THE EFFECT OF TARGETED VOICE MANIPULATIONS ON LONG-TERM ACOUSTIC CHARACTERISTICS

Alžběta Houzar, Tomáš Nechanský & Radek Skarnitzl

Institute of Phonetics, Charles University, Czech Republic
alzbeta.houzar@ff.cuni.cz; tomas.nechansky@seznam.cz; radek.skarnitzl@ff.cuni.cz

ABSTRACT

The human voice is characterized by enormous flexibility and plasticity, which is highly relevant for speaker identification. Even though naturalistic voice disguise in the forensic context tends to manifest little sophistication, some perpetrators have been shown to change atypical or more speech parameters. This paper aims to find how and whether fifteen targeted manipulations in the articulatory and phonation domains affect the characteristics of the fundamental frequency, long-term formant values, harmonicity, and the spectral slope when compared to the ten Czech male speakers' habitual voice. Whilst we anticipated that formant values would be influenced by articulatory settings rather than phonatory modifications; and that, on the other hand, the spectral slope and the harmonics-to-noise ratio would mostly reflect voice quality changes, our results turned out to be less straightforward, and it would be misleading to draw such general conclusions. Possible reasons for these findings are discussed.

Keywords: voice manipulation, fundamental frequency, formants, harmonicity, spectral slope

1. INTRODUCTION

In everyday life, a speaker's voice characteristics change due to a wide range of factors. These include, first, behavioural factors stemming from speakers themselves, such as using different speech styles [1, 2], being under the effect of various affective states and especially emotions [3, 4], speaking in a loud voice [5] or using whispered [6] speech. The sound of our voice also changes when we are speaking in different languages [7, 8] or even accents [9]. The person to whom we are talking may also affect our speech production, in what is referred to as phonetic accommodation, or vocal convergence toward one's conversation partner [10]. The second group of factors may be described as physiological; these include the effects of alcohol consumption [11], vocal fatigue [12], or the time within the female menstrual cycle [13] on speech; all these are also known to affect the sound of a speaker's voice. Naturally, the voice changes with time as well; various studies have examined the effect of different time spans on the voice, from changes within one day [14] to changes across a number of years [15].

All these shifts in our voices, all this within-speaker variability is made possible by the tremendous plasticity of the speech production mechanism [16, 17]: while our physiology imposes some limits on the sound of our voice, it allows for great variability, with countless degrees of freedom, that we use on a daily basis to express the various components of communicative intent [18]. This intra-speaker variability is, of course, of vital importance in the forensic phonetic context: an analyst comparing two voices and deciding on or against their identity must be aware of possible within-speaker differences.

One behavioural factor, which has not been listed above and which is also crucial in forensic voice comparison, is intentional voice disguise. Voice disguise refers to a speaker's deliberate attempt to modify the sound of their voice and thus to conceal their vocal identity. Forensic speech examiners encounter voice disguise particularly in crimes where the perpetrator suspects that they may be recognized (for example, in an anonymous blackmail telephone) or that the telephone call may be recorded (when calling an emergency line, for instance) [19].

When disguising their voice, perpetrators typically rely on shifting their speaking fundamental frequency (typically upwards for male speakers and downwards for females [19, 20]) placing a foreign object in front of or into their mouth (such as holding a tin can in front of their mouth as a resonator, covering their mouth with a handkerchief or cloth, or holding a pen between their teeth [21]), on imitating a regional dialect or foreign accent, changing the temporal patterning of their speech (e.g., monotonous syllable durations, unnatural pausing), or changing their voice by modifying some articulatory or phonatory settings [21, 22]. Fortunately, sophisticated complex shifts, involving more of the above-mentioned voice disguise strategies, are quite rare.

Outside of the forensic context, studies instructing speakers to change their voice as much as they can seem to confirm the rather simple strategies adopted by most speakers. However, a paper on speakers of Czech [23] identified several speakers who performed multiple, sophisticated modifications and still managed to sound natural; these changes resulted in significantly lower recognizability of the speakers. In his seminal study, Nolan examined the effect of targeted voice manipulations on the speech signal,

using his own voice [18]. In this study, we focus on targeted voice disguise in speakers of Czech and examine the effect of fifteen modifications on long-term formant, spectral, harmonic and f_0 characteristics.

We hypothesize that (1) formant values will mostly be affected by articulatory modifications, as they are considered an acoustic correlate of articulation setting, with F3 being relatively most stable [24]; (2) spectral and harmonic characteristics will depend mainly on phonatory modifications; and (3) f_0 will remain relatively stable across most settings, as speakers were instructed to only perform the targeted manipulation and not change other aspects of their speech like pitch.

2. METHOD

2.1. Material

Ten male speakers of Common Czech were analyzed in this study. All of them are experienced voice users, trained in phonetics, and they were generally able to perform the targeted voice manipulations. They were instructed to read a short text (translation of the Rainbow Passage) in their habitual voice, and fifteen more times, always performing a different modification.

The modifications were chosen mostly based on the SVPA scheme [25]:

- lip-spreading and lip-rounding
- closed jaw and open jaw
- palatalization and pharyngealization
- nasalization and denasalization
- pressed, breathy, whispery, and creaky voice

In addition, we included three combinations of one phonatory and one articulatory modification:

- spread lips and breathy voice
- labialized (rounded lips) and whispery voice
- open jaw and creaky voice

The recordings were obtained at the sound-treated recording studio of the Charles University's Institute of Phonetics, using 48kHz sampling frequency and 16-bit quantization. The speakers were told to repeat a passage in case they failed to maintain the targeted voice manipulation; the recordings were later edited so as to contain the best realization of each sentence vis-à-vis the intended modifications. As two speakers were not able to perform all the modifications, we analyzed 157 recordings in total (10 speakers * 16 versions – 3 not performed).

2.2. Analyses

Having extracted vocalic streams in Praat Vocal Toolkit [26], we dealt with the following acoustic signal domains: the fundamental frequency (f_0), long-term formants (LTF), harmonics-to-noise ratio (HNR), and long-term average spectrum (LTAS). All values were obtained automatically in Praat [27].

As for f_0 values, the median and alternative baseline (the level below which 7.64% of f_0 values fall [28]), were extracted in the 60–350 Hz range. Regarding LTF [29], first to third formant values were measured every 10 ms with the default 'robust' settings; then, all values below the 5th and above the 95th percentile were removed. For HNR [30], the floor of 60 Hz and standard cross-correlation settings for harmonicity extraction were used. Finally, as concerns LTAS, we focused on three spectral slope parameters: the Hammarberg index [31], which reports the difference between spectral amplitude (dB) maxima in the 0–2 and 2–5 kHz ranges; the alpha ratio [32], which describes the energy (Pa^2·s) ratio between the 0–1 and 1–5 kHz ranges; and the BgNoF0 index [33], which corresponds to the energy (Pa^2·s) ratio between the 350–1100 and 2300–5500 Hz ranges, thus without (most of) the fundamental frequency and F2 ranges.

To assess whether a modification had influenced the speech parameters in a consistent way across all speakers, we conducted the Wilcoxon paired test and applied the Bonferroni correction.

For visualization purposes, the data were normalized, so that the manipulated voice values show the difference from the habitual voice values.

3. RESULTS

3.1. The fundamental frequency

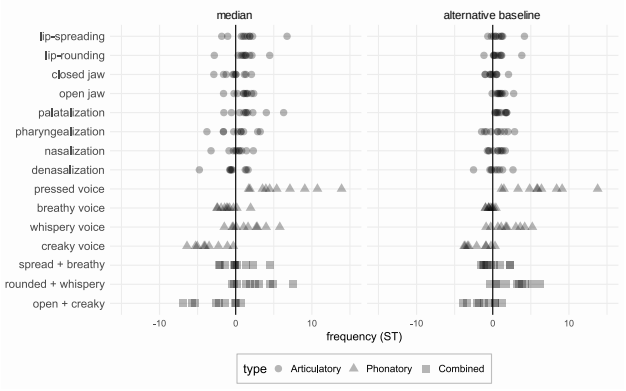

Figure 1: f_0 median and alternative baseline shifts (in ST) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

As is shown in Figure 1 and also in Table 1 below, there is a consistent shift of both the median and the alternative baseline in pressed phonation, which has also been confirmed statistically. As expected, a certain tendency could be observed in creaky voice as well; however, only a marginally significant result was returned regarding the f_0 median. It is noteworthy that the combined modifications copy the single

phonatory trends. As for the articulatory settings, only palatalized speech resulted in a homogeneous and statistically significant rise of the baseline value.

3.2. Long-term formants

Contrary to our assumptions, we witness shifts in long-term first to third formant values (LTF1–3) in relation to not only articulatory, but also phonatory manipulations. Even though LTF3 seems relatively stable in Figure 2, there are 5 significant comparisons (see Table 1). A significant increase/decrease in LTF1 has been revealed in 10 modifications; and 6 settings regularly influenced LTF2. Interestingly, lip-rounding caused a significant drop in all LTF values; and breathy phonation (both on its own and in combination with lip-spreading) lowered LTF1 but raised LTF2 and LTF3 medians.

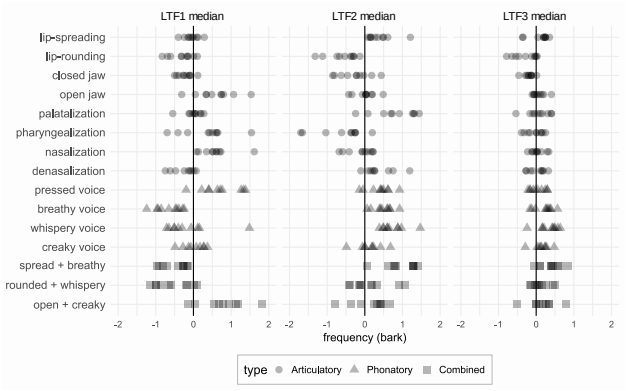

Figure 2: Long-term first to third formant shifts (in barks) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

3.3. Harmonics-to-noise ratio

Unsurprisingly, Figure 3 shows that speakers appear to have randomly oscillated around the HNR mean when performing articulatory manipulations. Regarding voice quality, which is more striking, the only significant exception is the breathy phonation alone and in combination; although creaky voice (again both single and combined) exhibits marginally significant results as well. As anticipated, creaky quality decreased HNR; nevertheless, breathy proved to have done the opposite.

3.4. The spectral slope

Finally, we were interested whether three spectral slope metrics would reflect differences between any of the manipulated and normal setting. As for BgNoF0, we cannot confirm this to be the case since no comparison turned out to be even marginally significant. Concerning the Hammarberg index and the alpha ratio, once again it was the breathy phonation that consistently shifted both indexes.

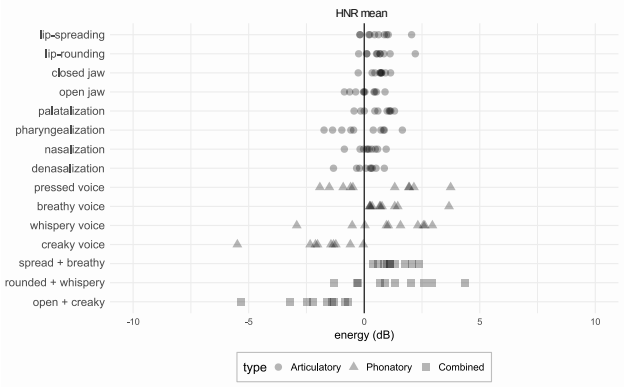

Figure 3: Harmonics-to-noise ratio shifts (in dB) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

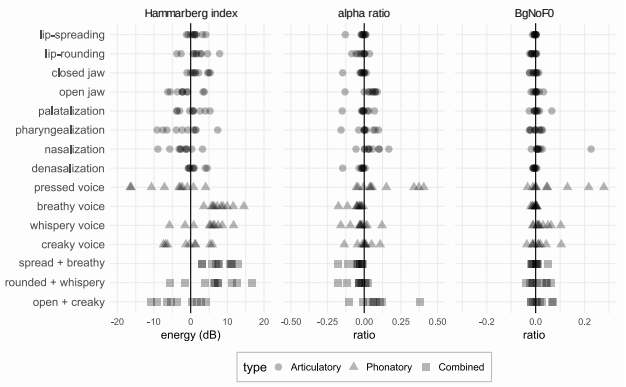

Figure 4: Spectral slope shifts (in Pa2·s and dB) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

To find out how all the voice manipulations shifted the studied acoustic parameters in individual speakers, see an interactive visualisation at http://tinyurl.com/voice-manipulations.

4. DISCUSSION

This study focused on targeted voice manipulations by experienced Czech voice users. The ability to perform (at least the majority of) these manipulations was crucial, and that is why the speaker sample is relatively limited: since performing these modifications is highly demanding, it would be unrealistic to obtain significantly more than ten speakers.

It should be pointed out that our results may have been affected by the different durations of extracted vocalic streams: due to the lack of periodic signal especially in phonatory settings, the durations ranged from 8.2 to 44.7 seconds (mean 29.5 s, SD 4.9 s).

As for the results, it seems that most manipulations brought along changes of parameters that were not predicted by our hypotheses. Some of

modification	median f_0	alternative baseline f_0	LTF1	LTF2	LTF3	HNR	Hammarberg index	alpha ratio	BgNoF0
lip-spreading	0.064	0.037	0.188	0.004	0.454	0.010	0.084	0.124	0.152
lip-rounding	0.064	0.049	**0.001**	**<0.001**	**<0.001**	0.010	0.193	0.124	0.027
closed jaw	0.722	0.922	**0.001**	0.007	**<0.001**	0.004	0.020	0.059	0.260
open jaw	0.084	0.004	**<0.001**	0.188	0.679	0.799	0.193	0.105	0.721
palatalization	0.049	**0.002**	0.934	**0.001**	0.599	0.024	0.625	0.906	0.770
pharyngealization	0.846	0.375	0.121	**<0.001**	0.421	0.846	0.275	0.415	1.000
nasalization	0.557	0.105	**<0.001**	0.041	0.934	0.193	0.064	0.037	0.037
denasalization	0.910	0.426	**0.002**	0.020	0.855	0.496	0.203	0.012	0.058
pressed	**0.002**	**0.002**	**<0.001**	0.004	0.978	0.322	0.037	0.084	0.124
breathy	0.064	0.049	**<0.001**	**<0.001**	**0.002**	**0.002**	**0.002**	**0.002**	0.286
whispery	0.084	0.014	0.064	**<0.001**	**0.001**	0.105	0.037	0.160	0.041
creaky	0.004	0.012	0.952	0.761	0.542	0.004	0.570	0.910	0.570
spread + breathy	1.000	0.846	**<0.001**	**<0.001**	**<0.001**	**0.002**	**0.002**	**0.002**	0.919
rounded + whispery	0.020	0.004	**0.001**	0.762	0.073	0.049	0.014	0.037	0.557
open + creaky	0.030	0.074	**0.001**	0.426	0.715	0.004	0.203	0.098	0.359

Table 1: Significant ($p \leq 0.0033$, in black) and marginally significant ($p \leq 0.0066$, in darker grey) p-values returned by the Wilcoxon paired test after Bonferroni correction ($\alpha = 0.05/15$).

these divergences should not be surprising, however, given the multiple interconnections of the phonatory and articulatory apparatus via the palatopharyngeus muscles, or of tongue position with nasality via the palatoglossus muscle [34]. For instance, most articulatory manipulations also triggered a change in fundamental frequency. This seems understandable with palatalization or pharyngealization, modifications which pull the larynx upwards or downwards, respectively, but is much less straightforward, for example, with lip-spreading.

One of our hypotheses stated that long-term formant values, being dependent on the shape of the vocal tract, would be mostly affected by the articulatory settings. Whereas we were able to confirm the widely known influence of lip-rounding, palatalization and pharyngealization, or jaw opening and closing on the drop or rise of either one or both F1 and F2, other findings were less common.

Firstly, formant values were significantly shifted in all settings but creaky phonation, which implies an articulation change above the larynx. Secondly, our data show an LTF1 increase in nasalization and a decrease in denasalization, which is probably caused by movements of the larynx when anchoring the velum to an opened and closed position, respectively. Thirdly, even though F3 is regarded as somewhat speaker-specific, it partially does reflect vowel quality. Based on our data, it is possible to confirm a stable F3 drop for lip-rounding and the closed jaw modification; with open articulation, the opposite trend is visible, though not significant (see Table 1).

Concerning phonatory modifications, it should be mentioned that even though HNR is usually measured in the middle third of sustained vowels, some researchers [35] do make use of long-term vocalic streams similarly to the presented paper. Creaky, whispery, and breathy phonation were anticipated to result in lower HNR since such voice qualities are expected to be less periodic than the modal voice. Whereas such a tendency was observed with creaky voice in our data (marginal significance), breathy phonation yielded the opposite result.

Spectral slope is considered another acoustic correlate of voice quality; hence, we assumed that phonatory settings would consistently shift the spectral slope values. Steeper spectral slope, with less energy in higher frequencies, has been reported for breathy phonation [33], which is also supported by our results. Conversely, creaky [31] and pressed [16] voice have been associated with flatter spectral slope, but neither has been confirmed in the present study.

To conclude, voice disguise is a phenomenon with which forensic voice comparison practitioners have to calculate [19, 22]. It would be beneficial, upon detection that a given case involves voice disguise, to be able to extrapolate acoustic values between speakers' disguised voice used in unknown recordings and their habitual voice obtained, for example, during interrogation. However, the multiple differences observed in our data beyond those that were predicted, as well as some acoustic manifestations which countered our hypotheses, do not warrant such a procedure as realistic. Nevertheless, our study provides another evidence of the fascinating plasticity of our voices.

5. ACKNOWLEDGEMENTS

This study was supported by the Grant Schemes at CU, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935; the last author was supported by the ERDF-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).

6. REFERENCES

[1] M. Jessen, "Forensic phonetics and the influence of speaking style on global measures of fundamental frequency," in *Formal linguistics and law*, G. Grewendorf and M. Rathert, Eds. Mouton de Gruyter, 2009, pp. 115–139.

[2] K. McDougall, and M. Duckworth, "Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English," *Int. J. of Speech, Lang. and Law*, vol. 25, pp. 205–230, 2018.

[3] E. J. Eriksson, R. D. Rodman, and R. C. Hubal, "Emotions in speech: Juristic implications," in *Speaker classification I*, C. Müller, Ed. Springer-Verlag, 2007, pp. 152–173.

[4] K. R. Scherer, "Acoustic patterning of emotion vocalization," in *Oxford handbook of voice perception*, S. Frühholz and P. Belin, Eds. Oxford University Press, pp. 61–91.

[5] H. Traunmüller, and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acoust. Soc. Am.*, vol. 107, pp. 3438–3451, 2000.

[6] Y. Swerdlin, J. Sith, and J. Wolfe, "The effect of whisper and creak on vocal tract resonances," *J. Acoust. Soc. Am.*, vol. 127, pp. 2590–2598, 2010.

[7] S. X. Chen, and M. H. Bond, "Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context." *Pers. Soc. Psychol. B.*, vol. 36, pp. 1514–1528, 2010.

[8] B. Lee, and D. Van Lancker Sidtis, "The bilingual voice: Vocal characteristics when speaking two languages across speech tasks," *Speech Lang. Hearing*, vol. 20, pp. 174–185, 2017.

[9] B. G. Evans, and P. Iverson, "Plasticity in vowel perception and production: A study of accent change in young adults, *J. Acoust. Soc. Am.*, vol. 121, pp. 3814–3826, 2007.

[10] J. S. Pardo, "On phonetic convergence during conversational interaction," *J. Acoust. Soc. Am.*, vol. 119, pp. 2382–2393, 2006.

[11] B. Baumeister, C. Heinrich, and F. Schiel, "The influence of alcoholic intoxication on the fundamental frequency of female and male speakers," *J. Acoust. Soc. Am.*, vol. 132, pp. 442–451, 2012.

[12] V. J. Boucher, and T. Ayad, "Physiological attributes of vocal fatigue and their acoustic effects: A synthesis of findings for a criterion-based prevention of acquired voice disorders," *J. Voice*, vol. 24, pp. 324–336, 2010.

[13] M. Hejná, "A case study of menstrual cycle effects: global phonation or also local phonatory phenomena?", *Proc. 19th ICPhS*, Melbourne, 2019, paper 13.

[14] M. Artkoski, J. Tommila, and A.-M. Laukkanen, "Changes in voice during a day in normal voices without vocal loading," *Log. Phon. Vocol.*, vol. 27, pp. 118–123, 2002.

[15] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *Int. J. of Speech, Lang. and Law*, vol. 24, pp. 177–199, 2017.

[16] J. Laver, *The phonetic description of voice quality*. Cambridge University Press, 1980.

[17] F. Nolan, "Degrees of freedom in speech production: An argument for native speakers in LADO," *Int. J. of Speech, Lang. and Law*, vol. 19, pp. 263–289, 2012.

[18] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge University Press,1983.

[19] A. Braun, "Stimmverstellung und Stimmenimitation in der forensischen Sprechererkennung," in *Das Phänomen Stimme: Imitation und Identität*, T. Kopfermann, Ed. St. Ingbert, 2006, pp. 177–181.

[20] H. J. Künzel, "Effects of voice disguise on speaking fundamental frequency," *Forensic Linguist.*, vol. 7, pp. 149–179, 2000.

[21] R. M. Figueiredo, and H. S. Britto, "A report on the acoustic effects of one type of disguise," *Forensic Linguist.*, vol. 3, pp. 168–175, 1996.

[22] H. Masthoff, "A report on a voice disguise experiment," *Forensic Linguist.*, vol. 3, pp. 160–167, 1996.

[23] A. Růžičková, and R. Skarnitzl, "Voice disguise strategies in Czech male speakers," *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, pp. 19–34, 2017.

[24] R. Skarnitzl, and T. Nechanský, "Segmental cues," in *Oxford handbook of forensic phonetics*, F. Nolan, K. McDougall, and T. Hudson, Eds. Oxford University Press, 2023 (in print).

[25] E. San Segundo, and J. Mompean, "A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity," *J. Voice*, vol. 31, pp. 644.e11–644.e27, 2017.

[26] R. Corretge, *Praat Vocal Toolkit*, retrieved from https://www.praatvocaltoolkit.com, 2022.

[27] P. Boersma, and D. Weenink, *Praat*, retrieved from http://www.praat.org, 2020.

[28] J. Lindh, and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Proc. Interspeech*, Antwerpen, 2007, pp. 2025–2028.

[29] F. Nolan, and C. Grigoras, "A case for formant analysis in forensic speaker identification," *Int. J. of Speech, Lang. and Law*, vol. 12, pp. 143–173, 2005.

[30] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, pp. 1544–1550.

[31] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolaryngol.*, vol. 90, pp. 441–451, 1980.

[32] J. Sundberg, and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *J. Acoust. Soc. Am.*, vol. 120, pp. 453–457, 2006.

[33] J. Volín, and J. Zimmermann, "Spectral slope parameters and detection of word stress," *Technical Computing Prague*, pp. 125–130, 2011.

[34] A. Marchal, *From speech physiology to linguistic phonetics*. John Wiley & Sons, 2009.

[35] V. Hughes, A. Cardoso, P. Foulkes, P. French, A. Gully, and P. Harrison, "Forensic voice comparison using long-term acoustic measures of laryngeal voice quality," *Proc. 19th ICPhS*, Melbourne, 2019, paper 639.