Classification of voice quality using neck-surface acceleration: Comparison with glottal flow and radiated sound

[*]Marcin Włodarczak, [†]Bogdan Ludusan, [‡]Johan Sundberg, and [*]Mattias Heldner, *‡*Stockholm, Sweden, and* †*Bielefeld, Germany*

Summary: Objectives. The aim of the present study is to investigate the usefulness of features extracted from miniature accelerometers attached to speaker's tracheal wall below the glottis for classification of phonation type. The performance of the accelerometer features is evaluated relative to features obtained from inverse filtered and radiated sound. While the former is a good proxy for the voice source, obtaining robust voice source features from the latter is considered difficult since it also contains information about the vocal tract filter. By contrast, the accelerometer signal is largely unaffected by the vocal tract and although it is shaped by subglottal resonances and the transfer properties of the neck tissue, these properties remain constant within a speaker. For this reason, we expect it to provide a better approximation of the voice source than the raw audio. We also investigate which aspects of the voice source are derivable from the accelerometer and microphone signals.

Methods. Five trained singers (two females and three males) were recorded producing the syllable [pæ:] in three voice qualities (neutral, breathy and pressed) and at three pitch levels as determined by the participants' personal preference. Features extracted from the three signals were used for classification of phonation type using a random forest classifier. In addition, accelerometer and microphone features with highest correlation with the voice source features were identified.

Results. The three signals showed comparable classification error rates, with considerable differences across speakers both with respect to the overall performance and the importance of individual features. The speaker-specific differences notwithstanding, variation of phonation type had consistent effects on the voice source, accelerometer and audio signals. With regard to the voice source, AQ, NAQ, L_1L_2 and CQ all showed a monotonic variation along the breathy − neutral − pressed continuum. Several features were also found to vary systematically in the accelerometer and audio signals: HRF, L_1L_2 and CPPS (both the accelerometer and the audio), as well as the sound level (for the audio). The random forest analysis revealed that all of these features were also among the most important for the classification of voice quality.

Conclusion. Both the accelerometer and the audio signals were found to discriminate between phonation types with an accuracy approaching that of the voice source. Thus, the accelerometer signal, which is largely uncontaminated by vocal tract resonances, offered no advantage over the signal collected with a normal microphone.

Key Words: phonation type classification−voice source−accelerometer−audio.

INTRODUCTION

As noted by Childers et al.,[1] voice quality is a notoriously elusive term with definitions ranging from speaker-specific long-term timbre of voice to a wider range of periodic oscillation produced by the "laryngeal constrictor".[2] The present work is concerned with short-term variation in voice characteristics attributable to phonation type. Conceived of in this way, voice quality is traditionally represented along a continuum ranging from breathy to neutral to pressed, corresponding to increasing vocal fold adduction and, consequently, increasingly constricted glottis. This, in turn, affects the glottal airflow (i.e. the voice source) and requires adjustment of subglottal pressure, such that when glottal adduction is increased, a stronger force is needed to keep vocal folds vibrating.

However, while varying voice quality involves modifications to the voice source, obtaining a robust estimate of the associated glottal flow waveform is by no means a solved problem, making studies of phonation type difficult, especially in running speech. Manual inverse filtering, which is a common method for analyzing the voice source,[3] is very time consuming and thus of limited use for analysis of large data sets. For this reason, a wide range of methods for capturing phonation characteristics have been proposed in the literature (see below). A particularly promising technique involves a miniature accelerometer attached to the tracheal wall, which captures surface vibrations corresponding to subglottal pressure variation. In the present investigation we compare voice source features, derived by inverse filtering the radiated sound, with the acoustic properties of the accelerometer signal from the tracheal wall as well as with those of the radiated audio signal captured with a normal microphone. Specifically, we examine voice quality in diminuendo sequences of syllable [pæ:] produced at three self-selected pitch levels (habitual, low and high) by trained

Accepted for publication June 30, 2022.
Declarations of interests: none.
From the *Department of Linguistics, Stockholm University, Sweden; †Faculty of Linguistics and Literary Studies, Bielefeld University, Germany; and the ‡Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden.
Address correspondence and reprint requests to Marcin Włodarczak, Department of Linguistics, Stockholm University, 106 91 Stockholm, Sweden. E-mail: wlodarczak@ling.su.se
Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2022.06.034

singers. Features extracted from the three signals were used in supervised (Random Forest) classification experiments. These were followed by a correlation analysis to establish which aspects of the voice source can be approximated with the microphone and accelerometer signals.

Previous work

Voice quality plays an important role in human communication. In addition to marking phonemic contrasts [4,5], voice quality contains information related to the speaker's vocal health[6,7] and it adds to the affective content of what is being said.[8–10] Voice quality dynamics is also involved in phenomena characteristic of spontaneous speech, such as disfluencies[11] and speaker convergence.[12] In addition, voice quality has been increasingly regarded as a prosodic feature, related to such suprasegmental phenomena as declination,[13] prosodic boundaries[14] and prosodic prominence.[15]

Given the relevance of voice quality in such diverse communicative and clinical contexts, much research has been done on automatic and instrumental discrimination between voice quality categories. Work on this topic has employed a wide range of measures across a multitude of domains, including temporal,[16] spectral[17–19] and cepstral features.[20–22] A phonation type which has attracted particular attention over the years is creaky voice (vocal fry), often studied in a clinical context.[23,24] Given that phonatory characteristics are difficult to infer from the acoustic signal, which includes resonances of the vocal tract, much work has been done on using techniques more closely related to the voice source, most particularly electroglottography (EGG),[25] and inverse filtering,[26,27] often done automatically.[28,29]

In an attempt to consolidate these diverse results, Borsky et al.[30] used mel frequency cepstral coefficients (MFCCs) to distinguish between neutral, breathy, strained (pressed) and rough (creaky) voice quality across the acoustic signal, the (inverse-filtered) voice source and the EGG signal. They found that static MFCCs and their first derivative performed best with the acoustic signal, followed by the voice source and EGG. The addition of the second derivative offered little to no improvement. Moreover, for the acoustic signal the study compared the accuracy obtained with MFCCs against three feature sets extracted with COVAREP,[31] related to glottal source, spectral envelope and harmonic model phase distortion. Two of these (harmonic model phase distortion features and glottal source features) outperformed the MFCC-based models. Finally, while combination of different signals improved classification accuracy considerably, especially for the acoustic signal, combining COVAREP features and MFCC offered no additional benefit over using the former set alone.

A more direct, while still unobtrusive, method of gaining information on the voice source is using a miniature accelerometer placed on the neck surface below the glottis,[32] which picks up skin vibrations caused by the air pressure oscillation in the trachea. Previous work has found a number of diverse applications of the accelerometer signal. Not surprisingly, given the absence of high frequency information, the accelerometer signal was found to be useful for measurement of fundamental frequency (f_o).[33,34] Even though SPL measurements based on the accelerometer signal were originally considered inaccurate,[35] more recent work has found that, given speaker-specific calibration, mean sound pressure level (SPL) can be estimated with accuracy of at least 2.8 dB.[36] Perhaps more interestingly, an even stronger link was found between the level of the accelerometer signal and subglottal pressure.[37] For some speakers, however, the relationship seems to vary with vocal intensity, with the accelerometer level being less sensitive to vocal effort in soft voice.[38] In addition, the accelerometer was found to be useful for estimating other characteristics of speech, such as nasality[32,39,40] or the relative strength of the fundamental.[41]

More pertinent to the aims of the present study, the accelerometer signal was also found to be suitable for classification of phonation type in sustained vowels.[22] Specifically, MFCCs calculated from the accelerometer waveforms discriminated between neutral, breathy, pressed and rough voice qualities with the accuracy of 80.2% and 89.5% at the word and utterance levels, respectively. Interestingly, the 0^{th} MFCC coefficient, which carries information about the overall spectral energy, achieved the accuracy of 60.7% on the word level and 68.4% on the utterance level on its own. While this suggests that signal level is particularly informative about phonation type, similar accuracy was achieved by increasing the number of MFCCs used (18 instead of 16) and removing the 0^{th} coefficients

These properties of the accelerometers, coupled with their unobtrusiveness, insensitivity to environmental noise as well as their privacy-respecting characteristics make them a particularly promising tool for continuous monitoring of vocal production in speakers with laryngeal dysfunctions or patients recovering after phonosurgeries.[42,43] For instance, Ghassemi et al.[44] demonstrated that f_o and SPL levels collected from the accelerometer signal can be used to automatically detect hypertension in patients with vocal fold nodules (but cf. Gelzinis et al.,[45] who found that a wide class of laryngeal disorders such as mass lesions of the vocal folds and paralysis can be accurately identified using the acoustic signal with the accelerometers offering no further gain).

Contribution of present work

The present work sets out to evaluate the usefulness of the accelerometer signal for classification of phonation type. Unlike other studies, we compare acoustic properties of the accelerometer signal with those of manually inverse filtered speech as well as raw microphone (audio) signal. In addition, rather than relying on measures whose relationship with voice source characteristics is difficult to establish (such as MFCCs), we use widely established acoustic features with known properties and well understood links to the voice source. To establish which aspects of the voice source can be accounted for by the accelerometer and the

audio signals, we combine automatic classification with a correlation analysis of features extracted from the three signals. Last but not least, we use trained singers, which allows us to evaluate independent contribution of f_o and SPL to voice quality variation, which is difficult to ascertain with untrained speakers.

MATERIAL AND METHODS

Material

Five speakers (two females and three males, mean age = 38, standard deviation = 6) participated in the recording. All participants were trained singers, experienced in the classical tradition. They were recorded using a head-mounted omnidirectional microphone (Sennheiser HSP2) positioned 6 cm from the mouth, and an accelerometer (Knowles BU-27135) attached to the tracheal wall below the cricoid cartilage with an adhesive disk.

The microphone levels were calibrated in dB SPL using a sound level meter. In addition, subglottal pressure (P_{sub}) was estimated from the oral pressure during the occlusion for the consonant /p/ by means of a plastic tube held in the corner of the mouth. The tube was connected to the Subglottal Pressure Monitor PG-60 (Glottal Enterprises), routed to a separate channel of the recording unit. For calibration the tube end was immersed into a glass of water at a measured depth under the water level which was announced in the recording.[1]

All signals were digitized using an Expert Sleepers ES-9 audio interface (48 kHz, 24 bit) with AC coupling on the microphone and accelerometer channels and DC coupling on the subglottal pressure channel.

The participants were instructed to produce diminuendo sequences of the syllable [pæ:] at three self-selected pitch levels (habitual, low and high) and with three phonation types (neutral, pressed and breathy). For each combination of pitch level and phonation type, three repetitions of a sequence were collected. A real-time display of intraoral pressure was monitored to ensure sufficient signal quality, otherwise the participants were asked to repeat the whole sequence. An overview of each participant's f_o, SPL and P_{sub} values is provided in Table 2 in the Appendix. In addition, averaged f_o values for each speaker, pitch condition and voice quality are presented graphically in Figure 8 in the Appendix.

The speakers were informed about the research aims and their written informed consent was obtained. For their participation they were rewarded with two cinema tickets. The recordings took place in a sound treated room in the Phonetics Laboratory at Stockholm University, Sweden.

[1]The accuracy of the pressure transducer was checked by immersing the end of the tube attached to the transducer at depths between 0 and 15 cm H_20. All readings agreed with the actual depth determined by means of a ruler. It was concluded that zero pressure and one more pressure lying within the typical range of phonatory subglottal pressures were enough for calibrating the pressure transducer.

Inverse filtering

The voice source was analyzed in terms of the waveform of the glottal airflow, or the flow glottogram, obtained by inverse filtering the audio signal, using the *Sopran* software.[46] A quasi-stationary portion of the vowel was selected and the *Inverse filter* option, which displays the waveform and the narrow-band spectrum in separate windows, was applied. The frequencies and bandwidths of the inverse filters were tuned manually by applying two criteria: (i) realistic formant frequencies, formant bandwidths and number of inverse filters (one per 1000 Hz) and (ii) minimization of flow ripple during the closed phase. When the inverse filters have been tuned, the resulting flow waveform, i.e., the flow glottogram and its derivative were saved to a stereo wave file.

Feature extraction

For analyzing the properties of flow glottogram the *Glottal flow parameter measurement* option of Sopran was used. After selecting one period, the following measures were obtained:

Fundamental frequency (f_o)
AC flow (AC_F): the difference between the maximum and the minimum of the glottal flow waveform.
Maximum flow declination rate (MFDR): the maximum negative slope of the glottal flow airflow waveform.
Amplitude quotient (AQ): the ratio between peak-to-peak flow amplitude and maximum flow declination rate (AC_F/MFDR).
Normalized amplitude quotient (NAQ): the ratio between amplitude quotient and period (AQ/T).
***L_1L_2*:** the level of the fundamental relative to the level of the second harmonic.
Closed quotient (CQ): the ratio between the closed phase duration and period.
Skewing quotient (SQ): the ratio between the opening and closing phase durations.

In addition, the corresponding level of the voice source spectrum fundamental (L_1) was measured by means of the *Spectrum* option of the Sopran software, applying a 30 Hz analysis bandwidth.

For the accelerometer and microphone signals, the following measures of voice quality were extracted using a custom Praat[47] script:

Smoothed cepstral peak prominence (CPPS): the amplitude of the first rahmonic relative to the regression line across the real cepstrum of the signal.[20]
Alpha ratio: the ratio of acoustic energy in the high (1-5 kHz) and the low (0-1 kHz) frequency bands: E_H/E_L, in dB.
***L_1*:** The level of the fundamental.
***L_1L_2*:** The level of the fundamental relative to the level of the second harmonic: $L_1 - L_2$.

Harmonic richness factor (HRF): The amplitude of the fundamental relative to the summed amplitudes of A_2 to A_{10}: $(A_2 + \ldots + A_{10})/A_1$, in dB[48].

In addition, f_o and the overall sound level (SL) were also extracted. The features were calculated over a 50-ms analysis window at the same time points as those used in the glottal flow analysis.

Data analysis

We analyzed our data by means of a supervised machine learning approach, Random Forest. The goal of this analysis is two-fold: First, to determine how well the three investigated voice quality types may be distinguished on the basis of the extracted features and second, to establish which of the different acoustic features discriminate better the considered voice quality classes.

Random Forest[49] uses the features and the class labels given in input to build a number of decision trees. The decision trees making up the Random Forest are fitted based on different randomly sampled subsets of the input data and individual data points are categorized as members of a particular class by means of majority voting. Unlike other popular analysis methods, such as regression, Random Forrest is largely unaffected by colinearity between independent variables and is thus particularly well suited to voice analysis, where correlated parameters are common. We ran three sets of experiments, each one using input features extracted from the different signals: the voice source signal, the accelerometer signal, and the microphone signal. In each experiment, a model was trained to discriminate between pairs of voice quality conditions: neutral-breathy and neutral-pressed, respectively. We also considered the breathy-pressed case as a control. Since this condition involves more substantial adjustments of phonatory settings, it was expected to result in higher classification accuracy compared to the contrasts including modal voice. All experiments were run on a per-speaker basis, by building a Random Forest consisting of 500 trees, using the R[50] library randomForest.[51]

The out-of-bag (OOB) error returned by the Random Forest classifier was employed for measuring the discrimination quality. The OOB error is determined by predicting the values of the input samples using the decision trees which were not fitted with that data. It is defined in Equation 1, based on the correctness (percentage of correct predictions out of the total number of samples predicted) of the classification, with a better discrimination being represented by a lower error value. For example, in a classification task with a 10% OOB error, 1 in 10 data points is classified as belonging to the wrong class, while the rest are assigned to the correct class. Since the focus of this study is comparing the usefulness of several types of signals for the classification of phonation type, we are interested in the relative performance (e.g., by comparing voice source and accelerometer characteristics) rather than the absolute performance of the classifier.

$$OOB\_error = 100 - correctness \quad (1)$$

We have chosen Random Forest for the analysis because the model can also return the importance of each feature used for learning, giving us insights into which features are the most useful in the discrimination of voice quality types. The importance of the features was defined as the total decrease in node impurities (measured by the Gini index) when splitting on that particular feature, averaged over all trees. It describes how informative a feature is for discriminating between phonation types, with more discriminative feature being assigned a higher importance score. Below, we illustrate the results (OOB error and importance of each feature) obtained for each speaker, individually, as well as an overall measure, computed as the average across our five speakers. To allow for easier comparison of importance scores across conditions, the per-speaker importance of each feature was normalized by dividing it by the sum of importance values of all features considered in that condition. Thus, the sum of importance scores of all features in each condition is equal to 1.

RESULTS

Classification experiments

We first investigate the classification performance obtained with the Random Forest model trained on the voice source-based features to discriminate the two considered conditions: neutral-breathy (NB) and neutral-pressed (NP). The results obtained are illustrated in Figure 1. One may note important individual variation, especially in the neutral-pressed condition, where the performance varies between less than 2%, for speaker M1, to more than 17%, for speaker M3. When compared to the control case (breathy-pressed, BP, not shown in Figure 1), these former two conditions exhibit an overall higher error rate than the latter one, in which the two voice quality types are more distinct from each other (OOB error ranging between 0 and 9.3%, with an average value of 5.7%).

Next, we looked at the mean error rates obtained in the two conditions when using features extracted from the three different signals considered here (see Figure 2). In general, none of the signals showed a clear advantage over the others. Testing the differences with Wilcoxon signed rank tests revealed that none of them are significant. The control case, the breathy-pressed discrimination, showed a lower error than the other two conditions, for all three signals (5.7%, 5.6% and 3.8% for the voice source, the accelerometer and the audio, respectively), but these differences were not found to be significant either.

Figures 3 and 4 display the importance of the voice source extracted features for discriminating between neutral and non-neutral phonation types, on a per-speaker basis. For the neutral-breathy discrimination (Figure 3) the algorithm seems to give rather different weights to the features, depending on the speaker that uttered the respective vowels.

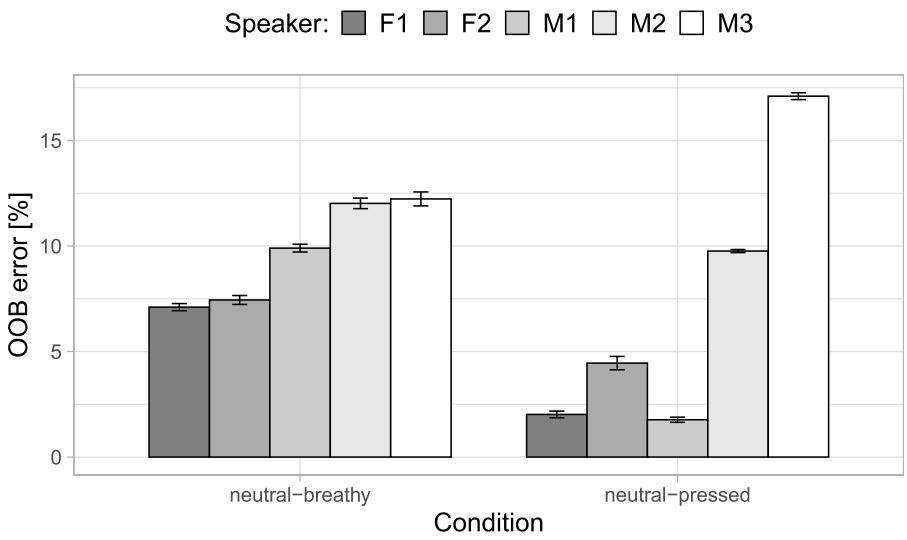

FIGURE 1. The OOB error rate for each of the five speakers in our corpus (F1-F2, M1-M3), in the two considered conditions (neutral-breathy and neutral-pressed), when using the features extracted from the voice source signal. The error bars represent the 95% confidence intervals.

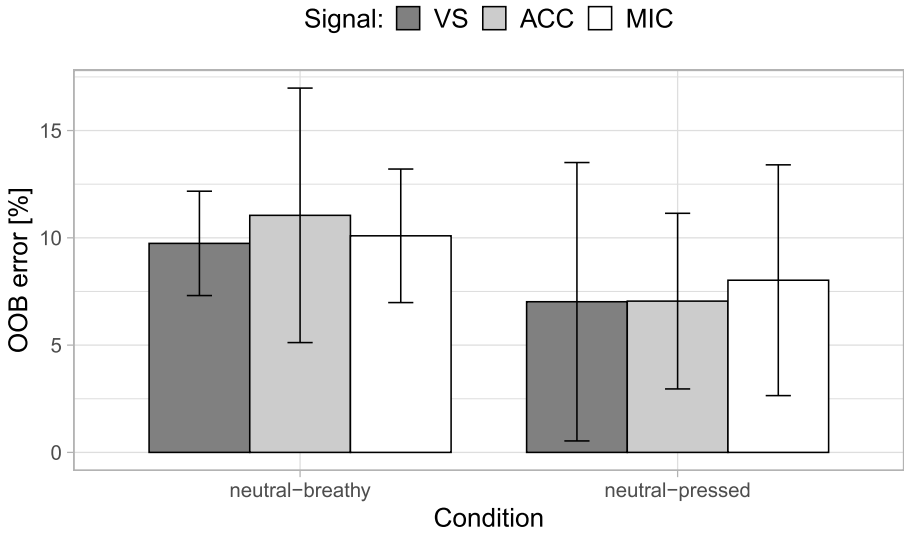

FIGURE 2. The mean OOB error rate across the five speakers in our corpus (F1-F2, M1-M3), for each of the two considered conditions (neutral-breathy and neutral-pressed) and three input signals (the voice source VS, accelerometer ACC, microphone MIC). The error bars represent the standard deviation across speakers.

Speaker F2 marks this distinction with changes that are mostly captured by three features (AQ, L_1L_2, CQ), while for the other speakers a more uniform distribution of the importance was observed. However, for each of them the algorithm finds one or two features which seem to be more helpful for the classification (e.g., AC_F and MFDR for F1, AC_F and NAQ for M1). Fundamental frequency is a highly discriminating feature for three speakers (F1, M2 and M3).

Similar to the neutral − breathy condition, the algorithm shows important differences in the ranking of the features for discriminating between neutral and pressed voice quality (Figure 4). For speakers F1 and M1, AQ and L_1, respectively, are assigned a much higher importance than the rest of the features. Again, three speakers (F2, M2 and M3) vary their f_o level consistently between the two voice qualities. In terms of other features exhibiting a high importance, we observe L_1L_2 for F1, M1 and M2, L_1 for speaker F2, SQ for M2 and AC_F for M3.

Lastly, we investigated the importance of each feature extracted from the three types of signals, averaged across speakers. The left panel of Figure 5 shows the ranking of the voice source-based features. It appears that f_o and L_1L_2 are highly ranked in both conditions, with other features being important for some conditions: AC_F for NB, AQ and L_1 for NP. Also in the case of features extracted from the accelerometer and audio signals (Figure 5 middle and right

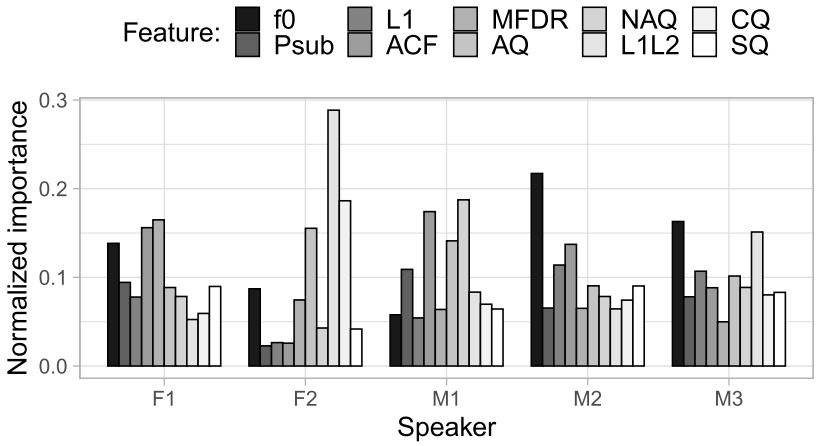

FIGURE 3. The normalized importance, as given by the Random Forest analysis, of the features extracted from the voice source signal, for each of the five speakers in our corpus (F1-F2, M1-M3), when the model is trained to discriminate neutral from breathy voice quality.

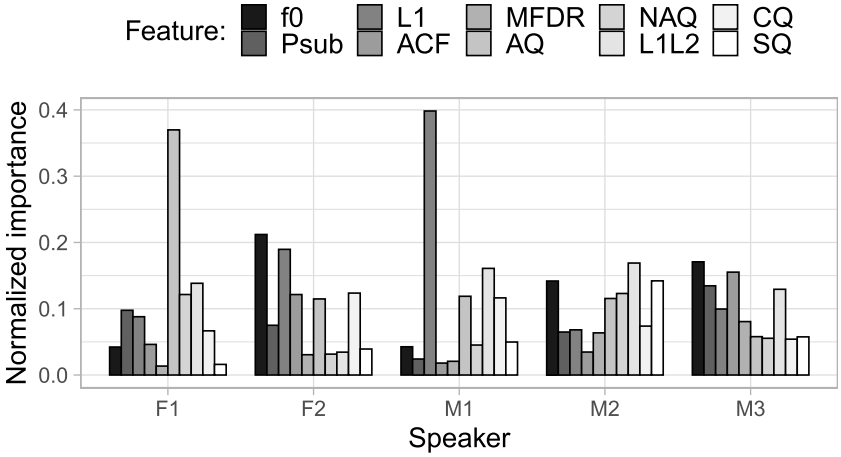

FIGURE 4. The normalized importance, as given by the Random Forest analysis, of the features extracted from the voice source signal, for each of the five speakers in our corpus (F1-F2, M1-M3), when the model is trained to discriminate neutral from pressed voice quality.

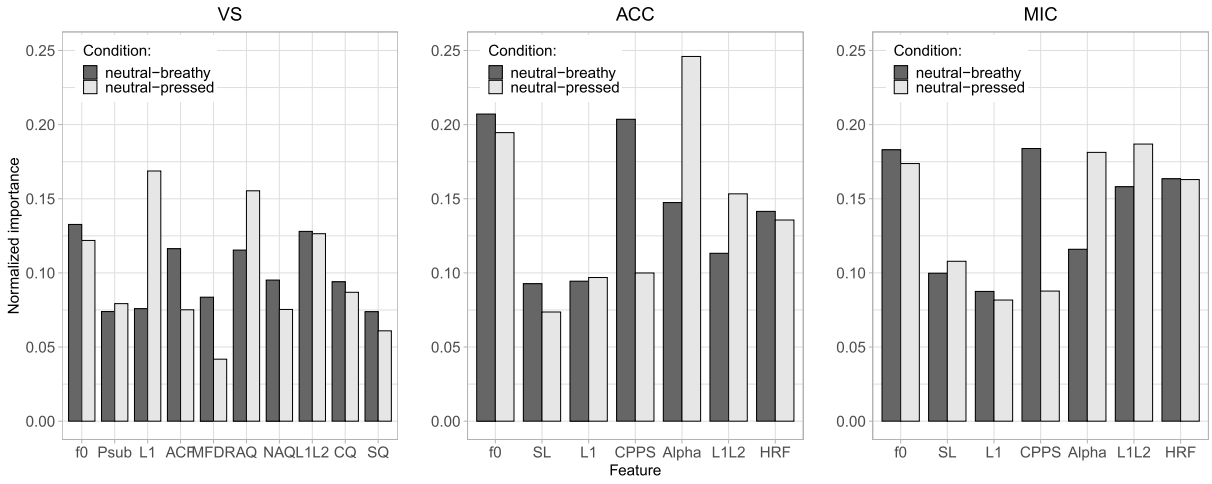

FIGURE 5. The mean normalized importance, as given by the Random Forest analysis, of the features extracted from the three input signals (the voice source VS, accelerometer ACC, microphone MIC), across the five speakers in our corpus (F1-F2, M1-M3), and for each of the two considered conditions (neutral-breathy and neutral-pressed).

panel, respectively) f_o plays an important role for discriminating neutral voice quality from the other two types. There are other similarities between the two signals, like the high ranking of CPPS in the NB case or of Alpha and L_1L_2 for NP.

Compared to the contrasts involving neutral phonation, the breathy − pressed condition shows on average much weaker reliance on f_o (see Figure 9 in the Appendix, where the importance of features in the breathy − pressed condition is displayed alongside the mean importance scores in the other two conditions). Instead, the discrimination between breathy and pressed voice relies mainly on AQ, NAQ, L_1L_2 and CQ (for the voice source) and on CPPS, Alpha, L_1L_2 and HRF (in the audio and accelerometer signals).

In order to evaluate the contribution of the voice source features on their own, the experiments were repeated without using f_o, sound level (for the accelerometer and audio signals) or P_{sub} (for the voice source signal). As expected, the mean error rates increased overall but were nevertheless comparable across the three signals (NB: 18.3%, 17.6%, 17.4%; NP: 12.9%, 12.4%, 11.7%, BP: 6.7%, 7.1%, 5.4%, for the voice source, the accelerometer and the audio respectively, none of the differences being significant). The relative importance of features also remained largely unchanged.

Relationships between the voice source, accelerometer, and audio signals

It is well-known that glottal adduction is the primary voice parameter controlling phonation type which, in turn, strongly affects flow glottogram parameters.[4,52] For example, the firmer the adduction, the longer the closed phase and the smaller the peak-to-peak flow amplitude. Hence, the flow glottogram can be assumed to reflect phonatory differences between phonation types. The question then becomes to what extent these differences are manifest in the accelerometer and microphone signals.

Given that several the voice source parameters are strongly correlated,[53] we computed the Spearman's rank correlation coefficient (ρ) between all the extracted features. In Table 1, pairwise correlations (positive or negative) of at least 0.7 between the voice source features (top), the voice source and accelerometer (middle) and the voice source and audio (bottom) are marked with a cross (for a full correlation matrix, see Table 3 in the Appendix). As expected, P_{sub} was strongly correlated with MFDR. Among the flow glottogram features the strongest correlations were found between L_1 and AC_F, AC_F and MFDR, MFDR and both AQ and NAQ, AQ and NAQ, NAQ and both L_1L_2 and CQ, and between L_1L_2 and CQ. Thus, all flow glottogram features except SQ were strongly correlated with at least one other flow glottogram feature.

TABLE 1.
Pairwise correlations (Spearman's ρ) between the voice source features (VS, top), the voice source and accelerometer features (ACC, middle), and the voice source and audio features (MIC, bottom). Correlations (positive or negative) of at least 0.7 are marked with a cross.

		f_o	P_{sub}	L_1	AC_F	MFDR	AQ	NAQ	L_1L_2	CQ	SQ
VS	P_{sub}										
	L_1										
	AC_F			×							
	MFDR		×		×						
	AQ					×					
	NAQ					×	×				
	L_1L_2							×			
	CQ							×	×		
	SQ										
ACC	f_o	×									
	SL										
	L_1										
	CPPS					×		×	×		
	Alpha						×				
	L_1L_2							×	×	×	
	HRF							×	×	×	
MIC	f_o	×									
	SL		×			×	×	×			
	L_1										
	CPPS							×	×		
	Alpha						×	×	×	×	
	L_1L_2						×	×	×	×	
	HRF					×	×	×	×	×	

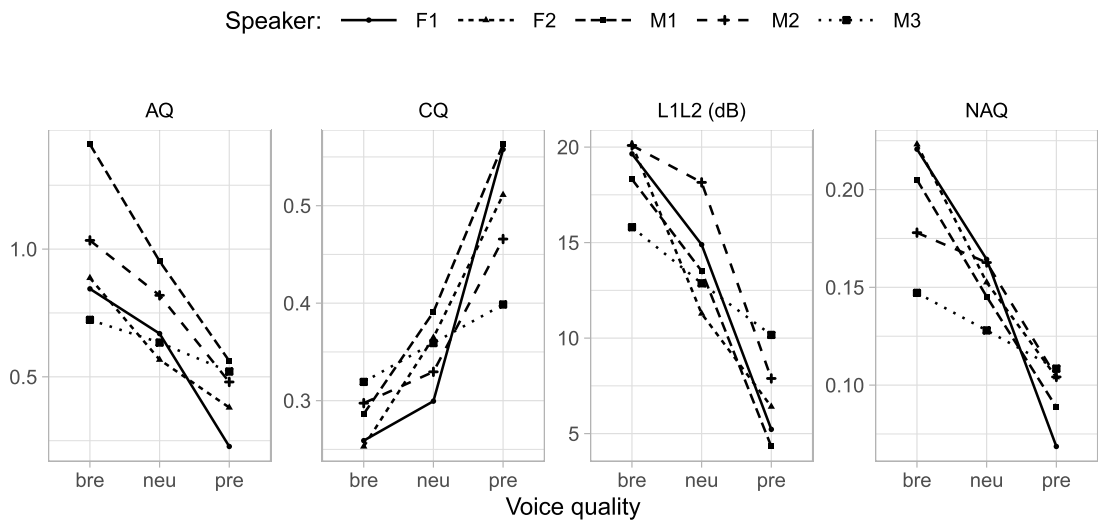

FIGURE 6. Averaged voice source features showing monotonic and consistent variation along the breathy (bre) – neutral (neu) – pressed (pre) continuum for the five participants

Since, unlike analyzing accelerometer and audio signals, flow glottogram analysis is cumbersome and time consuming, an important task of the present investigation was to analyze the relationship between the voice source features and audio and accelerometer signal features. To further elucidate the results of the classification experiments in the previous section and identify the glottal flow information, which can be retrieved from these signals, we examined the relationships between those flow glottogram features that showed a monotonic and similar variation along the breathy – neutral – pressed continuum for the five participants.

Four voice source features met these criteria: AQ, NAQ, L_1L_2 and CQ (see Figure 6). Therefore, these features are assumed to be particularly relevant for separating the three phonation types concerned. This is indeed evidenced by the fact that these features also showed higher importance than the other features in the classification task (Figure 5).

In the accelerometer and audio signals, the following features exhibited monotonic change along the breathy – neutral – pressed continuum: CPPS, L_1L_2 and HRF in the accelerometer signal, and with sound level, CPPS, L_1L_2 and HRF in the audio signal. Figure 7 displays these accelerometer and audio features averaged over condition for each participant. Table 1 reveals that these features were also highly correlated with the four features exhibiting consistent pattern in the voice source (AQ, NAQ, L_1L_2 and CQ). The other audio parameter highly correlation with these voice source features, Alpha (also included in Figure 7), showed a monotonic pattern for three out of five speakers.

DISCUSSION

The error rates obtained in the Random Forrest classification experiments revealed that features extracted from the accelerometer and the audio signals were approximately as good for discriminating phonation type as the voice source features. Phonation type is primarily controlled by laryngeal adjustments and subglottal pressure, which jointly determine voice source properties. Therefore, it was expected that the accelerometer signal would be more closely related to the voice source than the audio since the latter is strongly influenced by the vocal tract transfer function. Not only was this not the case but some of the correlations in Table 3 were also lower than expected. This could be due to subglottal resonances as well as by the transfer function of the tracheal wall tissues, which complicate the relationship between the source and the tracheal wall vibration. Since the effects of these factors are relatively constant within speakers, it should be possible to compensate for them by filtering (see Zanartu et al.[54] for an inverse filtering model designed for the accelerometer signal). However, since the the accelerometer signal was as good at discriminating between phonation types as the voice source, it is unlikely that further processing would lead to improved classification accuracy.

The audio signal's good performance relative to the other two signals can be assumed to be due to effects of the vocal tract shape accompanying the laryngeal adjustments needed for changing phonation type. Indeed, we have observed consistent differences in frequencies of the first formant across the three voice qualities. The first formant frequency tended to be lower in breathy and higher in pressed phonation as compared with neutral phonation. It seems likely that this effect was caused by changes of larynx height, high-effort phonation typically being associated with a raised larynx. A rise of the first formant frequency tends to raise the overall sound level of a vowel. This may have caused the consistent increase of the Alpha feature of the audio signal observed along the breathy – neutral – pressed continuum.

All three analyzed signals showed a great inter-participant variation, both with respect to discrimination accuracy

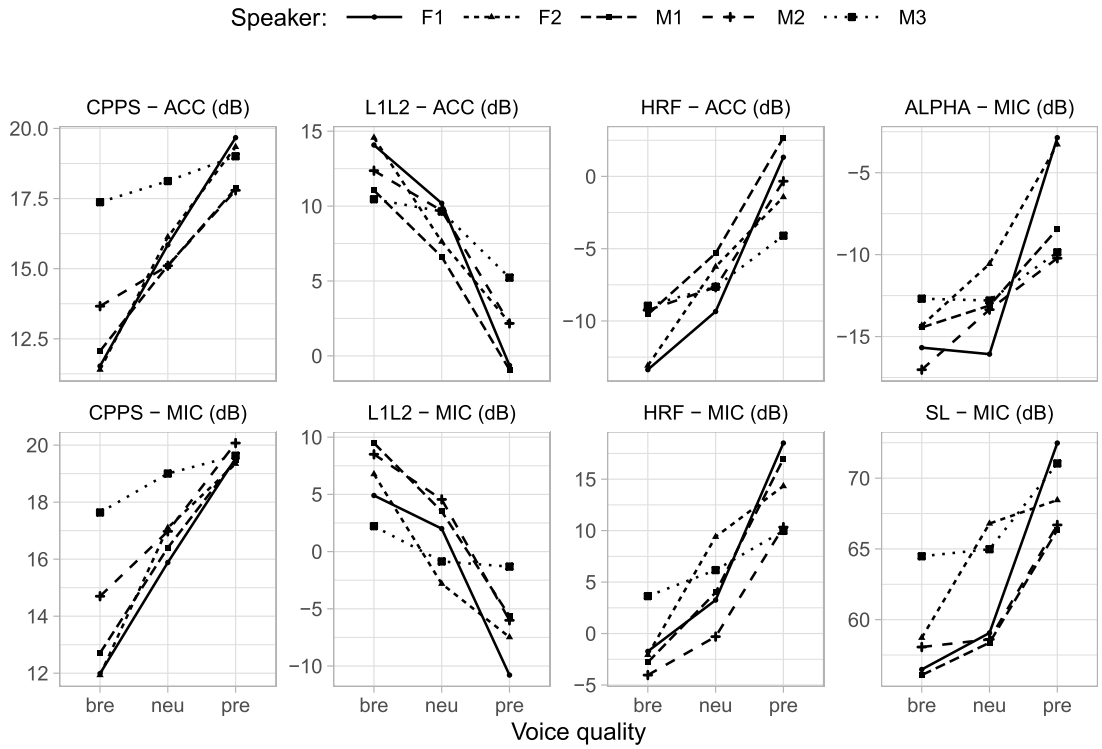

FIGURE 7. Averaged accelerometer (ACC) and audio (MIC) features showing monotonic and consistent variation along the breathy (bre) − neutral (neu) − pressed (pre) continuum for the five participants

and feature importance. Gender differences (two of the participants were female and three were male) could be a contributing factor in that regard, with significant gender variation in both vocal fold and vocal tract lengths. The former should affect the relation between glottal area and glottal flow, and consequently the relation between the latter and P_{sub}. Another explanation could be that all participants had substantial experience of choral or solo singing. Thus, their voice control was particularly well developed in neutral phonation as opposed to pressed or breathy phonation, which are generally discouraged in vocal practice. The observed interpersonal variability of voice properties also suggests a preference for speaker-dependent models, confirmed by increase in OOB error when data from all speakers were pooled (not reported in the present paper).

In spite of the individual differences, the voice source features AQ, NAQ, L_1L_2 and CQ were found to vary systematically with voice quality along the breathy − neutral − pressed continuum. Of these, the three former were typically lower in neutral than in breathy and still lower in pressed phonation, while the opposite applied to CQ. This is in agreement with previous findings.[4,52,55]

As shown in Table 1, these features were correlated with CPPS, L_1L_2 and HRF calculated from both the accelerometer and the audio signals. Both L_1L_2 and HRF are known to depend on the relative amplitude of the fundamental, which, in turn, has been found to be stronger in breathy than in neutral and stronger in neutral than in pressed.[56] Moreover, the amplitude of the voice source fundamental has been found to correlate with the peak-to-peak amplitude of the glottal flow pulse (AC_F),[57–59] with amplitude being dependent on glottal adduction: the firmer the adduction, the smaller the amplitude and the weaker the fundamental. Therefore, it is not surprising that accelerometer and audio L_1L_2 and HRF were found to be important to voice quality distinction. Similarly, CPPS was originally designed to be used on non-inverse filtered speech, which might explain its performance in this scenario.

Given that the Random Forest classifier can identify complex patterns in the data, the relationship between distribution of an individual feature and its importance score is not straightforward. However, features showing clear separation across the predicted categories can be expected to score high on importance unless their effect is overshadowed by another correlated feature. For this reason, the features varying systematically across phonation types were generally found to achieve high importance in the discrimination task.

According to the Random Forrest analysis pitch was identified as an important feature for voice quality classification. This may seem somewhat surprising, as participants were asked to keep the same pitch for the three qualities in each voice quality condition. Yet, as can be seen in Figure 8, breathy and pressed phonation were often accompanied by substantial modifications of f_o. Curiously, f_o was found to play a much smaller role in the discrimination between breathy and pressed phonation than in the other two

conditions. Again, this may be explained by the results shown in Figure 8, demonstrating that breathy and pressed phonation types often involved modification of f_o in the same direction relative to neutral. Consequently, the other features were allowed to come to the fore. Notably, the features that were assigned high importance were again among the features which were found to vary systematically between phonation types, confirming their relevance to phonation type discrimination.

CONCLUSIONS

Variation of voice quality is an integral part of the expressive repertoire of vocal production which possesses clinically relevant implications. For this reason, there is a pressing need for robust methods for estimation and classification of voice quality, preferably using a feature set with known properties and systematic relationship with the voice source. In this study, we attempted to make a small step towards this goal by comparing temporal, spectral and cepstral properties of the voice source, the audio and the neck-surface acceleration collected in a controlled experiment with trained singer participants.

Overall, we found that the classification algorithm rendered comparable error rates for these three signals. This was contrary to the expectation that features extracted from the voice source should result in the highest accuracy, and that the skin acceleration signal should outperform the audio signal. In addition, several voice source features (AQ, NAQ, L_1L_2 and CQ) were found to vary systematically with phonation type, a variation mirrored in certain accelerometer signal features (HRF, L_1L_2 and CPPS) and audio signal features (HRF, L_1L_2, CPPS and sound level).

To optimize the accuracy of the voice source data, we limited the analyses to the vowel [æ:], in which the first and second formant frequencies are high and wide apart, thus facilitating the tuning of the inverse filters. This limitation raises the question whether a material consisting of connected speech would yield a similar classification accuracy of voice quality across the different feature sets. Given that inverse filtering of such material is difficult, comparison against voice source features might be impossible. Similarly, the audio signal of running speech would present additional challenges for voice quality classification due to the articulatory variation. For these reasons, neck-surface acceleration might offer an advantage in that scenario. We leave this question open for future research.

Acknowledgements

This research is supported by the Swedish Research Council grant *Prosodic functions of voice quality dynamics* (VR 2019-02932) to the first author.

APPENDIX

TABLE 2.
Mean f_o (Hz), minimum and maximum P_{sub} values, and minimum and maximum $SPL_{@30cm}$ values for each speaker, phonation type and pitch condition.

				P_{sub} (cm H_20)		$SPL_{@30cm}$ (dB)	
Speaker	Phonation type	Pitch	Mean f_o (Hz)	Min	Max	Min	Max
F1	Breathy	Low	166.9	3.7	13.9	65.9	79.7
F1	Breathy	Mid	261.9	5.5	12.5	65.9	82.3
F1	Breathy	High	352.9	4.6	16.2	65.1	84.9
F1	Neutral	Low	182.5	2.5	17.8	63.6	89.7
F1	Neutral	Mid	232.9	3.2	22.1	69.1	92.3
F1	Neutral	High	312.4	3.9	22.8	59.8	91.6
F1	Pressed	Low	230.0	14.8	27.9	85.6	92.5
F1	Pressed	Mid	307.4	15.9	26.8	81.5	93.7
F1	Pressed	High	411.3	19.5	28.0	81.3	97.0
F2	Breathy	Low	181.6	7.9	18.6	67.4	73.1
F2	Breathy	Mid	241.6	5.5	14.3	68.8	81.4
F2	Breathy	High	357.1	4.3	15.6	73.3	88.0
F2	Neutral	Low	186.9	3.6	3.6	69.4	92.4
F2	Neutral	Mid	283.5	6.6	9.9	73.7	93.0
F2	Neutral	High	419.4	8.4	19.0	85.7	95.3
F2	Pressed	Low	222.8	7.6	20.6	75.7	91.9
F2	Pressed	Mid	277.1	10.3	21.1	75.9	92.8
F2	Pressed	High	369.5	9.2	26.9	75.1	94.8
M1	Breathy	Low	104.5	3.7	16.4	62.2	81.7
M1	Breathy	Mid	154.1	4.9	21.4	64.2	88.0
M1	Breathy	High	221.8	6.5	20.8	60.2	90.8
M1	Neutral	Low	106.5	5.0	12.2	64.6	80.5
M1	Neutral	Mid	153.9	5.0	12.6	67.6	87.1
M1	Neutral	High	215.3	4.4	16.8	69.4	92.4
M1	Pressed	Low	110.2	5.5	13.2	67.5	86.6
M1	Pressed	Mid	131.5	6.7	16.5	76.5	93.8
M1	Pressed	High	277.0	9.2	26.0	78.7	93.7
M2	Breathy	Low	124.6	5.4	13.6	62.6	75.1
M2	Breathy	Mid	185.9	10.2	23.5	68.7	83.0
M2	Breathy	High	258.5	9.8	32.0	68.2	85.0
M2	Neutral	Low	145.1	4.1	16.6	63.2	83.9
M2	Neutral	Mid	200.0	5.1	27.2	64.0	89.3
M2	Neutral	High	308.8	14.3	33.6	75.8	98.8
M2	Pressed	Low	167.6	12.5	21.4	77.6	85.1
M2	Pressed	Mid	221.0	11.9	28.4	77.9	87.6
M2	Pressed	High	286.8	14.4	35.5	82.5	98.1
M3	Breathy	Low	160.0	5.4	15.1	69.2	82.7
M3	Breathy	Mid	198.0	5.5	16.9	70.7	86.7
M3	Breathy	High	246.0	7.4	20.3	82.4	94.5
M3	Neutral	Low	155.0	4.2	10.9	68.9	85.4
M3	Neutral	Mid	184.0	6.0	11.9	68.5	89.9
M3	Neutral	High	235.0	8.1	15.4	78.6	96.4
M3	Pressed	Low	176.0	7.0	13.9	74.8	87.4
M3	Pressed	Mid	213.0	9.6	19.2	80.1	89.9
M3	Pressed	High	222.0	7.3	17.8	76.2	94.5

TABLE 3.
Pairwise correlations (Spearman's ρ) between the voice source (VS), accelerometer (ACC) and audio (MIC) features.

		VS									ACC							MIC						
		P_{sub}	L_1	AC_F	MFDR	AQ	NAQ	L_1L_2	CQ	SQ	f_o	SL	L_1	CPPS	Alpha	L_1L_2	HRF	f_o	SL	L_1	CPPS	Alpha	L_1L_2	HRF
VS	f_o	0.4	-0.28	-0.19	0.25	-0.52	0.08	0.03	-0.05	0.06	0.98	0.46	0.47	-0.05	0.37	0.16	-0.23	0.98	0.37	0.47	-0.15	0.22	-0.4	0.12
	P_{sub}		0.26	0.49	0.8	-0.69	-0.53	-0.46	0.48	0.39	0.4	0.67	0.58	0.43	0.66	-0.46	0.45	0.4	0.76	0.52	0.38	0.63	-0.61	0.56
	L_1			0.83	0.43	0.08	-0.09	0.04	-0.04	0	-0.27	0.05	0.07	0.15	-0.1	-0.08	0.11	-0.27	0.23	0.52	0.13	0.02	0.12	0
	AC_F				0.71	-0.18	-0.31	-0.17	0.14	0.21	-0.17	0.26	0.23	0.37	0.1	-0.27	0.3	-0.17	0.48	0.58	0.34	0.22	-0.12	0.24
	MFDR					-0.79	-0.73	-0.57	0.54	0.49	0.25	0.49	0.4	0.7	0.62	-0.57	0.56	0.26	0.89	0.59	0.62	0.67	-0.66	0.72
	AQ						0.78	0.69	-0.66	-0.47	-0.51	-0.48	-0.36	-0.69	-0.8	0.58	-0.55	-0.52	-0.85	-0.34	-0.62	-0.77	0.86	-0.83
	NAQ							0.85	-0.84	-0.54	0.07	-0.21	-0.07	-0.84	-0.68	0.82	-0.83	0.07	-0.72	-0.04	-0.81	-0.75	0.75	-0.9
	L_1L_2								-0.88	-0.42	0.03	-0.15	-0.01	-0.72	-0.69	0.86	-0.84	0.02	-0.64	0.08	-0.73	-0.72	0.81	-0.89
	CQ									0.36	-0.05	0.16	0.03	0.67	0.68	-0.79	0.79	-0.04	0.6	-0.07	0.68	0.71	-0.72	0.82
	SQ										0.06	0.29	0.2	0.5	0.4	-0.49	0.5	0.06	0.38	0.05	0.5	0.41	-0.47	0.48
ACC	f_o											0.47	0.49	-0.05	0.37	0.17	-0.24	1	0.37	0.5	-0.15	0.23	-0.39	0.11
	SL												0.97	0.21	0.42	-0.14	0.15	0.47	0.45	0.45	0.18	0.33	-0.35	0.23
	L_1													0.08	0.31	0.02	-0.02	0.49	0.35	0.5	0.06	0.22	-0.23	0.09
	CPPS														0.58	-0.71	0.72	-0.05	0.71	0.16	0.93	0.69	-0.66	0.81
	Alpha															-0.63	0.61	0.37	0.72	0.19	0.57	0.82	-0.79	0.77
	L_1L_2																-0.99	0.16	-0.56	0.11	-0.72	-0.66	0.68	-0.81
	HRF																	-0.23	0.54	-0.12	0.74	0.64	-0.65	0.79
MIC	f_o																		0.37	0.49	-0.15	0.23	-0.39	0.11
	SL																			0.59	0.63	0.76	-0.8	0.83
	L_1																				0.07	0.21	-0.21	0.11
	CPPS																					0.64	-0.64	0.79
	Alpha																						-0.75	0.81
	L_1L_2																							-0.89

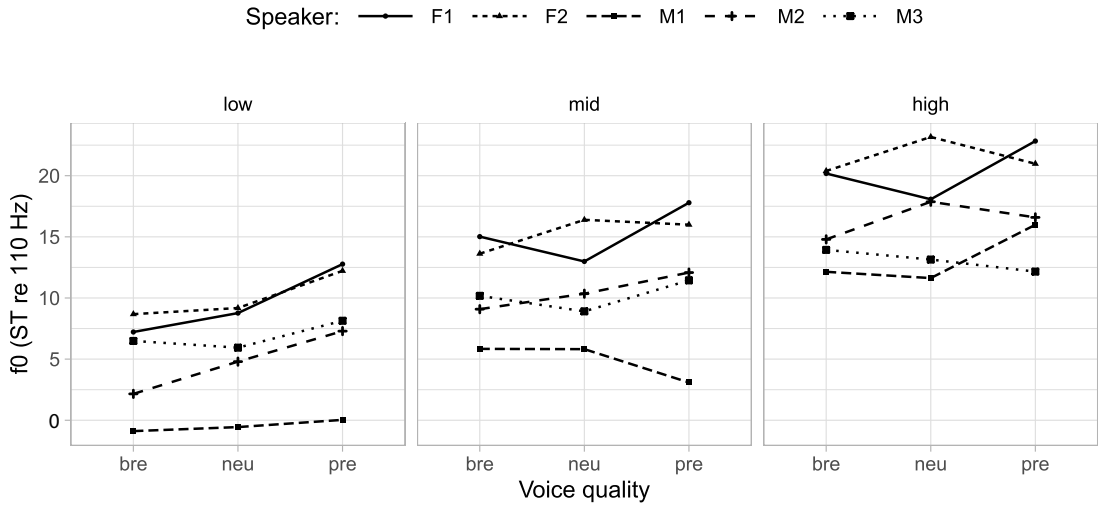

FIGURE 8. Averaged values of f_o for each speaker, pitch condition (*low, mid* and *high*) and voice quality (*bre*: breathy, *neu*: neutral, *pre*: pressed).

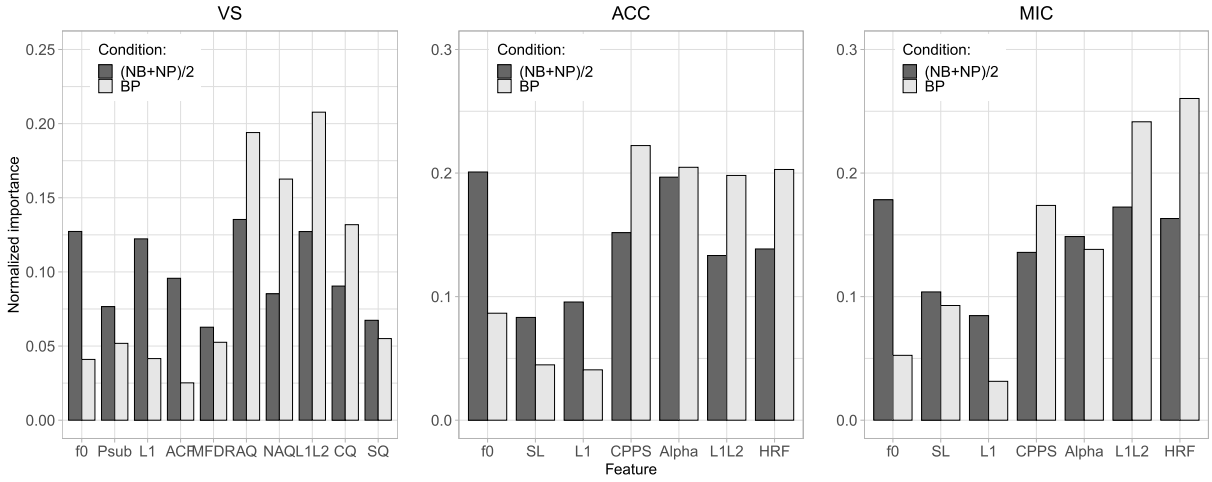

FIGURE 9. The mean normalized importance, as given by the Random Forest analysis, of the features extracted from the three input signals (the voice source VS, accelerometer ACC, microphone MIC), across the five speakers in our corpus (F1-F2, M1-M3). Displayed are the average of the importance in the neutral-breathy and neutral-pressed conditions, (NB+NP)/2, and the importance for the breathy-pressed condition.

REFERENCES

1. Childers DG, Lee CK. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*. 1991;90 (5):2394–2410.
2. Esling JH, Moisik SR, Benner A, Crevier-Buchman L. *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press; 2019.
3. Sundberg J. Objective characterization of phonation type using amplitude of flow glottogram pulse and of voice source fundamental. *Journal of Voice*. 2020;36(1):4–14.
4. Gordon M, Ladefoged P. Phonation types: A cross-linguistic overview. *Journal of Phonetics*. 2001;29(4):383–406.
5. Kuang J, Keating P. Vocal fold vibratory patterns in tense versus lax phonation contrasts. *Journal of the Acoustical Society of America*. 2014;136(5):2784–2797.
6. Maryn Y, Weenink D. Objective dysphonia measures in the program Praat: Smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice*. 2015;29(1):35–43.
7. Sauder C, Bretl M, Eadie T. Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and analysis of dysphonia in speech and voice (ADSV). *Journal of Voice*. 2017;31(5):557–566.
8. Airas M, Alku P. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*. 2006;63(1):26–46.
9. Gobl C, Chasaide AN. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*. 2003;40(1 −2):189–212.
10. Scherer KR, Sundberg J, Tamarit L, Salomão GL. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*. 2015;29 (1):218–235.
11. Shriberg EE. Phonetic consequences of speech disfluency. In: *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 1999)*. 1999:619−622. Florence, Italy

12. Levitan R, Hirschberg J. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: *Proceedings of Interspeech 2011*. 2011:3081–3084. Florence, Italy
13. Chasaide AN, Yanushevskaya I, Gobl C. Prosody of voice: Declination, sentence mode and interaction with prominence. In: *Proceedings of the XVIIIth International Congress of Phonetic Sciences (ICPhS 2015)*. 2015. Glasgow, UK
14. Carlson R, Hirschberg J, Swerts M. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*. 2005;46(3–4):326–333.
15. Ludusan B, Wagner P, Włodarczak M. Cue interaction in the perception of prosodic prominence: The role of voice quality. In: *Proceedings of Interspeech 2021*. 2021:1006–1010.
16. Vishnubhotla S, Espy-Wilson C. Automatic detection of irregular phonation in continuous speech. In: *Proceedings of Interspeech 2006*. 2006:949–952. Pittsburgh, PA
17. Kane J, Gobl C. Identifying regions of non-modal phonation using features of the wavelet transform. In: *Proceedings of Interspeech 2011*. 2011:177–180. Florence, Italy
18. Székely E, Kane J, Scherer S, Gobl C, Carson-Berndsen J. Detecting a targeted voice style in an audiobook using voice quality features. *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)* 2012 :4593–4596.
19. Kane J, Gobl C. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*. 2013;21(6):1170–1179.
20. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech Language and Hearing Research*. 1996;39(2).
21. Heman-Ackah YD, Michael DD, Goding GS. The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice*. 2002;16(1):20–27.
22. Borsky M, Cocude M, Mehta DD, Zañartu M, Gudnason J. Classification of voice modes using neck-surface accelerometer data. In: *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017:5060–5064. New Orleans, LA.
23. Drugman T, Kane J, Gobl C. Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language*. 2014;28(5):1233–1253.
24. Ishi CT, Sakakibara K-I, Ishiguro H, Hagita N. A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing*. 2008;16(1):47–56.
25. Borsky M, Mehta DD, Gudjohnsen JP, Guðnason J. Classification of voice modality using electroglottogram waveforms. In: *Proceedings of Interspeech 2016*. 2016:3166–3170. San Francisco CA, USA
26. Childers DG, Ahn C. Modeling the glottal volume-velocity waveform for three voice types. *The Journal of the Acoustical Society of America*. 1995;97(1):505–519.
27. Székely E, Cabral JP, Cahill P, Carson-Berndsen J. Clustering expressive speech styles in audiobooks using glottal source parameters. In: *Proceedings of Interspeech 2011*. 2011:2409–2412.Florence, Italy
28. Alku P. Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication*. 1992;11(2–3):109–118.
29. Cabral JP, Renals S, Richmond K, Yamagishi J. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis* 2007. Bonn, Germany
30. Borsky M, Mehta DD, Van Stan JH, Gudnason J. Modal and nonmodal voice quality classification using acoustic and electroglottographic features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2017;25(12):2281–2291.
31. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP — A collaborative voice analysis repository for speech technologies. In: *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*. IEEE; 2014:960–964.
32. Stevens KN, Kalikow D, Willemain T. A miniature accelerometer for detecting glottal waveforms and nasalization. *Journal of Speech, Language, and Hearing Research*. 1975;18(3):594–599.
33. Askenfelt A, Gauffin J, Kitzing P, Sundberg J. Electroglottograph and contact microphone for measuring vocal pitch. *Speech Transmission Laboratory, Quarterly Progress and Status Report*. 1977;4:13–21.
34. Sundberg J. Chest wall vibrations in singers. *Journal of Speech and Hearing Research*. 1983;26(3):329–340.
35. Coleman RF. Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice. *Journal of Voice*. 1988;2(3):200–205.
36. Švec JG, Titze IR, Popolo PS. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *The Journal of the Acoustical Society of America*. 2005;117(3):1386–1394.
37. Fryd AS, Van Stan JH, Hillman RE, Mehta DD. Estimating subglottal pressure from neck-surface acceleration during normal voice production. *Journal of Speech, Language, and Hearing Research*. 2016;59(6):1335–1345.
38. McKenna VS, Llico AF, Mehta DD, Perkell JS, Stepp CE. Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers. *Journal of Speech, Language, and Hearing Research*. 2017;60(12):3404–3416.
39. Horii Y. An accelerometric measure as a physical correlate of perceived hypernasality in speech. *Journal of Speech, Language, and Hearing Research*. 1983;26(3):476–480.
40. Lippman R. Detecting nasalization using a low-cost miniature accelerometer. *Journal of Speech, Language, and Hearing Research*. 1981;24(3):314–317.
41. Mehta DD, Espinoza VM, Van Stan JH, Zañartu M, Hillman RE. The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *The Journal of the Acoustical Society of America*. 2019;145(5):EL386–EL392.
42. Mehta DD, Van Stan JH, Zañartu M, Ghassemi M, Guttag JV, Espinoza VM, et al. Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology*. 2015;3.
43. Llico AF, Zañartu M, Gonzalez AJ, Wodicka GR, Mehta DD, Van Stan JH, et al. Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *Journal of the Acoustical Society of America*. 2015;138(1):EL14–9.
44. Ghassemi M, Van Stan JH, Mehta DD, Zañartu M, Cheyne II HA, Hillman RE, et al. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules. *IEEE Transactions on Biomedical Engineering*. 2014;61(6):1668–1675.
45. Gelzinis A, Verikas A, Vaiciukynas E, Bacauskiene M, Minelga J, Hållander M, et al. Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders. In: *Proceedings of the 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*. IEEE; 2014:125–132.
46. Granqvist S. Sopran [Computer program]. 2022. https://tolvan.com/index.php?page=/sopran/sopran.php.
47. Boersma P., Weenink D.. Praat: doing phonetics by computer. 2021. Computer program, http://www.praat.org/.
48. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. *Journal of Speech, Language, and Hearing Research*. 1990;33(2):298–306.
49. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5–32.
50. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/
51. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18–22.https://CRAN.R-project.org/doc/Rnews/
52. Millgård M, Fors T, Sundberg J. Flow glottogram characteristics and perceived degree of phonatory pressedness. *Journal of Voice*. 2016;30(3):287–292. https://doi.org/10.1016/j.jvoice.2015.03.014.

53. Sundberg J. Flow glottogram and subglottal pressure relationship in singers and untrained voices. *Journal of Voice*. 2018;32(1):23–31.
54. Zañartu M, Ho JC, Mehta DD, Hillman RE, Wodicka GR. Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration. *IEEE Transactions on Audio, Speech and Language Processing*. 2013;21(9):1929–1939.
55. Alku P, Bäckström T, Vilkman E. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*. 2002;112(2):701–710. https://doi.org/10.1121/1.1490365.
56. Kreiman J, Shue Y-L, Chen G, Iseli M, Gerratt BR, Neubauer J, et al. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *Journal of the Acoustical Society of America*. 2012;132(4):2625–2632.
57. Gauffin J, Sundberg J. Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech, Language, and Hearing Research*. 1989;32(3):556–565. https://doi.org/10.1044/jshr.3203.556.
58. Sundberg J. Objective characterization of phonation type using amplitude of flow glottogram pulse and of voice source fundamental. *Journal of Voice*. 2022;36(1):4–14. https://doi.org/10.1016/j.jvoice.2020.03.018.
59. Sundberg J, Thalén M, Alku P, Vilkman E. Estimating perceived phonatory pressedness in singing from flow glottograms. *Journal of Voice*. 2004;18(1):56–62. https://doi.org/10.1016/j.jvoice.2003.05.006.