THE EFFECTS OF THE ONLINE VISUALIZATION OF ACOUSTIC-PROSODIC FEATURES OF SPEECH ON SPEAKERS' PRODUCTIONS

Kerstin Fischer[1] *and Oliver Niebuhr*[2]

[1]*Department of Design and Communication,* [2]*Centre for Industrial Electronics*
University of Southern Denmark
[kerstin,olni]@sdu.dk

Abstract: In this paper, we investigate the role of concomitant, visual feedback on speakers' speech productions. While the effects of visualizations of prosodic features have been addressed in prior work, usually very broad comparisons were made, such as between audio- and audio-visual feedback, or the effect under consideration did not concern speech production, but rather elicited speakers' preferences. In the current study, we compare speakers' productions when using two different versions of the same visualization tool; specifically, the tool visualizes speakers' intonation contours, emphasis, and, in one condition, also creaky voice. The results show that speakers produce the target utterances with falling contours significantly better when using the tool that visualizes creaky voice. These findings suggest that the visualization of acoustic-prosodic features can have a significant impact on speakers' linguistic productions.

1 Introduction

In this paper, we investigate the effects of the online visualization of phonetic / prosodic features on speakers' speech productions. Within our recent project, we have developed an online tool that provides speakers with visual feedback on their speech productions with the aim to sensitize young speakers of Scandinavian languages with respect to the prosodic properties of their neighboring languages.

In the current study, we investigate whether such a tool can indeed influence people's productions. Specifically, we evaluate the effects of online, concomitant visualization by presenting participants with one of two versions of the tool: In one condition, participants trained the use of falling intonation with the visualization tool that included a visualization of creaky voice, whereas in the other condition, participants trained with the same tool, just without a visualization of creaky voice. Our hypothesis is that the visual online feedback that provides clear information of a target pronunciation, here on the use of creaky voice, will help participants to improve their speech productions significantly.

2 Previous Work

Many prosodic notation systems have been developed over the course of the past fifty years, often from different theoretical perspectives and not always for pedagogical purposes (cf. [1]). Correspondingly, only few notation systems have been evaluated with respect to whether they enable language learners to adapt their own productions based on the visualized information. Schaefer et al. [2], for instance, study the effects of different visualizations, yet they do not investigate whether these visualizations facilitate learners' productions. Specifically, in their study, participants were presented with videos showing either pitch or loudness on the y-axis or pitch on the y-axis and loudness in terms of dot-size. They asked the participants which kinds of presentations they preferred; participants indicated that they preferred the dynamic

presentation of pitch on the y-axis and loudness in terms of dot size. However, preference judgments may not inform us about the extent to which the notation system can be processed by learners online and lead to better productions.

That visual feedback can significantly improve learners' segmental production has been demonstrated, for instance, by Offerman & Olsen [3], who allowed students in the experiment group to practice the voice onset time of plosives in Spanish while watching visual representations (spectograms) of both a native speaker's and their own productions. Concerning suprasegmental features, some studies have addressed the effect of visualizations either without a control group (e.g. [4]), or compared to other modalities; for instance, De Bot & colleagues (e.g. [5-6]) test their version of audio-visual feedback in comparison with audio feedback only. Such a procedure is however slightly problematic because also memory issues can cause the advantage of the audio-visual feedback (instead of just auditory feedback) in the practicing sessions. It is thus preferable to employ the same modality in order to test the effects of specific visualizations.

More recently, the effectiveness of the Pitcher and its online equivalent, the Web-Pitcher, visualization and feedback tools for prosodic learning, has been repeatedly demonstrated in production experiments ([7-8]). However, the experiments had a within-subjects design and were hence not fully able to separate familiarization effects from the effects of the provided prosodic feedback.

To sum up, previous work has documented potential benefits of the visualization of prosodic information, yet either in comparison to baseline conditions that differed from the experimental condition in several ways or without assessing whether speakers really benefit from the feedback. This research gap is addressed in the current study.

3 The Visualization Tool

The online visualization tool was developed to support young Scandinavians in getting acquainted with their neighboring languages in the framework of a project funded by Nordplus. The tool is freely available under http://nordplus.sonoware.de. It provides visual feedback on spoken utterances with different display options; for instance, it allows the user to choose the degree of stylization, the visual display of prosodic stress, or the display of raw data. Furthermore, it indicates creaky voice in the form of a green line at the bottom of the window. For the current study, we used two different versions of the tool in order to determine the effect of visual feedback: one with and one without the detection and separate visualization of creaky voice. In this way, participants in both conditions saw exactly the same tool just with one specific feature added in one condition. This ensures that participants have essentially the same experience, and that they receive feedback in the same modality, while allowing us to study the effects of visual feedback on one aspect of prosodic implementation, namely the depth of the final fall, which, if fully executed, can result in creaky voice at the ends of utterances (or major prosodic-phrase boundaries, cf. [9]).

4 Research Subject

Previous work has shown that deep falls are associated with charismatic speech [7]. However, creak phonation, also known as glottalization, fulfills many other functions in speech communication as well, see Davidson [9] for a recent overview. In informal, spontaneous speech (both in German and in English varieties), it represents an allophonic variant of voiceless stop consonants, the phoneme /t/ in particular. Furthermore, it serves in many of the world's languages as an acoustic cue to phrase, turn, and topic boundaries [9]. In tone languages like Mandarin or Cantonese, it is involved in signaling tonal contrasts. In the Scandinavian languages, for which the visualization tool has been developed, we find creak as a phonetic correlate of Danish stød

[10] as well as in connection with word-accent I in varieties of Swedish [11]. Beyond its use as a linguistic marker of phonemes, tones or prosodic structure, creaky voice is often found under certain sociophonetic conditions (e.g., in the speech produced by young US women) or in combination with expressing affects or emotions like irritation in Vietnamese, embarrassment in Japanese or sarcasm and disgust in Germanic languages [12-13].

In summary, creak phonation is a phonetic feature with a high functional load, where the form-function links differ across the languages of the world, including the Scandinavian ones. That is, language learners have to practice to employ creaky voice properly in their new language. Yet, there is currently no visualization or feedback tool that can assist them in practicing. It is mainly for these reasons that the creaky-voice detector was developed and selected for testing in the present experiment.

5 Method

The study was carried out as a between-subject experiment.

5.1 Speech Material

The speech material elicited for the experiment consists of 20 isolated German sentences that should be read with a final fall, so deep that it reaches until the bottom of the speaker's pitch range, where it then turns into creaky voice phonation. The sentences are ordered in such a way that it becomes increasingly difficult to produce them with final falls that end in creak. For example, one level of difficulty is added by turning confirmative statements (*Das sehe ich genauso*, 'I totally agree') into contradicting statements (*Das sehe ich ganz anders*, 'I completely disagree'), for which speakers are often reluctant to fall so low that their voice becomes creaky. Similarly, matter-of-fact statements (*Dafür haben wir aktuell keine Zeit*, 'We don't have time for that at the moment') are turned into apologetic statements (*Entschuldigung, aber dafür haben wir leider keine Zeit*, 'I am sorry, but we don't have time for that, unfortunately'), which also typically end without creak and higher in pitch than their matter-of-fact counterparts. Another level of difficulty was related to prosodic structure. Specifically, in one set of sentences, the nuclear pitch-accented syllable (which typically shows a high tonal accent in German, cf. [14]) was at least three syllables away from the end of the sentence, hence leaving sufficient time for the speakers to produce a terminal low sentence-final fall. By contrast, in another set of sentences, the nuclear pitch-accented syllable was in final or penultimate position in the sentence.

5.2 Participants

Participants were recruited online via the social networks of the researchers in the research group; the sample is thus a convenience sample that mostly comprises the friends and family of the respective student assistants as well as a group of students in a developmental psychology class. The mean age of the participants is 23.3, with a range from 19 to 29. Twenty-one identify as female, five as male, while none preferred not to say. Nine participants saw the tool in the creaky voice condition, whereas 17 saw it without the creaky voice detector.

The experiment was carried out by five different experiment leaders: the two first authors, a friendly colleague, as well as two student assistants. In order to ensure that all participants received the same instructions, a script with the instructions to the participants was created, which the experiment leaders used to recruit the participants, and which was used in the online questionnaire to guide the participants through the experiment.

5.3 Procedure

When entering the online survey, participants were first asked to fill out a consent form, which informed them about the goal of the study, namely to investigate to what extent the visualizations by the tool help them to produce utterances with deep falls. It also provided them with a motivation for practicing this particular prosodic feature, namely that it may contribute to them sounding more charismatic. Furthermore, the consent form contained information about data storage and handling in accordance with GDPR.

Once participants had given their informed consent, about participants were asked to read a set of sentences with falling intonation and to record themselves while doing so. Then the participants were introduced to the respective version of the tool and instructed how they can record themselves and see the visualizations of their utterances. The participants in the creaky voice condition were pointed to the green line as an indicator that the utterance is falling deeply enough. After participants had read the sentences, they were asked to save the recordings in the tool, to download them to their own computers and to send them to the respective researcher for analysis. The elicited 20 sentences per participant were saved in a single audio file in m4a format, but at a low compression level that should not have significant negative effects on f0 (see [15]).

5.4 Data Analysis

The 20 elicited sentences per participant were acoustically analyzed automatically in PRAAT by means of the script of de Jong & Wempe [16]. Some data had to excluded from the analysis because some speakers produced rises instead of falls. Since f0 measurements, and those related to local f0 minima in particular, are typically quite error-prone, the script's output underwent a plausibility check for each participant. To that end, a Pitch Object was computed for the audio file of each participant in PRAAT. Sentence-final octave errors in the Pitch Object were manually corrected, and then the script's output was compared on a visual basis to the f0-values measured in the Pitch Object. In case of conflicting values, the f0-value in the script's output was replaced by the one determined manually in the Pitch Object.

Then, three measures, i.e. dependent variables, were determined: (1) the *Fall Range* in semitones (st), (2) the final *Fall Velocity* in semitones per second (st/s), and (3) the presence/absence of sentence-final creak phonation, determined according to the principles explained in Keating et al. [17]. Measures (1)-(2) were derived from the raw f0 measurements. The Fall Range represents the difference between the speaker's mean f0 level and the f0 minimum produced at the end of the respective sentence. The Fall Velocity is the semitone difference between the last and third-last measurable f0 value divided by the time interval in between these two values. The total number of sentences ending in creaky voice was counted per participant and is henceforth referred to as the *Creak Count*.

The data per participant of Fall Range, Fall Velocity, and Creak Count were statistically analyzed using the non-parametric equivalent of a multivariate ANOVA, i.e. a Kruskal-Wallis test for k independent samples. This more conversative test was selected with respect to the small participant samples.

6 Results

The Kruskal-Wallis test for k independent samples resulted in two significant effects of the pitch-visualization difference on the prosodic measures. The two effects concerned the Final Velocity ($H[1] = 8.96$, $p = 0.003$) and the Creak Count ($H[1] = 5.19$, $p = 0.023$). The effect related to Fall Range approached significance ($H[1] = 3.07$, $p = 0.080$).

As Figure 1(a) shows, the participants' sentence-final falls ended with a slope of about -10 st/s when there was no detection and visualization of creaky voice. By contrast, if there was such a detection and visualization, the final fall was realized about 50% faster or, since the underlying time interval between the two f0 measures was constant in all cases, about 50% steeper (-15 st/s) – and turned significantly more often into creaky voice at its end, according to the Kruskal-Wallis test. In fact, Figure 1(b) shows that the number of sentences ending in creaky voice was almost twice as high in the condition with than in the condition without the detection and visualization of creaky voice (10.4 vs. 5.1).

Two further points are worth noting in connection with this difference. First, even in the condition with the detection and visualization of creaky voice, the mean Creak Count across all participants was 10.4, which is still well below 20, i.e. the maximum Creak Count corresponding to total number of elicited sentences. Thus, the creaky-voice feedback we gave to speakers was significantly more effective than no feedback. Indeed, 10.4 out of 20 represents a higher frequency of occurrence than is typically reported in connection with low-falling phrase-final intonations in West Germanic languages [18-19]. Yet, there is obviously still considerable room for improvement as to the reliable elicitation of (sentence-final) creak by means of real-time acoustic feedback concepts. Second, although the phonetics literature often points out gender differences in the occurrence and use of creak (e.g., [9]), our own speaker samples contain indications for such a difference only in the condition without but not with the creak detector (due to the small sample sizes, we did not conduct any statistical tests). In the condition without the creak detector, phrase-final creak occurred on average more often for male than for female speakers (see [9], [19]).

The Fall Range did not differ significantly but the results on this dependent variables match with the overall pattern in that we found an on average about 1.3 st higher Fall Range in the condition with than without the detection and visualization of creaky voice (7.31 vs. 5.98 st). Given that the frequency range in which modal phonation turns into creak is to some extent biomechanically determined and hence fairly fixed for both male and female speakers (cf. [9]), the combined results of Fall Range and Fall Velocity suggest that the participants in the creak-visualization condition produced their sentences at a lower mean f0 level than those in the condition without creak visualization. Comparisons of absolute f0 measurements between the two samples indeed support that assumption.

Furthermore, we see in Figure 2 for one example participant (JUL) in the creak-visualization condition that there was something like a learning effect in the sense that both the sentence-final f0 minima and the sentences' mean f0 levels decreased successively across the 20 sentences. Thus, speakers got better at producing sentence-final creak as they proceeded in the list of sentences.

That gave the impetus to conduct a closer inspection of the results data: We divided the audio files for all 17 and 9 participants of the two conditions along half of the sentences and then measured the f0 minima, the mean f0 levels as well as the three dependent variables separately for the initial 10 and final 10 sentences. The paired measurements per participant were then checked for significant differences using Wilcoxon signed-rank tests (due to the small sample size no paired-sample t-test was used). The p-values of the multiple tests were corrected using the Benjamini-Hochberg method. In fact, on this basis we find a significant learning effect in the sense of Figure 2 (i.e. $p < 0.05$ for at least one of the above measures) for 5 of the 9 participants (55.5%) in the creak-visualization condition, but only for 6 of the 17 participants (35.3%) in the condition without creak-visualization.

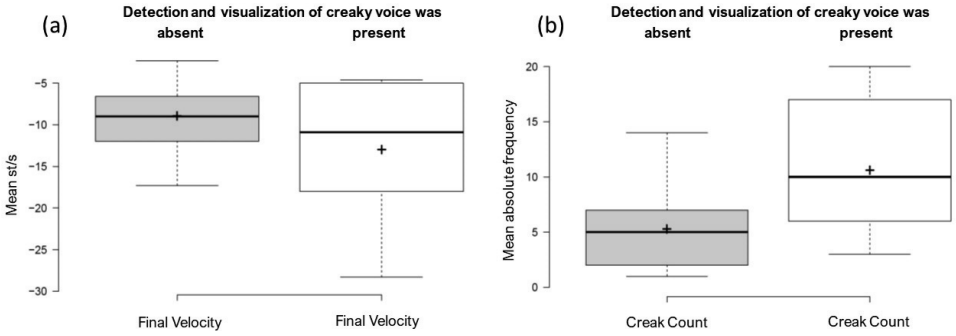

Figure 1 - Box-plot summaries of significant differences between the sentence-final intonation realizations with (N = 17) and without (N = 9) the detection and visualization of creaky voice.

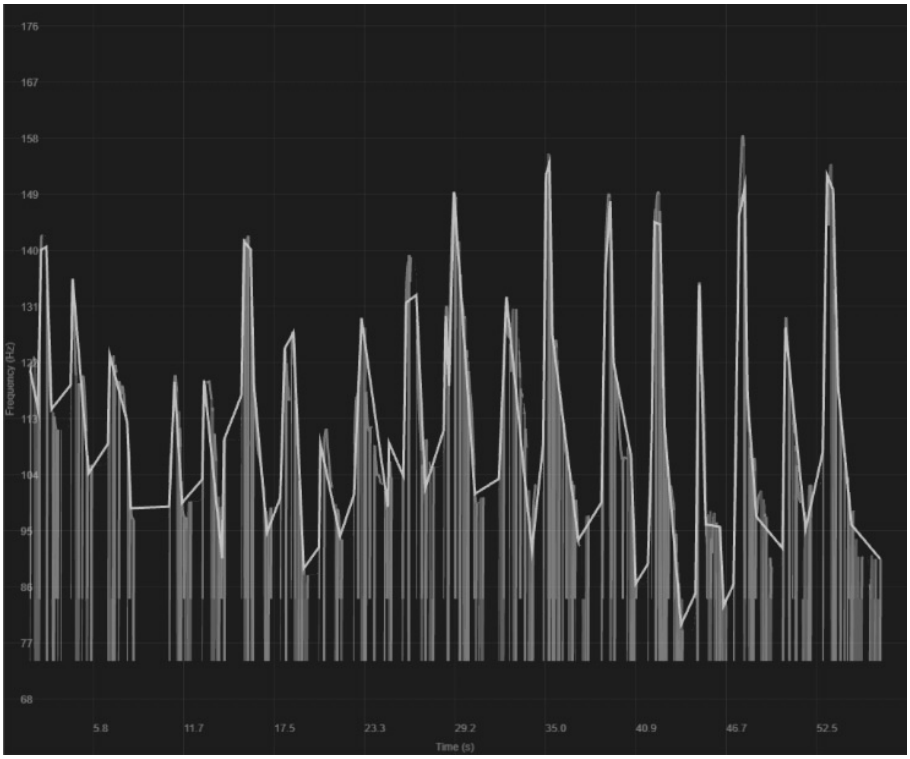

Figure 2 - Example of the 20 sentences realized with creaky voice detector by male participant JUL

7 Discussion

This study was carried out with a relatively small number of participants of a relatively restricted age range, who were recruited by word-of-mouth by several different researchers. These factors may limit the generalizability of the results to other populations, who are possibly not so used to online tools or who may also have a different educational background than the participants recruited, who were mostly students.

Nevertheless, for the respective population, we could show that providing concomitant visual feedback on speakers' utterances does have a positive effect on their productions, such that speakers were more able to produce intonation contours as instructed. Since all participants used the same visualization tool except for the presentation of creaky voice, we can rule out confounding factors like the modality in which feedback is presented. Furthermore, since we analyzed speakers' productions, we can show that the effect of the feedback provided is not only preferred, but also facilitates speakers' productions.

While creaky voice itself is an important feature whose appropriate use needs to be acquired when learning a new language, in the current study, it also served as very specific feedback for the tool users, namely whether their goal, to produce deep falling intonation contours, was achieved. Thus, the visualization of creaky voice served here as an indication of the target, which may have enforced the facilitative effect of the visualization. Whether the visualization of other acoustic-prosodic features will be equally beneficial will have to be determined on a case-by-case basis in future studies.

8 Conclusion

The results of the comparison of speakers' productions of sentences with falling intonation contours when exposed to two versions of the visualization tool developed in our Nordplus project indicated significant facilitative effects of visual feedback on the acoustic-prosodic features of their utterances. While the number and diversity of participants was somewhat restricted, we can conclude that at least for the population investigated, concomitant visual feedback of creaky voice leads to better speech productions.

Acknowledgements

This work was made possible through the funding by Nordplus for the project "Development of a Tool for the Visualization of Intonation." We would furthermore like to thank Rosalyn Langedijk, Marie U. Leistner and especially Nils F. Tolksdorf for their help with the data elicitation, as well as Ali Asadi for his support in the paper production.

References

[1] FISCHER, K., NIEBUHR, O., ALM, M. & SCHÜMCHEN, N., (submitted): *Intuitive Visualization of Intonation for Foreign Language Learners*.

[2] SCHAEFER, R. S., BEIJER, L. J., SEUSKENS, W., RIETVELD, T. C. M. & SADAKATA, M.: *Intuitive visualizations of pitch and loudness in speech*. Psychonomic Bulletin & Review, 23(2), 548-555, 2015.

[3] OFFERMAN, H. M., & OLSON, D. J.: *Visual feedback and second language segmental production: The generalizability of pronunciation gains*. System, 59, 45-60, 2016.

[4] LEVIS, J., & PICKERING, L.: *Teaching intonation in discourse using speech visualization technology*. System, 32(4), 505-524, 2004.

[5] DE BOT, K.: *Visual feedback on Intonation I: Effectiveness and induced practice behaviour*. Language and Speech, 26(4), 331-350, 1983. doi:10.1177/002383098302600402

[6] WELTENS, B. & DE BOT, K.: *The visualization of pitch contours: Some aspects of its effectiveness in teaching foreign intonation*. Speech Communication, 3, 157-163, 1984.

[7] NIEBUHR, O. & NEITSCH, J.: *Digital Rhetoric 2.0: How to Train Charismatic Speaking with Speech-Melody Visualization Software*. In: A. Karpov, R. Potapova (eds), *Speech & Computer*. Lecture Notes in Computer Science Vol. 12335, pp. 357-368. New York: Springer Nature, 2020.

[8] NIEBUHR, O.: *Computer-assisted prosody training: Improving public speakers' vocal charisma with the Web-Pitcher*. Revista da Abralin, 20(1), 1-29, 2021.

[9] DAVIDSON, L.: *The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages*. Wiley Interdisciplinary Reviews: Cognitive Science, 12(3), e1547, 2021.
[10] GRØNNUM, N., & BASBØLL, H.: *Danish Stød–Towards simpler structural principles?*. In Understanding Prosody, 27-46, De Gruyter, 2012.
[11] LUNDMARK, M. S., AMBRAZAITIS, G., & EWALD, O.: *Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents*. In Interspeech 2017: Situated interaction, 3236-3240, 2017.
[12] GOBL, C., & CHASAIDE, A. N.: *The role of voice quality in communicating emotion, mood and attitude.* Speech communication, 40(1-2), 189-212, 2003.
[13] YUASA, I. P.: *Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?*. American Speech, 85(3), 315-337, 2010.
[14] PETERS, B.: *Weiterführende Untersuchungen zu prosodischen Grenzen in deutscher Spontansprache*. Prosodic structures in German spontaneous speech (AIPUK 35a), 203-345, 2005.
[15] GE, C., XIONG, Y., & MOK, P. (2021). *How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements*. In 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, 1683-1687, 2021
[16] DE JONG, N.H. & WEMPE, T.: *Praat script to detect syllable nuclei and measure speech rate automatically*, Behavior research methods, 41(2), 385-390, 2009.
[17] KEATING, P. A., GARELLEK, M., & KREIMAN, J.: *Acoustic properties of different kinds of creaky voice*. Proc. 18th International Congress of Phonetic Sciences, Glasgow, UK, 2-7, 2015.
[18] KÖSER, S.: *Phrasen-finale Phonationsänderungen und ihre Rolle beim turn taking*. In: Barth-Weingarten, D. & Szczepek Reed, B. (Hrsg.): *Prosodie und Phonetik in der Interaktion – Prosody and phonetics in interaction*. Mannheim: Verlag für Gesprächsforschung, S. 20-45, 2014.
[19] BECKER, K., KHAN, S., & ZIMMAN, L.: *Creaky voice in a diverse gender sample: Challenging ideologies about sex, gender and creak in American English*. New Ways of Analyzing Variation, 44, 2015.