Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality

Jitka Vaňková, Radek Skarnitzl

Institute of Phonetics, Faculty of Arts, Charles University in Prague, Czech Republic

jitka.vanka@gmail.com, radek.skarnitzl@ff.cuni.cz

Abstract

This study focuses on short-term acoustic correlates of voice quality. It assesses the within-speaker stability (across different speaking styles) and between-speaker variability of measurements which compare the amplitudes of various spectral events – H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2* and H1*-A3*. Although speakers do differ with regard to the compactness of the parameters in read and spontaneous speaking styles, the parameters H1*-H2*, H1*-A1* and H1*-A2* appear both considerably stable for one speaker in different speaking styles and efficient in between-speaker comparisons. Though not directly applicable in forensic settings, these glottal parameters outperformed vowel formants in classification using LDA.

Index Terms: voice quality, spectrum, speaking styles, Czech

1. Introduction

Voice quality has long been recognized as an independent and full-fledged prosodic dimension [1]. It is a multidimensional phenomenon, which has made it difficult to describe in other than negative terms (i.e., what it is not) [2]. Since Laver's seminal work [3], voice quality has been defined in two ways: narrowly, referring only to the vibration of the vocal folds and its perceptual impact; and broadly, referring also to the perceptual impact of the movements and longer-term settings of supraglottal organs. In this study, we are interested in the narrower sense of the term voice quality, in phonatory modifications.

Since the perceptual evaluation of voice quality is far from straightforward (see e.g. [4], [5]), various acoustic correlates of different aspects of voice quality have been proposed. In this regard, we may talk about both long- and short-term acoustic manifestations of voice quality. The long-term average spectrum (LTAS) shows the frequency distribution of the speech signal over a longer (typically at least 30 seconds) stretch of speech [6], [7], [8]. By averaging over a long portion of speech, spectral differences due to individual segments are evened out, and the method thus yields information pertaining to general voice quality. LTAS has been successfully applied in various phonetic and speech pathological tasks (see e.g. [8] for a summary), with various parameterizations of the LTAS being proposed. Most of these reflect spectral slope, or spectral tilt, in other words the energy decrease with increasing frequency (e.g., the alpha value [9], the Hammarberg index [10], or more recent attempts [11]). In addition, the prominence of a specific peak in the LTAS – called the singer's or speaker's formant – has been correlated with qualities like resonance or sonority of the voice [12], [13].

Short-term manifestations of voice quality may be extracted from individual speech sounds, typically vowels. Jitter and shimmer quantify the degree of fluctuations of the voice source, in the frequency and amplitude domain, respectively; however, it has been suggested that these do not constitute useful correlates of voice quality [14]. Another group of parameters concerns the degree of additive noise in voice: harmonicity, or harmonics-to-noise ratio (HNR) whose measurement has been proposed in various domains [15], [16], [17], the glottal-to-noise excitation ratio (GNE) [18], and others. The last group of parameters to be mentioned here reflects, similarly to the LTAS, spectral slope. In this case, however, it is the spectral slope of a single vowel that is quantified [19], by means of comparing the amplitudes of various events in the acoustic spectrum (see [20] or [21] for an extensive review). The parameter H1-H2 (the amplitude of the first harmonic relative to that of the second) has been correlated with the open quotient. H1-A1 (the amplitude of the first harmonic relative to that of the first formant) is regarded as an indication of F1 bandwidth (B1), which is in turn an indication of the degree to which the glottis fails to close completely during the closing phase. H1-A3 (the amplitude of H1 relative to that of F3) is a reflection of spectral slope. These measures on the speech spectrum are most frequent, although others have also been proposed, such as H2-H4 or H1-A2 (e.g. [22], [23], [24]).

In the current study, we are interested in the stability of voice quality parameters across different speaking styles within one speaker and, at the same time, in inter-speaker differences. We will focus on those parameters which rely on the comparison of the amplitudes of various harmonics (or spectral peaks) – i.e. H1-H2, H1-A1 etc. – one of the reasons for this choice being the fact that Hanson [20] does mention a considerable degree of speaker specificity of these parameters but does not explore this question further.

Theoretically, a low degree of within-speaker variability and a high degree of inter-speaker variability may be useful in phonetic speaker recognition [25], [26], [27]. However, it appears that direct applicability of even such a positive finding in most forensic phonetic casework is problematic, as illustrated by Nolan [28]. The greatest drawback consists in the fact that most recordings of unknown speakers are telephone speech in which (at least) the first two harmonics of male voices are lost. In addition, the widespread use of mobile phones leads to various kinds and levels of background noise. While it is also true that voice quality may differ significantly with speaking style and a mismatch in speaking style thus may lead to false eliminations [28], we believe that the within-speaker variability of parameters like H1-H2 or H1-A3 – which would belong to the long-term segmental strand in Nolan's model [29] – does merit further investigation.

10.21437/SpeechProsody.2014-205

In her study, Hanson [20] suggests that, in order to enable comparison of these measures across different speakers (and vowels), the amplitudes of the first and second harmonic, H1 and H2, need to be corrected for the boosting effects of the first formant (frequency and bandwidth), while F3 amplitude needs to be corrected for the boosting effects of the lower formants. The corrected values are then denoted with an asterisk, thus for instance H1*-H2*, H1*-A1, or H1*-A3*. It will be these corrected measures that will be applied in this study. Specifically, we are interested in the stability of these measures within speakers across speaking styles, as well as in differences across speakers. The performance of the target measures will be compared with that of mean formant values, which will serve as a sort of benchmark here.

2. Method

2.1. Material & subjects

The material for this study was taken from the VASST corpus, which focuses on the variability of speaking styles and which has been collected in various regions of the Czech Republic. Recordings were obtained in quiet rooms in people's homes via a professional portable recorder Edirol HR-09, with a 48-kHz sampling frequency (later down-sampled to 32 kHz). In the present study, we analyzed recordings of spontaneous and read speech produced by six adult male native speakers of Czech aged 28–65 (mean age = 40).

The spontaneous speech sample involved a semi-structured interview in which the speaker was encouraged to speak freely about selected topics. As for the read speech sample, the speakers were asked to read a coherent text in a natural way after sufficient preparation.

We analyzed the Czech open central monophthongs /a/ and /aː/ in various consonantal contexts – only vowels in the context of /ɦ/ were excluded, as the glottal fricative may introduce additional breathiness into the spectrum of the vowel. The boundaries of the target segments were manually adjusted following the suggestions of [30] in Praat [31]. Each token was marked for syllable status with respect to word stress and for its position in utterance (final or non-final). For each speaker and style, we analyzed 50 vowel items, yielding the total of 600 tokens (6 speakers × 2 styles × 50 items). 10 of those had to be removed from analyses since they were not assigned glottal parameter values (see below).

2.2. Parameter extraction and analyses

All parameter values (spectral magnitudes of H1, H2, H4, A1, A2 and A3, as well as the formant frequencies of F1–F4) were automatically extracted by VoiceSauce (VS) [32], [33], a free stand-alone software, using the labelled Praat TextGrids.

To locate and measure the harmonics, VS relies on the extraction of F0. The default algorithm for F0 extraction in VS is STRAIGHT [34], which was also used in our study. In traditional FFT analysis, changing the analysis window can change the features of the extracted spectrum. Here, amplitudes of the harmonics are computed pitch-synchronously (over a 3-cycle window), which eliminates much of the variability in spectra computed over a fixed time window. The method is equivalent to using a very long FFT window but enables considerably more accurate measurements without relying on large FFT calculations [32], [33]. As only male voices were examined, the settings were slightly adjusted: maximum F0 was lowered to 400 Hz and minimum F0 raised to 60 Hz. All other default settings have been preserved.

As for the formant frequencies (F1–F4), they were likewise automatically extracted by VS, using the default algorithm for formant detection, the Snack Sound Toolkit [35]. Snack is an algorithm based on LPC, which uses as defaults the covariance method, pre-emphasis of 0.96, window length of 25 ms, and frame shift of 1 ms, so as to match the F0 estimation by the STRAIGHT algorithm [36].

From the values extracted at 1-ms intervals, the mean value was computed from the middle third (33–67%) of each vowel. Subsequently, 10% of the lowest and 10% of the highest values of the four formants and F0 were manually checked for extraction errors and, if necessary, corrected by direct estimation from the spectrogram (in the case of formants) and the waveform (for F0). Extraction mistakes were not numerous, apart from one speaker whose vowels occasionally manifested diplophonia [37]. The corrected F0 values were then reloaded into VS and the glottal parameters of the corresponding items computed again.

The computation of glottal parameters in VS differs slightly from that mentioned in [20]. VS uses for the corrected measures an algorithm developed by Iseli et al. [38] where H1*-H2* is corrected for the boosting effect of not only F1 but also F2, and F1 through F3 are used for the computation of H1*-A3*. In addition, VS computes H1*-A1* (*cf.* H1*-A1 in [20]).

To investigate the within- and between-speaker variability of these parameters, we performed several analyses. First, the stability of the glottal parameters within a speaker and across the two speaking styles was examined by means of the Kolmogorov-Smirnov (K-S) test, which compares two distributions of values. Second, the K-S test was also applied, in pairwise comparisons, to examine between-speaker variability. Finally, the effectiveness of the glottal parameters to discriminate between speakers was compared to that of formant frequencies by means of Linear Discriminant Analysis (LDA).

3. Results

The main aim of this study was to assess the stability of short-term voice quality parameters across different speaking styles (read and spontaneous) within one speaker, as well as their between-speaker variability. For these purposes, the Kolmogorov-Smirnov (K-S) test was used as it is also sensitive to differences in the general shapes of the distributions (such as differences in dispersion and skewness) in the compared samples.

3.1. Within-speaker stability

Let us first have a look at how stable the parameters are within one speaker. Figure 1 displays for each parameter which of the speakers (labelled S1–S6) did not yield any significant differences across the two styles ($p > 0.05$; above the line,

denoted with a +) and which of the speakers did yield significant differences ($p < 0.05$; below the line, denoted with a –). As we can see, speakers differ with respect to parameter stability: while the parameter differences in the two speaking styles are always insignificant for S1 – i.e., the values are stable in the two styles – S4, on the other hand, yields significant differences in 4 out of the 5 parameters. The figure also suggests that the most stable parameter is H1*-A2* followed by H1*-H2* and H1*-A1*, while H2*-H4* and H1*-A3* appear the least successful in expressing within-speaker stability of our sample.

	H1*-H2*	H2*-H4*	H1*-A1*	H1*-A2*	H1*-A3*	
				S1		
	S1		S1	S2		
	S2	S1	S3	S3	S1	
	S3	S3	S5	S5	S2	
+	S6	S4	S6	S6	S5	+
–	S4	S2	S2	S4	S3	–
	S5	S5	S4		S4	
		S6			S6	

Figure 1. *Within-speaker stability of voice quality parameters in the two analyzed speaking styles: insignificant differences between the styles appear above the line (also denoted with a +), significant ones below the line (with a –).*

Example distributions are given in Figures 2 and 3; Figure 2 shows the values of parameter H1*-A1* (in dB) for speaker S5 whose distribution did not differ across the two speaking styles ($p > 0.05$), while Figure 3 shows the values of H1*-H2* for speaker S4 which differed significantly ($p < 0.05$).

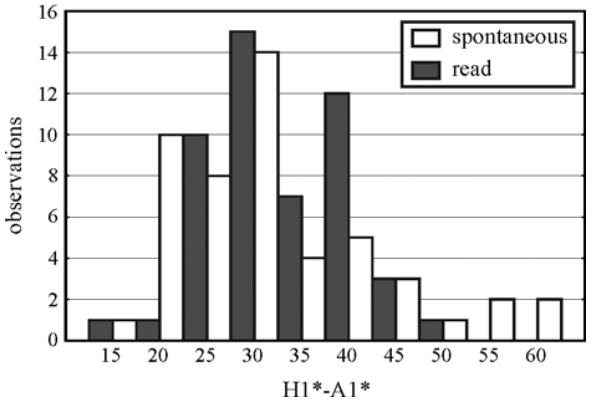

Figure 2. *The distribution of speaker S5's H1*-A1* (in dB) in read and spontaneous speech.*

3.2. Between-speaker variability

We were further interested to what extent these parameters can capture differences between speakers. To illustrate this, we present the results of pairwise comparisons between speakers for the most successful parameter in expressing between-speaker variability, H1*-H2*, in Table 1. The table shows that all 15 possible comparisons – having 6 speakers allows 15 pairwise comparisons – are statistically significant (denoted by an *), most of them highly significant (denoted by **). Moreover, speakers S1 and S5 show statistically highly significant differences from all other speakers, thus being clearly discriminated by their distribution of H1*-H2* values from the others.

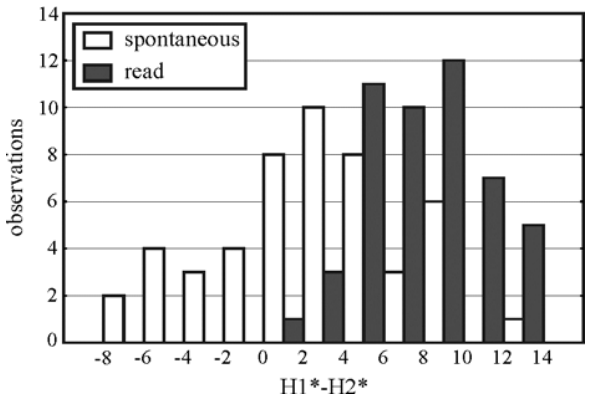

Figure 3. *The distribution of speaker S4's H1*-H2* (in dB) in read and spontaneous speech.*

	S1	S2	S3	S4	S5	S6
S1	X	**	**	**	**	**
S2		X	*	**	**	**
S3			X	**	**	**
S4				X	**	*
S5					X	**
S6						X

Table 1. *Between-speaker variability of H1*-H2* in pairwise comparisons (** $p < 0.001$; * $p < 0.05$).*

Not only H1*-H2* but also the other parameters appear to be efficient in expressing between-speaker differences. The overview of between-speaker pairwise comparisons for all 5 parameters is presented in Table 2. As already stated above, H1*-H2* is the most successful parameter in this respect, though it can be seen that also H1*-A1* yields statistically significant differences for all possible comparisons. H1*-A2* and H1*-A3* perform only slightly worse (one statistically insignificant comparison), while H2*-H4* turns out to reflect between-speaker variability the least, with 4 of the 15 comparisons being statistically insignificant.

	$p < 0.001$	$p < 0.05$	$p > 0.05$
H1*-H2*	13	2	0
H2*-H4*	8	3	4
H1*-A1*	12	3	0
H1*-A2*	14	0	1
H1*-A3*	14	0	1

Table 2. *Significance levels of between-speaker pairwise comparisons for all analyzed parameters.*

Figures 4 and 5 again provide example distributions. Figure 4 shows a similar distribution of H2*-H4* of speakers S1 and S3, while Figure 5 shows distinct distributions of H1*-H2* of speakers S1 and S6 (note that the depicted parameters did not differ in these two speakers across the two speaking styles; see Figure 1). Speaker S6's voice thus appears to be breathier, as suggested by the positive values of H1*-H2* [20].

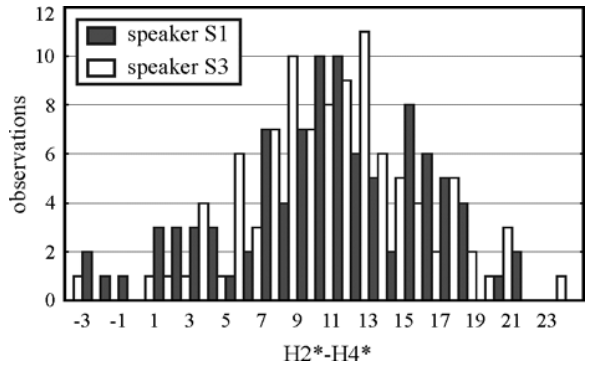

Figure 4. *The distribution of H2*-H4* (in dB) of speakers S1 and S3.*

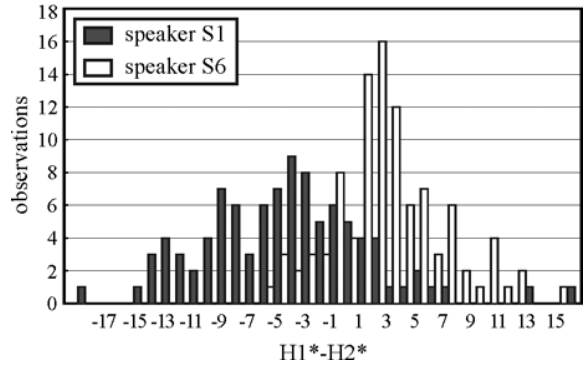

Figure 5. *The distribution of H1*-H2* (in dB) of speakers S1 and S6.*

3.3. Comparison with formant frequencies

Our final objective was to compare the effectiveness of the glottal parameters in discriminating between speakers with that of formants (F1–F4), which are often employed for these purposes and were thus used as a benchmark. This comparison was based on Linear Discriminant Analysis (LDA).

The classification rates for the two sets of parameters are presented in Table 3. The glottal parameters perform slightly better (52.6%) than formants (48.3%), both being well above the chance classification rate of 16.7%. Also, both models are statistically highly significant: F (25,2148) = 36.7; $p < 0.001$ for the glottal parameters, and F (20,1927) = 32.1; $p < 0.001$ for the formants. Table 3 also reveals considerable differences between the discriminability of individual speakers, as well as differences in the performance of the two models for the six speakers, which is most marked for speaker S1.

The values of Wilks' λ complement these tendencies: overall λ equals 0.267 for the glottal parameters and 0.385 for formants, which indicates that the variability in our data is better accounted for by the former.

Speaker	Glottal parameters	Formants
S1	62.5	12.5
S2	39.2	56.1
S3	24.2	39.4
S4	51.5	74.5
S5	84.0	72.0
S6	53.5	34.3
Total	**52.6**	**48.3**

Table 3. *Classification rates (in %) for the glottal parameters and formant values.*

The values of Wilks' λ for the individual parameters also support our findings that the most useful parameter for differentiating between speakers in our model is H1*-H2* as its removal would impair its efficiency the most, while H2*-H4* contributes to its efficiency the least. As for formants, F3 appears to be most and F1 least useful.

4. General discussion and Conclusions

This study analyzed short-term voice quality parameters from the viewpoint of their within- and between-speaker variability, as well as of their potential to discriminate between speakers.

As for within-speaker stability, our results suggested considerable differences between speakers with regard to their compactness in read and spontaneous speaking styles (*cf.* speakers S1 and S4 in Fig. 1). The results also indicate that H1*-A2* is the most stable parameter, followed by H1*-H2* and H1*-A1*. The same parameters also manifested high between-speaker variability (Table 2). This finding points to the importance of the relative amplitude of H1 for distinguishing voice quality (*cf.* [20], [23], [24]).

Between-speaker comparisons also revealed that some speakers are clearly differentiated from all others, specifically speaker S1 and S5 (see Table 1 and also Table 3). Interestingly, S1 was our youngest subject (28 years old), while S5 was our oldest one (65 years old). Our results are thus in accordance with previous studies [38] which showed an age dependency of H1*-H2* and H1*-A3*.

The comparison of the speaker-discriminating potential of the glottal parameters and formant values suggested that the glottal parameters slightly outperform formants overall, though individual differences may be observed. By way of conclusion, let us repeat, however, that the findings cannot be directly applicable in forensic casework due to the band-limited telephone signal, and that our main point of interest was the stability of the voice parameters across the two speaking styles; in this respect, we believe, our study indicated their usefulness for future research.

5. Acknowledgements

This research was supported by the Czech Science Foundation (GACR 406/12/0298) and the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation.

6. References

[1] Campbell, N. and Mokhtari, P., "Voice quality: the 4th prosodic dimension", Proc 15th ICPhS, Barcelona, 2417-2420, 2003.

[2] Kreiman, J. and Sidtis, D., "Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception", Oxford: Wiley-Blackwell, 2011.

[3] Laver, J., "The Phonetic Description of Voice Quality", Cambridge: Cambridge University Press, 1980.

[4] Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A. and Berke, G. S., "Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research", J Speech Hearing Res, 36: 21-40, 1993.

[5] Bele, I.V., "Reliability in perceptual analysis of voice quality", J Voice, 19: 555-573, 2005.

[6] Löfqvist, A., "The long-time-average spectrum as a tool in voice research", J Phon, 14: 471-475, 1986.

[7] Master, S., De Biase, N., Pedrosa, V. and Chiari, B. M., "The long-term average spectrum in research and in the clinical practice of speech therapists", Pró-Fono Revista de Atualização Científica, 18: 111-120, 2006.

[8] Leino, T., "Long-term average spectrum in screening of voice quality in speech: Untrained male university students", J Voice, 23: 671-676, 2009.

[9] Frøkjær-Jensen, B. and Prytz, S., "Registration of voice quality", Brüel Kjær Technological Review, 3: 3-17, 1976.

[10] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. and Wedin, L., "Perceptual and acoustic correlates of abnormal voice qualities", Acta Otolaryngologica, 90: 441-451, 1980.

[11] Volín, J., Weingartová, L. and Skarnitzl, R., "Spectral characteristics of schwa in Czech accented English", Research in Language, 11: 31-39, 2013.

[12] Bele, I. V., "The speaker's formant", J Voice, 20: 555-578, 2006.

[13] Leino, T., Laukkanen, A.-M. and Radolf, V., "Formation of the actor's/speaker's formant: A study applying spectrum analysis and computer modeling", J Voice, 25: 150–158, 2011.

[14] Kreiman, J. and Gerratt, B. R., "Jitter, shimmer, and noise in pathological voice quality perception", Proc VOQUAL'03, Geneva, 57-61, 2003.

[15] Yumoto, E., Gould, W. J. and Baer, T., "Harmonics-to-noise ratio as an index of the degree of hoarseness", J Acoust Soc Am, 71: 1544-1550, 1982.

[16] de Krom, G., "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", J Speech Hearing Res, 36: 254-266, 1993.

[17] Yu, A.-T. and Wang, H.-C., "New speech harmonic structure measure and its applications to speech processing", J Acoust Soc Am, 120: 2938-2949, 2006.

[18] Michaelis, D., Gramss, T. and Strube, H. W., "Glottal-to-noise excitation ratio – a new measure for describing pathological voices", Acta Acustica, 83: 700-706, 1997.

[19] Weingartová, L. and Volín, J., "Spectral measurements of vowels for speaker identification in Czech", Studie z aplikované lingvistiky, 1/2013: 21-36, 2013.

[20] Hanson, H. M., "Glottal characteristics of female speakers: Acoustic correlates", J Acoust Soc Am, 101: 466-481, 1997.

[21] Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y. and Slifka, J., "Towards models of phonation", J Phon, 29: 451-480, 2001.

[22] Kreiman, J., Gerratt, B. R. and Antoñanzas-Barroso, N., "Measures of the glottal source spectrum", J Speech Language Hearing Res, 50: 595-610, 2007.

[23] Keating, P. A. and Esposito, C., "Linguistic Voice Quality", UCLA Working Papers in Phonetics, 105: 85-91, 2007.

[24] Keating, P., Esposito, C., Garellek, M., Khan, S. ud D. and Kuang, J., "Phonation contrasts across languages", UCLA Working Papers in Phonetics, 108: 188-202, 2010.

[25] Nolan, F., "Speaker Recognition and Forensic Phonetics". In W. J. Hardcastle and J. Laver [Eds], Handbook of Phonetic Sciences. Oxford: Blackwell, 744-767, 1997.

[26] Hollien, H., "Forensic Voice Identification", San Diego: Academic Press, 2002.

[27] Jessen, M., "Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs", München: LINCOM, 2013.

[28] Nolan, F., "Forensic speaker identification and the phonetic description of voice quality". In W. Hardcastle and J. Beck [Eds], A Figure of Speech. Mahwah, New Jersey: Erlbaum, 385-411, 2005.

[29] Nolan, F., "The phonetic bases of speaker recognition", Cambridge: Cambridge University Press, 1983.

[30] Machač, P. and Skarnitzl, R., "Principles of Phonetic Segmentation", Praha: Epocha, 2009.

[31] Boersma, P. and Weenink, D., "Praat - Doing phonetics by computer" (Version 5.3.53.). Online: http://www.praat.org, accessed on 11 July, 2013.

[32] Shue, Y., Keating, P., Vicenik, C. and Yu, K., "VoiceSauce: A program for voice analysis", Proc 17th ICPhS, Hong Kong, 1846-1849, 2011.

[33] Shue, Y., "VoiceSauce: A program for voice analysis" (Version 1.14). Online: http://www.seas.ucla.edu/spapl/voicesauce/, last updated on May 30, 2013, accessed on 7 October, 2013.

[34] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A., "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based F0 extraction", Spe Com, 27: 187-207, 1999.

[35] Sjölander, K. (2004), Snack sound toolkit. KTH Stockholm, Sweden. Online: http://www.speech.kth.se/snack, accessed on 10 October, 2013.

[36] VoiceSauce Manual. Online: http://www.seas.ucla.edu/spapl/voicesauce/documentation/parameters.html#formants, accessed on 20 November, 2013.

[37] Cavalli, L. and Hirson, A., "Diplophonia Reappraised", J Voice, 13: 542-556, 1999

[38] Iseli, M., Shue, Y.-L. and Alwan, A., "Age, sex, and vowel dependencies of acoustic measures related to the voice source", J Acoust Soc Am, 121: 2283–2295, 2007.