.: Harmonics-to-noise ratio as an index of the degree of hoarsness*, J. Acoust. Soc. Am.,* 71(6), p. 1544–1550, 1982.

[35] Esposito, C. M.: The effects of linguistic experience on the perception of phonation, *J. of Phonetics*, 38, p. 306–316, 2010.

[36] Garellek, M., Keating, P.: The acoustic consequences of phonation and tone interactions in Jalapa Mazatec, *J. of IPA,* 41(2), 2011.

Akustické listy: ročník 25, číslo 1–2 červen 2019 ISSN: 1212-4702
Vydavatel: Česká akustická společnost, z. s., Technická 2, 166 27 Praha 6
Počet stran: 20 Počet výtisků: 200
Redakční rada: M. Brothánek, O. Jiříček, R. Čmejla, J. Volín
Jazyková úprava: R. Svobodová, M. Tharp
Uzávěrka příštího čísla Akustických listů je 30. srpna 2019.