INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Testing acoustic voice quality classification across languages and speech styles

Bettina Braun[1], *Nicole Dehé*[1], *Marieke Einfeldt*[1], *Daniela Wochner*[1], *Katharina Zahner-Ritter*[2]

[1]University of Konstanz, Department of Linguistics, Germany
[2]University of Trier, Department II, Phonetics, Germany

{bettina.braun, nicole.dehe, marieke.einfeldt, daniela.wochner}@uni-konstanz.de,
k.zahner-ritter@uni-trier.de

Abstract

Many studies relate acoustic voice quality measures to perceptual classification. We extend this line of research by training a classifier on a balanced set of perceptually annotated voice quality categories with high inter-rater agreement, and test it on speech samples from a different language and on a different speech style. Annotations were done on continuous speech from different laboratory settings. In Experiment 1, we trained a random forest with Standard Chinese and German recordings labelled as modal, breathy, or glottalized. The model had an accuracy of 78.7% on unseen data from the same sample (most important variables were harmonics-to-noise ratio, cepstral-peak prominence, and H1-A2). This model was then used to classify data from a different language (Icelandic, Experiment 2) and to classify a different speech style (German infant-directed speech (IDS), Experiment 3). Cross-linguistic generalizability was high for Icelandic (78.6% accuracy), but lower for German IDS (71.7% accuracy). Accuracy of recordings of adult-directed speech from the same speakers as in Experiment 3 (77%, Experiment 4) suggests that it is the special speech style of IDS, rather than the recording setting that led to lower performance. Results are discussed in terms of efficiency of coding and generalizability across languages and speech styles.

Index Terms: voice quality, phonation type, acoustic measures, random forest, cross-linguistic generalization, infant-directed speech, German, Chinese, Icelandic

1. Introduction

In a broad sense, voice quality includes glottal (e.g., f0, phonation) and supralaryngeal properties (e.g., nasalization) of a speaker; in a narrow sense, voice quality refers only to the glottal characteristics [1], more precisely termed 'phonation type' or 'phonatory quality' (for a historical overview of descriptions of voice quality, see [2, section 1.3]. In this paper, we focus on phonation type. Phonation is strongly affected by the biological characteristics of the glottis, the length and structure of the vocal folds and their tenseness, as well as muscular tension in the cricothyroid and crico-arythenoid muscles, cf. [1], leading to the perceived long-term characteristics of a person's voice [3, 4]. In addition to indexical information, differences in voice quality can signal lexical contrasts (the difference between (glottalized) /t/ and (non-glottalized) /d/ in some American English varieties [5], phonemic contrasts of vowels in Gujarati [6], cf. [7]), illocution type (breathier voice in rhetorical than in information-seeking questions [8, 9]), irony and sarcasm [10, 11], as well as information-structure [e.g., 12 on Finnish focus marking]. Voice quality also plays a role in signalling emotions [13]; as a social marker for gender identity [14 for English, 15 for Japanese] and often occurs in infant-directed speech (IDS) [16-19].

For the classification of non-pathological voice quality, as e.g., in linguistic and paralinguistic research, voice quality classification often uses the labels modal voice, breathy voice, or glottalized voice. These categories have been introduced in [1], among other categories. [1] describes neutral (modal) voice quality as being derived from periodic and efficient way of vocal fold vibration without audible friction noise. Some researchers locate modal voice in the typical fundamental frequency (f0) range of the speaker; fundamental frequencies that exceed the typical f0-range are called falsetto, those below the typical f0-range creak. Breathy voice is characterized by auditory glottal frication, glottalized voice quality by partly irregular, low frequency vocal fold vibration [1].

The use of perceptual classification in linguistic and paralinguistic question has the advantage that it represents a valid classification with functional value. The flip side is that classification may be circular (the perceived linguistic or emotional category may influence the choice of voice quality) and is more subjective [20]. Unlike in pathological settings or the forensic domain, coders are typically not extensively trained for different phonation types [21]. Acoustic measures, on the contrary, have the advantage that they can be easily extracted from sound recordings (unlike inverse filtering or physiological measures) and they appear to be more objective than perceptual classification (nevertheless, perceptual classification of voice quality may be useful when it comes to the interpretation of linguistic or paralinguistic functions [22]). There are a large number of studies that relate acoustic measures to perceived voice quality, including some meta studies. Measures related to breathy voice are the periodicity in the signal, such as harmonics-to-noise ratio, hnr [23], the amplitude of the first harmonics H1 (i.e., f0), which is measured relative to the amplitude of the second harmonic, H1-H2 [6, 7, 23] or relative to the amplitude of the second or third formant, H1-A2 or H1-A3 [6, 14, 24], and cepstral-peak prominence, cpp [16, 23]. Many of these studies investigate the acoustic characteristics of sustained vowels [25, 26] (which have the advantage that the phonation and fundamental frequency is time-invariant and there are no effects of speech rate, prominence and phrasing), or pathological voice quality [21, 27].

In this paper, our focus is on voice quality from *continuous speech*, collected in various projects at our Lab, in which non-modal voice quality on different constituents of an utterance signaled a linguistic contrast. In Experiment 1, we train a random forest classifier to predict breathy, glottalized, and modal vowels in different position of utterances. We then test the *cross-linguistic generalizability* of the statistical model by applying it to a similar speech style in Icelandic (Experiment 2). Next, we test the same model with speech samples from a *different speech style* (German IDS, Experiment 3) and to adult-

http://dx.doi.org/10.21437/Interspeech.2021-315

directed speech (ADS) from the same speakers (Experiment 4), which allows us to compare lab speech (Exp. 1 and 2) to spontaneous speech (Exp. 3 and 4). Given that biological sex affects voice quality and acoustic correlates [28, 29], we focused on biologically female speakers in this paper.

2. Experiment 1

We used a series of acoustic markers of voice quality to predict three perceptual voice quality labels (breathy, modal, glottalized). Data selection was based on the goal of assembling a data set in which modal and non-modal (breathy and glottalized) phonation type occurred almost equally frequently. The rationale for this decision was that modal voice quality is the default in non-pathological voices and hence more frequent than the non-modal voice qualities breathy and glottalized voice. This imbalanced distribution may lead to models that are particularly tuned to modal voice.

2.1. Methods

2.1.1. Materials

The main data are drawn from [30], which consists of 1737 vowels utterances from 10 female speakers from Beijing. The recordings were done in a quiet room using a headset microphone Shure SM10A and were digitized onto a computer with 44.1 kHz, 16 Bit. Speakers read contexts that prompted either an information-seeking question or a rhetorical question. The speech material consisted of *wh*-questions (e.g., '*Who likes lemons?'*) and polar questions (e.g., '*Does anyone like lemons?'*) and the sentence-final word was balanced for lexical tone. One native-speaker labeler annotated the voice quality of the vowels of the first and last word (50% of the labels were annotated by one of the authors with 95% accuracy (kappa κ = 0.94, almost perfect agreement [31]). In total, 132 vowels were glottalized, 1605 were modal (there were no breathy instances). We hence included 109 breathy-voice vowels (produced by 10 female speakers) from a similar experiment for German [9]. The German recordings were done in a sound-attenuated cabin at the PhonLab of the University of Konstanz with the same recording device as in [30]. The recordings were annotated for voice quality on all content words by one annotator (agreement for 20% of the data by a second annotator was 89.7%, κ = 0.71, which is substantial [31]). Finally, to increase the data set, we included another 64 breathy voice and 191 modal voice vowels of one speaker that was used in a perception experiment [22]. She was recorded in a sound-attenuated room at the PhonLab of the University of Konstanz, using an MXL 990 condenser microphone and a Tascam HDP2 portable stereo audio recorder (44.1 kHz, 16 Bit).

2.1.2. Acoustic measures

We extracted acoustic measures that are frequently used in the literature to distinguish pathological and non-pathological voice qualities. We adapted the algorithms described in VoiceSauce [32] for praat [33] to extract measures that operationalize periodicity, the relative height of the first harmonics H1, spectral tilt and cpp at the center of vowels that were manually labelled for voice quality (40ms window). Prior to analyses all recordings were down-sampled to 16kHz, converted to mono and scaled to a peak amplitude of 0.9. The following measures were extracted:

- Harmonics-to-noise-ratio (hnr), shimmer and jitter (extracted using the Voice Report in praat [34]).
- H1-H2 in dB: difference in amplitude for first and second harmonics [6, 35]. The amplitudes of H1 and H2 were the maximum in the long-term averaged spectrum (standard settings in praat) in the respective frequency range of f0 (H1) and the second harmonics (H2) +/- 10%. F0 was extracted via a praat pitch object with the standard settings (range 75-550Hz). If f0 could not be extracted, the value was set to NA.
- H1-A2 in dB: difference in amplitude for first harmonic and second formant [24]. The second format was extracted via a burg-formant object in praat with the standard settings at the midpoint of the vowel. If f0 could not be extracted, the value was set to NA.
- H1-A3 in dB: difference in amplitude for first harmonic and third formant [35]; the extraction was analogous to H1-A2.
- H1*-A3* in dB: difference in amplitude for first harmonic and the third formant, corrected for formant position following the procedure in [36].
- Cepstral-peak prominence, cpp [16, 37]: The height (i.e., "prominence") of that peak relative to a regression line through the overall cepstrum. The data was multiplied with a Hamming window and then transformed into the cepstral domain. The cpp is the maximum around the quefrency of the pitch period. This peak was then normalized relative to the linear regression line (calculated between 1 ms and the maximum quefrency).
- Glottal to noise excitation ratio, gne [38]: the maximum of a harmonicity object (gne) that was extracted with the standard settings in praat in the 40ms window (rectangular). This measures is often used for the classification of pathologically breathy voice [39].
- b1 and b2 in Hz [40]: bandwidth of the first and second formant (in Hz) at the center of the vowel, extracted from the burg formant object.
- Binary variable if f0 could be measured or not (f0_yesno), and whether shimmer could be measured or not.

Since voice quality may change within the utterance, global measures as in [41] could not be used.

2.1.3. Random forest

To determine the acoustic cues that are most important for the perceptual voice quality classification, we first assembled a balanced data set. From the 1796 modal voice vowels we randomly selected 400 vowels, slightly more than non-modal phonation stimuli (balancing speaker identity, experimental origin and position in the utterance).

Table 1: Number of annotated vowels used for training the classifier, split by voice quality label

	Breathy voice	**Glottalized voice**	**Modal voice**	**total**
Full set	173	132	1796	2101
Full set (balanced)	173	132	400	705
Training set	138	101	325	564
Test set	35	31	75	141

To train the random forest, we randomly selected 80% of the data for training and 20% for test (again balancing the above

factors), see Table 1. The random forest included all the acoustic variables introduced in 2.1.2, using the R-package *randomForest* [42]. The number of trees was set to 1000, mtry (the number of acoustic variables selected at each step was set to 8). Random forests extract the importance of the individual variables using the Gini-index [42].

2.2. Results

Figure 1 shows the importance of the acoustic features (left top corner), A-C displays the most important measures across the three voice quality labels. The Gini-index shows that hnr was ranked very high, followed in importance by cpp and the relative amplitude of the first harmonics in relation to the second formant in dB (H1.A2.dB). The confusion matrix for the test set is shown in Table 2. The overall accuracy was 78.7% (breathy voice 79.3%, glottalized voice 95%, modal voice 75%), κ=0.60.

Table 2: *Confusion matrix of predicted labels (columns) and actual labels (rows) for Experiment 1.*

	breathy voice	**glottalized voice**	**modal voice**
breathy	23	0	6
glottalized	1	19	0
modal	11	12	69

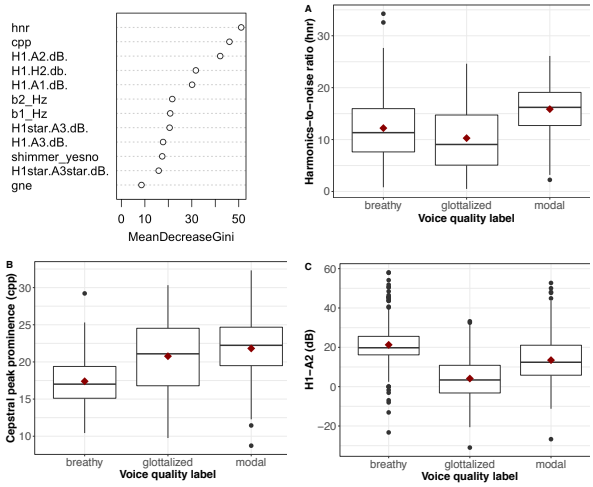

Figure 1: *Results of the random forest for the balanced data set of Experiment 1. Top-right: Mean decrease in Gini, A-C: most important acoustic measures.*

2.3. Discussion

The accuracy of predicting three classes of voice quality labels was nearly 79%. The most important acoustic variable for prediction were hnr, cpp, and H1-A2, which is in line with previous findings. For instance, [23] found the highest correlation with cpp. The variance explained increased to 94% with further acoustic measures (including a breathiness index and the amplitude of first harmonic). [14] report a correlation of 83% between first harmonic amplitude and breathiness ratings, [23] a correlation of 66% for the same measure, H1-H2 (they attribute their lower value to the higher proportion of male speakers in their sample).

Note that the model did not contain speaker identity, information on the vowel quality or language, and was trained on continuous speech data with different speaking rates and intonation contours. Information on these factors may likely improve the model fit, but at the same time limit the generalizability of the model.

3. Experiment 2

Exp. 2 tested whether the acoustic model from Exp. 1 can be applied to the classification of voice quality of Icelandic question data (similar linguistic structure, but a different language and recording environment). Icelandic is a Germanic intonation language, which, unlike German, exhibits word-initial stress with very few exceptions and final falling contours across utterance types [43].

3.1. Methods

3.1.1. Data

The data were drawn from [44]. Eleven female speakers were recorded in a sound-attenuated room (with the same materials and recording device as for the German and Chinese speakers in Exp. 1). Similar to the German data of Exp. 1, voice quality was annotated on the stressed syllable of the three content words. We extracted acoustic measures from 1243 vowels, 23 with breathy voice quality (1.9%), 114 with glottalized voice (9.2%), and 1106 with modal voice quality (89%). Interrater reliability for 82 vowels (7% of the data) resulted in an agreement of 90% (κ = 0.71).

3.1.2. Analysis

We extracted the same acoustic measures as in Exp. 1. We then used the model from Exp. 1 to predict the voice quality labels.

3.2. Results

Table 3 shows the confusion matrix for Exp. 2. The overall accuracy of the prediction was 78.6% overall (69.6% for breathy voice, 44% for glottalized voice and 82.4% for modal voice stimuli), κ=0.29.

Table 3: *Confusion matrix for Experiment 2.*

	breathy voice	**glottalized voice**	**modal voice**
breathy	16	0	7
glottalized	22	51	42
modal	133	62	910

3.3. Discussion

The model trained on German question data generalized well to the Icelandic question data, as indicated by an equally high accuracy (78.8% in Exp. 1 vs. 78.6% in Exp. 2). Perceptual evaluation of a sample of misclassifications suggests that the 40ms analysis window is sometimes too large (including neighboring voiceless sounds, leading to misclassification as breathy) or too small (missing vowel-initial glottalization, leading to misclassification as modal).

4. Experiment 3

Exp. 3 tested the applicability of the model in Exp.1 to German IDS. IDS has been argued to be more breathy and variable than ADS [16, 17], and hence might be more challenging to predict than ADS read speech.

4.1. Methods

The German IDS data was collected and annotated in [18, 19]. It contained 449 utterances from eleven mothers interacting with their children. They were recorded in a quiet room at the University of Konstanz. The speakers wore DPA Microphones Headset Microphone (DPA 4088), connected to a Sennheiser Bodypack Transmitter (SK100) and an Olympus Linear PCM Recorder (LS-11). Recordings were in a 16 bit format at 44.1 kHz. Speakers opened a treasure box and extracted one of eight objects at a time and talked about them to their two-year old children. The stressed vowel of the target objects and of the vowels of the word preceding and following the target object were annotated for voice quality. In total, there were 996 vowels, for which acoustic measures could be extracted, 689 with modal voice (69.2%), 151 with breathy voice (15.1%) and 156 with glottalized vowels (15.7%). Interrater agreement for the total set of 1028 vowels was 81.0% ($\kappa = 0.59$). The analysis was the same as for Exp. 1 and 2.

4.2. Results

Table 4 shows the confusion matrix for Exp. 3. Overall accuracy was 71.7% (40% for breathy, 37% for glottalized and 87% for modal voice), κ=0.23. As in Exp. 2, the less frequent non-modal voice qualities had lower classification rates than the more frequent modal voice.

Table 4: *Confusion matrix for Experiment 3.*

	breathy voice	**glottalized voice**	**modal voice**
breathy	60	8	83
glottalized	11	57	88
modal	76	16	597

4.3. Discussion

IDS was classified less well with the laboratory (read speech) model of Exp. 1 than the Icelandic question data (71.7% vs. 78.6% accuracy), despite the fact that the model in Exp. 1 was partly trained on German. The lower classification may be due to the different communicative style (ADS in Exp. 1 and 2 vs. IDS in Exp. 3), due to the difference in spontaneity (read speech in Exp. 1 and 2 vs. spontaneous speech in Exp. 3) or due to the difference in recording setting (sound attenuated cabin in Experiments 1 and 2 vs. a quiet room in Exp. 3). To investigate these issues, we tested spontaneous ADS data from the same speakers in the same room as in Exp. 3 (recorded in the same experimental session).

5. Experiment 4

5.1. Methods

Exp. 4 used 433 vowels of continuous speech in which the speakers of Exp. 3 talked to an adult (experimenter, child not present). There were 33 breathy (7.7%), 81 glottalized (18.8%) and 316 modal voice stimuli (73.5%). Interrater agreement for 78 vowels was 96.2% ($\kappa = 0.90$).

5.2. Results and discussion

The overall accuracy was 77.2% (κ=0.26, see Table 5) and was hence comparable to Exp. 1 and 2. This suggests that the low accuracy in Exp. 3 is likely due to the special speech style used in IDS, rather than the spontaneity or the recording setting.

Table 5: *Confusion matrix for Experiment 4.*

	breathy voice	**glottalized voice**	**modal voice**
breathy	14	3	16
glottalized	7	46	28
modal	30	14	272

6. General Discussion

This paper presents a random forest model to map acoustic data to perceptual voice quality classifications. This is important as voice quality does not only signal indexical information, but also linguistic contrasts. The model had an accuracy of nearly 80% on an unseen sample of data from the trained population (German and Chinese question data) and it transferred well to Icelandic question data and to German spontaneous ADS. The accuracy of automatic classification was a bit lower for German IDS. There are a number of potential factors that may explain the differences in accuracy (and kappa) across experiments, such as the unbalanced distribution of the different voice quality labels or the differences in interrater accuracy and reliability (kappa). Post-hoc correlation analyses did not show any significant correlations between these factors (all $p > 0.2$). Thus, for the current data, neither human classification accuracy (or reliability) nor the proportion of modally voiced vowels seem to affect machine classification accuracy and reliability. The lack of a correlation with the proportion of modally voiced stimuli is positive, since a skewed distribution can hardly be avoided in natural data. We successfully avoided distributional effects by using a balanced set of voice quality labels for training.

It is noteworthy that voice quality labels in IDS were hardest to predict. It is conceivable that the higher and more variable f0 and vowel formants in IDS compared to ADS [46, 47] affected our acoustic measures [48] which, in turn would suggest a register-specific acoustic profile of voice quality.

Further research with different languages and speech styles is necessary to investigate whether the model generalizes more broadly or where its limits are. Furthermore, we will look more closely into the misclassifications to detect potential patterns and issues (cf. the window size, the perception of voice quality in different consonantal environments), which will allow to improve accuracy and kappa across corpora, in particular for less laboratory speech styles.

7. Conclusions

We presented a classifier, trained on a small set of perceptually annotated voice quality labels from German and Chinese continuous, scripted speech, which generalized well to another language (Icelandic) and to more spontaneous productions in German, despite different recording settings. Generalization to another speech style, German IDS, was harder and further research will investigate on how to improve generalization across different speech styles.

8. Acknowledgements

The project was funded by the German Research Foundation (DFG): BR 3428/4-1/2 and DE 876/3-1/2. We thank Tianyi Zhao, Justin Hofenbitzer, Johanna Schnell, Friederike Hohl, Naomi Reichmann and Elena Schweizer for help with literature research and/or for annotating data. Furthermore, we thank Carmen Widera on discussion on the statistical analysis.

9. References

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

[2] J.H. Esling, et al., "*Voice Quality. The Laryngeal Articulator Model"*. Cambridge, UK: Cambridge University Press, 2019.

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

[4] D. Abercrombie, "*Elements of General Phonetics"*. Edinburgh: Edinburgh University Press, 1967.

[5] M. Garellek, "*Perception of glottalization and phrase-final creak"*. The Journal of the Acoustical Society of America, 137(2): p. 822-831. 2015.

[6] E. Fischer-Jorgensen, "*Phonetic analysis of breathy (murmured) vowels in Gujarati"*. Indian Linguistics, 28: p. 71-139. 1967.

[7] P. Ladefoged, "*The linguistic use of different phonation types"*, in *Vocal fold physiology: Contemporary research and clinical issues*, D.M. Bless and J.H. Abbs, [Eds], College Hill: San Diego. p. 351-360, 1983.

[8] N. Dehé and B. Braun, "*The prosody of rhetorical questions in English"*. English Lang. and Linguistics, 24(4): p. 607-635. 2019.

[9] B. Braun, et al., "*The prosody of rhetorical and information-seeking questions in German"*. Language and Speech, 62(4): p. 779–807. 2019.

[10] O. Niebuhr. "*"A little more ironic" Voice quality and segmental reduction differences between sarcastic and neutral utterances"*. in *7th International Conference on Speech Prosody 2014*. Dublin, Ireland, 2014.

[11] S. Li, et al., "*The role of voice quality in Mandarin sarcastic speech: An acoustic and electroglottographic study"*. Journal of Speech, Language, and Hearing Research. 2020.

[12] A. Arnhold, "*Complex prosodic focus marking in Finnish: expanding the data landscape"*. J. Phonetics, 56: p. 85-109. 2016.

[13] C. Gobl and A. Ní Chasaide, "*The role of voice quality in communicating emotion, mood and attitude"*. Speech Communication, 40: p. 189-212. 2003.

[14] D.H. Klatt and L.C. Klatt, "*Analysis, synthesis, and perception of voice quality variations among female and male talkers"*. The Journal of the Acoustical Society of America, 87: p. 820-857. 1990.

[15] R.L. Starr, "*Sweet voice: The role of voice quality in a Japanese feminine style"*. Language in Society, 44: p. 1-34. 2015.

[16] K. Miyazawa, et al., "*Vowels in infant-directed speech: More breathy and more variable, but not clearer"*. Cognition, 166: p. 84-93. 2017.

[17] T. Benders, E. Tobin, and A. Szakay. "*Infant-Directed Speech may not be across-the-board breathy, but has a variable voice quality"*. in *17th Australasian International Conference on Speech Science and Technology*. Sydney, Australia, 2018.

[18] K. Busse, "Are there differnences between German IDS and ADS compared to Japanese", BA-thesis, Linguistics, Konstanz, 2019

[19] S. Willenborg, "Prosodische Aspekte der Eltern-Kind-Interaktion: Stimmqualität und ihre Variabilität", BA-thesis, Linguistics, Konstanz, 2020

[20] J. Kreiman, D. Vanlancker-Siditis, and B.R. Gerrat. "*Defining and measuring voice quality"*. in *Voice Quality: Functions, Analysis and Synthesis*. Geneva, Switzerland, 2003.

[21] B. Barsties, et al., "*The effect of visual feedback and training in auditory-perceptual judgment of voice quality"*. Logopedics Phoniatrics Vocology. 2015.

[22] M. Kharaman, et al. "*The processing of prosodic cues to rhetorical question interpretation: Psycholinguistic and neurolinguistics evidence"*. in *Proceedings of Interspeech*. Graz, Austria, 2019.

[23] J. Hillenbrand, R.A. Cleveland, and R.L. Erickson, "*Acoustic correlates of breathy voice quality"*. Journal of Speech and Hearing Research, 37(as): p. 769-778 1994.

[24] M. Garellek and P.A. Keating, "*The acoustic consequences of tone and phonation interaction in Jalapa Mazatec"*. Journal of the International Phonetic Association, 41: p. 185-205. 2011.

[25] R.P. Clapham, et al., "*The Relationship Between Acoustic Signal Typing and Perceptual Evaluation of Tracheoesophageal Voice Quality for Sustained Vowels"*. Journal of Voice, 29(4): p. 517.e23-517.e29. 2015.

[26] A.G. Askenfelt and B. Hammarberg, "*Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures"*. J. Speech & Hearing Res., 29: p. 50-64. 1986.

[27] Y. Maryn, et al., "*Acoustic measurement of overall voice quality: A meta-analysis"*. Journal of Acoustic Society of America, 126(5): p. 2619–2634. 2009.

[28] T. Jayakumar, J.J. Benoy, and M. Yasin, "*Effect of age and gender on acoustic voice quality index across lifespan: A cross-sectional study in Indian population"*. Journal of Voice. in press.

[29] M. Iseli, Y.-L. Shue, and A. Alwan, "*Age, sex, and vowel dependencies of acoustic measures related to the voice source"*. The Journal of the Acoustical Society of America, 121(4): p. 2283–2295. 2007.

[30] K. Zahner, et al. "*The prosodic marking of rhetorical questions in Standard Chinese"*. in *10th International Conference on Speech Prosody*. Tokyo, Japan, 2020.

[31] J.R. Landis and G.G. Koch, "*The measurement of observer agreement for categorical data"*. Biometrics, 33: p. 159-174. 1977.

[32] Y.-L. Shue, et al., "*VoiceSauce: A program for voice analysis"*. The Journal of the Acoustical Society of America, 126(4). 2009.

[33] P. Boersma and D. Weenink, "*Praat: doing phonetics by computer"*: http://www.praat.org/, retrieved 11 May 2018, 2018

[34] P. Boersma, "*Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ration of a sampled sound"*. Proceedings of the Institute of Phonetic Sciences, University of Amsterdam, 17: p. 97-110. 1993.

[35] E.B. Holmberg, et al., "*Comparison among aerodynamic, electroglottographic, and acoustic spectral measures of female voice"*. J. Speech and Hearing Research, 38: p. 1212-1223. 1995.

[36] C. Mooshammer, "*Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German"*. Journal of Acoustic Society of America, 127(2): p. 1047-1058. 2010.

[37] J.M. Hillenbrand, "*Some effects of intonation contour on sentence intelligibility"*. Journal of the Acoustical Society of America Journal, 114: p. 2338 –2338. 2003.

[38] M. Frohlich, D. Michaelis, and W. Strube. "*Acoustic "breathiness measures" in the description of pathological voices"*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seattle, WA, USA, 1998.

[39] J.I. Godino-Llorente, et al., "*The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders"*. Journal of Voice, 24(1): p. 47-56. 2010.

[40] F. Burkhardt. "*Rule-based voice quality variation with formant synthesis"*. in *Interspeech*. Brighton, UK, 2009.

[41] B.B.v. Latoszek, B. Lehnert, and B. Janotte, "*Validation of the acoustic voice quality index version 03.01 and acoustic breathiness index in German"*. Journal of Voice, 34(1): p. 157.e17-157.e25. 2020.

[42] A. Liaw and M. Wiener, "*Classification and Regression by randomForest"*. R News, 2(3): p. 18-22. 2002.

[43] K. Árnason, "*The phonology of Icelandic and Faroese"*. Oxford: Oxford University Press, 2011.

[44] N. Dehé and B. Braun, "*The intonation of information-seeking and rhetorical questions in Icelandic"*. Journal of Germanic Linguistics, 32(1): p. 1 - 42. 2020.

[45] P.A. Keating, M. Garellek, and J. Kreiman. "*Acoustic properties of different kinds of creaky voice"*. in *18th International Congress of the Phonetic Sciences*. Glasgow, UK, 2015.

[46] T. Benders, "*Mommy is only happy! Dutch mothers' realization of speech sounds in infant-directed speech expresses emotion, not didactic intent"*. Infant Behav and Dev., 36(4): p. 847-862. 2013.

[47] M. Kalashnikova, C. Carignan, and D. Burnham, "*The origins of babytalk: smiling, teaching or social convergence"*. Royal Society Open Science, 4: p. 170306. 2017.

[48] J.J. Kuang, "*Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice"*. Journal of Acoustic Society of America, 142. 2017.