EXPLAINING VOICE CHARACTERISTICS TO NOVICE VOICE PRACTITIONERS — HOW SUCCESSFUL IS IT?

Jana Wiechmann[1], Frederik Rautenberg[2], Petra Wagner[1], Reinhold Häb-Umbach[2]

[1]Phonetics Work Group, Faculty of Linguistics and Literary Studies, Bielefeld University,

[2]Department of Communications Engineering, Paderborn University
jana.wiechmann@uni-bielefeld.de, rautenberg@nt.upb.de, petra.wagner@uni-bielefeld.de, haeb@nt.upb.de

ABSTRACT

Human voices are notoriously difficult to characterize. A suitable and consistent description of voice characteristics is crucial in many applied disciplines such as speech therapy or forensics. The present study examines the ability of novice voice practitioners (students of clinical linguistics) to characterize voices before and after an expert explanation of laryngeal, supralaryngeal and prosodic voice features. Results show that even short expert explanations lead to a higher agreement between expert and novices. Especially voice characteristics related to laryngeal and supralaryngeal settings remain a major challenge to identify. We suggest that voice conversion technology may be employed in the future to assist the explanation of voice characteristics.

Keywords: voice quality, voice characteristics, explaining, phonetics pedagogy

1. INTRODUCTION

An adequate characterization of voices and voice quality is crucial across different fields, e.g., phonetics, vocal coaching, speech therapy or forensics, and has led to different approaches for doing so [1, 2, 3]. However, the characterization of voices is not only a challenge for novices in these fields. Even experts rarely achieve a high inter-rater agreement [4, 5, 6].

One of the most popular approaches to voice characterization goes back to [7], based on which the Vocal Profile Analysis (VPA) scheme was developed [8]. The VPA focuses on the articulatory settings underlying voice characteristics, and has been adapted to simplify its usage [9, 10], for example by relying on binary rather than multilevel rating scales or reduced settings. Voice characterizations in speech therapy typically employ perceptual characterizations of voices with the help of multilevel rating scales, and focus on aspects of phonation, or voice quality, i.e., the GRBAS [11, 12] or the RBH scale in the German speaking community [13, 14]. Other approaches include prosodic features or more global parameters such as overall tension. The handling of intermittent features and of the parallel occurrence of features also differs between approaches [10, 15]. Especially in forensic phonetics, factors beyond voice quality are important. Therefore, speaking style, but also segmental aspects such as dialect and pronunciation, or even background noise in recordings, are taken into account [2].

As mentioned above, all existing perceptual assessment tools yet seem to be difficult to use in a consistent way. One reason for this might be a different understanding of the terminology, or different internal standards between raters. It has also been discussed that voice as a multidimensional phenomenon is difficult to assess in general, leading to high cognitive load. For this reason, complex assessment tools, especially those employing multilevel scales, can be overwhelming [10, 4]. To overcome some of these difficulties, calibration or training sessions on example voices were suggested [16] and the usage of prototypical anchor voices for individual voice features [4, 16]. Given that even prototypical voices are likely to differ in multiple dimensions of voice, a recent follow-up idea of this is the usage of state-of-the-art voice conversion technology to provide pairs of example voices differing only in one feature of interest [17]. Despite the above-mentioned difficulties, it is crucial for voice practitioners to learn about voice characterizations as part of their profession. We want to understand better, which aspects need to be improved in the training of voice characterizations, and what can be reached in single explanation sessions. The aim of this study therefore is

1. to assess if a short, single expert explanation using prototypical voices and imitation can help in learning how to classify voices,

2. to evaluate if the explanation leads to a higher consistency among raters regarding individual voices,
3. to discover which voice characteristics are easier or more difficult to classify.

To this end, we asked novice voice practitioners to characterize voices before and after a short, systematic, expert explanation, and compared their performances to those of experts.

2. METHODS

2.1. Selection of voice characteristics

We defined a catalogue[1] of 20 voice features to be explained and assessed (cf. Table 1). We decided to take into account a wide set of voice characteristics, including laryngeal settings related to phonation, supralaryngeal settings related to the resonance characteristics of the vocal tract, as well as prosodic characteristics related to (dynamic) loudness and pitch. Our selected features needed to fulfill two main criteria:

1. They can be expressed by fairly common adjectives in German (our participants' native language), to guarantee their fundamental conceptual familiarity to non-experts.
2. They have fairly well-understood underlying articulatory settings and perceptual characteristics, which are necessary for their explanation.

2.2. Participants and speech materials

20 students of clinical linguistics (female, mean age = 22.5) participated in the study and received monetary compensation. 20 voices (10 female, 10 male) were selected from The Nautilus Speaker Characterization (NSC) Corpus [18] for classification. The set of chosen voices covered the full set of characteristics in our feature set, and included 2 relatively "neutral" voices. From the corpus, semi-spontaneous simulated telephone calls of around 30 seconds were used for characterization. While none of the voices can be considered pathological, some of them contain characteristics that can feature in pathological voices.

2.3. Expert rater agreement and gold standard

To set a gold standard for voice characterization, two voice experts classified the 20 voices independently with respect to the 20 voice features in a binary manner. Afterwards, they jointly discussed controversial cases, to develop a more common understanding of the feature set. Subsequently, the experts rated the voices again. Inter-expert agreement before the discussion was Cohen's $\kappa = 0.61$, corresponding to 'substantial agreement'. After the discussion, their agreement was 'almost perfect' ($k = 0.85$). The ratings of the first author and expert explainer, were lastly defined as the gold standard (also cf. Table 1) to which the participants' ratings were subsequently compared.

2.4. Study phases

The study was divided into three steps: (1) A pre-explanation phase, in which participants classified a set of 10 voices, (2) an explanation phase, during which the expert explained the selected voice features (cf. Section 2.1), and (3) a post-explanation phase, during which participants rated the same set of voices again. During the voice ratings (pre- and post-explanation phase), each participant listened to 10 voices (out of 20) in a pseudorandomized order (each participant started with a different voice, male and female voices alternated). Voices were distributed across participants so that every voice was classified 10 times before and 10 times after the explanation. Participants were allowed to listen to each voice twice if necessary, and were instructed to mark a voice feature on a prepared list if they perceived it, even if it was only present temporarily or not very prominent. There was no minimal or maximal amount of features that listeners could select. Participants could also add features that they heard, but which were not listed. During the explanation phase, the expert explained how the features emerge anatomically, and how they manifest acoustically and perceptually. Additionally, the expert imitated the voice characteristics and played prototypical examples. To increase consistency across explanations, the expert followed a script, but explanations remained spontaneous and interactive: Participants were allowed to ask clarification questions, comment and provide feedback. The explanations lasted around 15-20 minutes each. Once the participants expressed that they felt confident enough to perform the classification task again, the post-explanation phase was initiated.

2.5. Rater agreement and consistency analysis

We used Cohen's κ to assess (1) the intra-rater agreement among participants before and after the explanation as well as (2) the agreement between the individual participants and the gold

Class	Voice Characteristics
laryngeal	creaky (10), hoarse (7), breathy (4), whispery (3), cracked (2), harsh (2), soft (2)
supralaryngeal	dark (2), bright (3), nasal (4), retracted tongue and/or raised larynx (4)
prosodic (pitch)	low (2), high (4), highly variable pitch (3)
prosodic (loudness)	loud (3), quiet (4), highly variable loudness (1)
mixed	shrill (1), resonant (4), tense (2)

Table 1: List of perceptual voice characteristics (English translations used in the perception study, classified into largyngeal, supralaryngeal, prosodic (pitch and loudness related) and mixed. Numbers in brackets indicate the total number of occurrences of each voice characteristic (based on the gold standard for our data set, cf. Section 2.3)

.

standard before and after the explanation. To assess the consistency among raters regarding the individual voices and voice characteristics, Fleiss' κ was calculated. Agreement was interpreted according to [19], differentiating the following agreement categories: 'almost none', 'slight', 'fair', 'moderate', 'substantial', 'almost perfect'. The statistical analyses were carried out using the *irr* package in R [20, 21].

3. RESULTS

3.1. Participants intra-rater agreement

We calculated the intra-rater agreement for each participant which varied from $\kappa = 0.27$ to $\kappa = 0.62$ ($M = 0.49$, $SD = 0.092$). Most raters achieved 'moderate agreement', one 'substantial', and three 'fair agreement' comparing pre and post per participant.

3.2. Participant-expert agreement

Twelve participants show a higher agreement with the gold standard after the explanation, two remain identical (while not showing a high intra-rater consistency, either: P15: $k = 0.48$, 'moderate', P18: $k = 0.27$, 'fair'), and six even decrease in their agreement with the expert (cf. Figure 1). The mean agreement with the gold standard increases minimally but not significantly (mean $\kappa = 0.26$ pre-explanation, mean $\kappa = 0.27$ post-explanation).

3.3. Consistency analysis per voice

Figure 2 shows the agreement across all participants per voice. Thirteen voices increase in agreement after the explanation, while seven voices decrease. Mean agreement across all voices before the explanation is $\kappa = 0.22$ and after $\kappa = 0.23$, increasing minimally but not significantly. Three

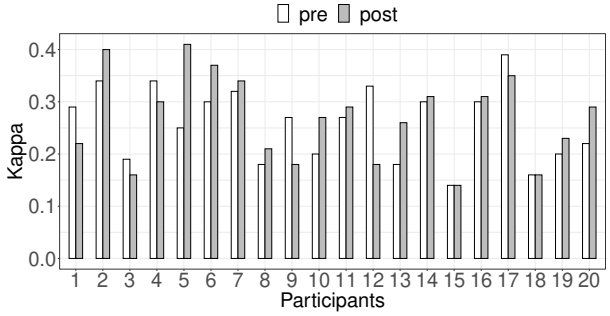

Figure 1: Agreement (Cohen's κ) between each participant and the expert before and after the explanation.

voices decrease in agreement, and even changed the agreement category after the explanation: Two of them went from 'moderate' to 'fair' and one from 'fair' to 'slight'. Five voices increased in agreement and changed the agreement category: One voice went from 'fair' to 'moderate', all others went from 'slight' to 'fair'. The remaining voices stayed within the same category (most of them 'fair'), one of them showing substantial improvement within the category. All other voices show little change. The voice showing the highest agreement score reaches this after the explanation ('moderate').

3.4. Consistency analysis per voice feature

Table 2 shows the agreement across all participants for features showing the lowest and highest agreement, or the strongest changes. In total, twelve features (out of twenty) have higher agreement after the explanation while eight have lower or almost identical agreement. Mean agreement across all voice features before the explanation is $\kappa = 0.22$, and after it $\kappa = 0.23$, showing minimal increase without being significant. With respect to agreement, two features changed the category: 'creaky' decreased from 'fair' to 'slight' and 'tense'

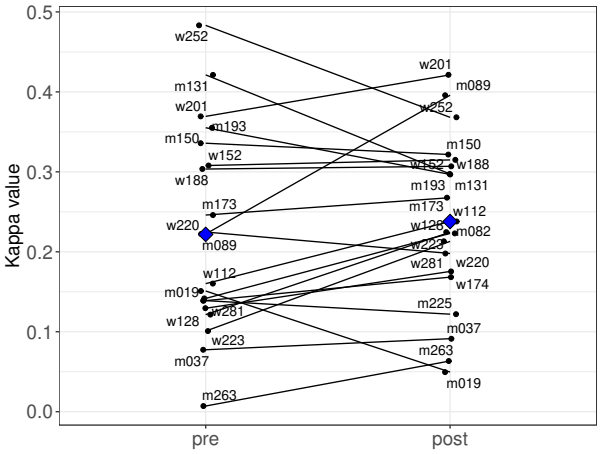

Figure 2: This figure shows the agreement (Fleiss' κ) across all participants per voice pre and post explanation. Means are indicated in blue.

Agreement	Feature	Pre	Post
highest	quiet	0.65	0.69
	loud	0.41	0.45
	high	0.32	0.35
lowest	low	0.13	0.11
	cracked	0.11	0.1
	breathy	0.1	0.09
highest increase	tense	0.03	0.21
	nasal	0.07	0.19
highest decrease	creaky	0.22	0.16
	hoarse	0.38	0.27

Table 2: This table shows the overall agreement (Fleiss' κ) of voice features before (pre) and after (post) the explanation, that showed the highest or lowest overall agreement, or the strongest increase or decrease in agreement.

increased from 'slight' to 'fair'. All other voice characteristics stayed within the same agreement category. Ten features stay within the category 'slight agreement', while six stay within 'fair'. Besides the features mentioned in Table 2, 'shrill', 'soft' and 'resonant' show a comparatively higher agreement than other voice features (above or equal to $\kappa = 0.26$).

4. DISCUSSION AND CONCLUSIONS

The first research question raised in this study was whether a single expert explanation of complex voice characteristics can actually help novice voice practitioners. The results show a clear tendency that the explanations indeed helped the majority of participants. Probably, the short amount of time available for the explanations has prevented a better success. Obviously, we cannot generalize our results to other settings, and the impact of the individual expert is left unexplored. With respect to our second question, our analysis revealed that the agreement differs vastly across different voices. This may have been caused by different levels of agreement for individual voice characteristics (research question 3). In fact, agreement is not distributed uniformly across these. Rather, some features were much more difficult to classify than others. In particular, our results show that prosodic features are rated much more consistently, with the notable exception of 'low'. Besides, 'hoarse', 'resonant' and 'shrill' achieved relatively high agreements, while 'tense' and 'nasal' showed a strong increase after the explanation. We argue that most of the characteristics that show high agreement/improvement are at least partially characterized by their supralaryngeal resonance characteristics ('tense', 'nasal') and/or prosodic features such as loudness ('shrill', 'resonant'). Purely laryngeal features show less agreement. 'Hoarse', a voice feature rather prominent in everyday experience and well represented in our data set, is an exception. Interestingly, it showed a strong decrease in agreement after the explanation. Another laryngeal feature, 'creaky', is also well represented in our data set, but decreases from 'fair' to only 'slight' agreement after the explanation.

Overall, laryngeal and, to a slightly lesser degree, supralaryngeal features seem to be more difficult to detect than prosodic ones. This analysis is difficult to uphold when looking at individual voices, some of which showed a high decrease in agreement despite being characterized by voice features which should be "easy" to classify - or vice versa. This leads us to conclude that tracing individual characteristics of voices may be a different challenge if these appear in combination. However, the complex potential combinations of voice features cannot be easily imitated or straightforwardly explained, not even by an expert, or covered by prototypical example voices. Summing up, a short explanation using prototypical examples and imitations of voice characteristics helps, but does not achieve a high agreement among participants. We next plan to examine whether explanatory success can be further enhanced by state-of-the-art voice conversion technology, which disentangles the relevant dimensions of voice characteristics, and may imitate and recombine these with a higher consistency than expert explainers.

ACKNOWLEDGEMENTS

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824

5. REFERENCES

[1] J. Kreiman and D. Sidtis, "Voices and listeners: Toward a model of voice perception," *Acoustics Today*, vol. 7, no. 4, pp. 7–15, 2011.

[2] W. J. Hardcastle and J. M. Beck, "Forensic speaker identification and the phonetic description of voice quality," in *A Figure of Speech*. Routledge, 2014, pp. 425–452.

[3] S. S. Hammer and A. Teufel-Dietrich, *Stimmtherapie mit Erwachsenen: was Stimmtherapeuten wissen sollten*. Springer-Verlag, 2017.

[4] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[5] J. Kreiman and B. R. Gerratt, "Sources of listener disagreement in voice quality assessment," *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1867–1876, 2000.

[6] A. Webb, P. Carding, I. J. Deary, K. MacKenzie, N. Steen, and J. A. Wilson, "The reliability of three perceptual evaluation scales for dysphonia," *European Archives of Oto-Rhino-Laryngology and Head & Neck*, vol. 261, no. 8, pp. 429–434, 2004.

[7] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[8] J. Laver, S. Wirz, J. Mackenzie, and S. Hiller, "A perceptual protocol for the analysis of vocal profiles," *Edinburgh University Department of Linguistics Work in Progress*, vol. 14, pp. 139–155, 1981.

[9] J. M. Beck, "Vocal profile analysis scheme: A userâs manual," *Queen Margaret University, Edinburgh*, 2007.

[10] E. San Segundo and J. A. Mompean, "A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity," *Journal of Voice*, vol. 31, no. 5, pp. 644–e11, 2017.

[11] M. Hirano, "Objective evaluation of the human voice: clinical aspects," *Folia Phoniatrica et Logopaedica*, vol. 41, no. 2-3, pp. 89–144, 1989.

[12] B. Ann and R. No, "Grbas evaluation of running speech and sustained phonations," *Ann, Bull. RILP No*, vol. 28, pp. 51–56, 1994.

[13] J. Wendler and L. C. Anders, "Hoarse voices-on the reliability of acoustic and auditory classifications," in *FOLIA PHONIATRICA*, vol. 38, no. 5-6. KARGER ALLSCHWILERSTRASSE 10, CH-4009 BASEL, SWITZERLAND, 1986, pp. 369–369.

[14] J. Wendler, W. Seidner, G. Kittel, and U. Eysholdt, "Lehrbuch der Phoniatrie und Pädaudiologie 3. aufl. s 323-364," 1996.

[15] E. San Segundo, P. Foulkes, P. French, P. Harrison, V. Hughes, and C. Kavanagh, "The use of the vocal profile analysis for speaker characterization: Methodological proposals," *Journal of the International Phonetic Association*, vol. 49, no. 3, pp. 353–380, 2019.

[16] T. L. Eadie and C. R. Baylor, "The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice," *Journal of voice*, vol. 20, no. 4, pp. 527–544, 2006.

[17] J. Wiechmann, T. Glarner, F. Rautenberg, P. Wagner, and R. Häb-Umbach, "Technically enabled explaining of voice characteristics," in *P&P 18 Abstract Booklet*. Bielefeld University, Germany, 2022. [Online]. Available: https://doi.org/10.11576/pundp2022-1038

[18] L. F. Gallardo and B. Weiss, "The nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, pp. 363–374, 1977.

[20] M. Gamer, J. Lemon, and I. F. P. Singh, *irr: Various Coefficients of Interrater Reliability and Agreement*, 2019, r package version 0.84.1. [Online]. Available: https://CRAN.R-project.org/package=irr

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[1] The text uses the English translations of the following German original descriptions: knarrend (creaky), heiser (hoarse), behaucht (breathy), geflüstert (whispery), brüchig (cracked), rau (harsh), gepresst (tense), klangvoll (resonant), sanft (soft), dunkel (dark), hell (bright), nasal (nasal), kehlig und/oder rückverlagert (retracted tongue and/or raised larynx), hoch (high), tief (low), laut (loud), leise (quiet), schrill (shrill), starke Lautstärkenvariabilität oder Tonhöhenvariabilität (variable loudness or pitch)