Interspeech 2022
18-22 September 2022, Incheon, Korea

Convolutional Neural Networks for Classification of Voice Qualities from Speech and Neck Surface Accelerometer Signals

Sudarsana Reddy Kadiri, Farhad Javanmardi and Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland
sudarsana.kadiri@aalto.fi, farhad.javanmardi@aalto.fi, paavo.alku@aalto.fi

Abstract

Prior studies in the automatic classification of voice quality have mainly studied support vector machine (SVM) classifiers using the acoustic speech signal as input. Recently, one voice quality classification study was published using neck surface accelerometer (NSA) and speech signals as inputs and using SVMs with hand-crafted glottal source features. The present study examines simultaneously recorded NSA and speech signals in the classification of three voice qualities (breathy, modal, and pressed) using convolutional neural networks (CNNs) as classifier. The study has two goals: (1) to investigate which of the two signals (NSA vs. speech) is more useful in the classification task, and (2) to compare whether deep learning -based CNN classifiers with spectrogram and mel-spectrogram features are able to improve the classification accuracy compared to SVM classifiers using hand-crafted glottal source features. The results indicated that the NSA signal showed better classification of the voice qualities compared to the speech signal, and that the CNN classifier outperformed the SVM classifiers with large margins. The best mean classification accuracy was achieved with mel-spectrogram as input to the CNN classifier (93.8% for NSA and 90.6% for speech).

Index Terms: Voice quality, neck surface accelerometer, Mel-spectrogram, computational paralinguistics, CNNs.

1. Introduction

Humans are capable of producing various voice qualities, such as breathy, harsh, and creaky by regulating the laryngeal muscles along with the respiratory force [1–3]. Voice quality is a perceptual attribute and is often defined as auditory coloring of a speaker's voice [1, 4, 5]. Voice quality plays a major role in conveying para-linguistic information such as emotions, health and personality [6–10]. Expressing politeness and intimacy often involves using breathy voice [11], and expressing emotions of high arousal such as anger and fear typically involves using tense voice [12, 13]. To produce phonological contrasts, voice quality is also used in certain languages [5, 14–17]. This study focuses on three voice qualities, namely, breathy, modal and pressed. Pressed and breathy are considered to be opposite ends of the voice quality continuum, while modal is in the middle of the continuum [18, 19]. More details about the functioning of the speech production mechanism in the generation of these three voice qualities can be found in [1, 4].

Voice quality is known to be closely associated with the vibration mode of the vocal folds and thereby with the characteristics of the glottal flow pulse [1, 4, 19]. The glottal pulse derived with glottal inverse filtering (GIF) has been shown to vary from a smooth symmetric waveform in breathy voice to an asymmetric waveform in pressed voice [19, 20]. These time-domain changes in the glottal pulse affect the spectral decay of the voice excitation in the frequency-domain [21, 22]. Several features have been developed to parameterize the glottal source waveform using time-domain and frequency-domain approaches [18, 19, 23]. Time-domain features (such as, the closing quotient, the speed quotient and the quasi-open quotient), as well as amplitude-based features (such as the normalized amplitude quotient) have been derived from glottal source waveform and its derivative [19, 24, 25]. Frequency-domain features (such as, the amplitude difference between the fundamental (F0) and second harmonic (H1-H2) [22], the harmonic richness factor (HRF) [2] and the parabolic spectral parameter (PSP) [26]) have been developed to capture the spectral decay of the glottal source waveform. Voice quality has also been studied by fitting the derivative of the estimated glottal source with the artificial Liljencrants-Fant (LF) model [13, 27]. In [8, 28–30], authors measured the impact of the glottal source on the spectrum of speech using features such as the F0, H1-H2, the spectral slope between 2 kHz and the fourth harmonic (H4), H2-H4, and the spectral slope between 5 kHz and 2 kHz. Low-frequency spectral density was proposed in [21] to characterise breathy voices, as spectral energy is larger at low frequencies. Sharper changes in the vocal fold closure of pressed voice were captured using the maximum dispersion quotient feature based on the linear prediction residual signal in [18].

Glottal source features were used with the mel-frequency cepstral coefficients (MFCCs) in [18, 23] for the automatic classification of voice qualities from speech signals. The experiments reported in [31, 32] used features derived from the approximate glottal source signals such as the linear prediction residual and the zero frequency filtered signal (ZFF signal) to characterize voice qualities in speech and singing. Features including the energy of the ZFF signal, the slope of the ZFF signal at zero crossings, the energy of excitation, and the loudness measure were used for analysis and automatic classification of voice qualities in [31, 32]. Moreover, cepstral coefficients derived from the zero-time windowing spectrum and the single frequency filtering spectrum were explored for the classification of voice qualities in speech and singing in [33].

Neck surface accelerometer (NSA) signals provide an alternative means to measure vocal fold vibration characteristics [34–36]. The NSA is a sensor that captures the vibration of the vocal folds during speech production. The NSA is mounted to below-glottis skin surface and hence signals collected with this sensor are less affected by the vocal tract compared to acoustic speech signals. NSA signals are justified to be used in analysis and classification of voice qualities because they are directly associated with vocal fold dynamics. NSA signals were shown to supplement acoustic speech signals in studying vocal dose and vocal hyperfunction in [37–40].

As per the authors' knowledge, there are only three studies in the automatic classification of voice qualities using the NSA signal. In [41], four voice qualities (breathy, modal, pressed, and rough) were classified using the Gaussian mixture model

10.21437/Interspeech.2022-10513

(GMM) classifier and MFCCs derived from the NSA signal. In [42], several features such as harmonics, jitter, shimmer and entropy derived from the NSA signal were used for discriminating three voice qualities (modal, breathy and pressed) using linear discriminant analysis, decision tree, support vector machine (SVM), K-nearest neighbours, and multi-layer perceptron classifiers. In [43], authors used several glottal source features and MFCCs with the SVM classifier for discriminating the same three voice qualities as in [42].

To the best of our knowledge, there is only one previous voice quality classification study comparing $simultaneous$ recordings of acoustic speech signals and NSA signals [43]. In their study, authors used SVM classifiers and MFCC features derived from source waveforms, extracted both from speech and NSA signals. However, there are no prior voice quality classification studies on using recent deep learning classifiers such as convolutional neural networks (CNNs) with popular spectral representations, such as spectrogram or mel-spectrogram. Hence, the aim of the current study is to investigate CNNs with spectrogram and mel-spectrogram as input feature representations in classification of three voice qualities (breathy, modal, pressed). In particular we are interested to compare how the deep learning -based CNN classifier based on spectrogram inputs performs in the automatic voice quality classification compared to state of the art, that is, SVM classifiers based on hand-engineered features (studied in [43]). Given the benefits shown by the NSA signal in [43], we also compare the classification performance between the NSA signal and the acoustic speech signal.

2. Voice Quality Database

In the current study, we use the voice quality database described in [42]. This database consists of five vowels ($[a]$, [æ], $[e]$, $[i]$ and $[u]$) uttered in three voice qualities (breathy, modal and pressed) [42]. The vowels were produced by 31 native Canadian English female speakers with an age range of 18-40 years. Each vowel was uttered three times in the three voice qualities, which resulted in a total of 1395 vowels ($5 \cdot 3 \cdot 3 \cdot 31$). The database includes simultaneous recordings of acoustic speech signals and NSA signals. The data was collected with a sampling frequency of 44.1 kHz, but it was down-sampled to 16 kHz in this study. All the speech signals were perceptually assessed by five speech language pathologists and the scores were analysed in terms of intra- and inter-rater reliability. The perceptual assessment resulted in 952 reliable signals, out of which 395 are breathy voices, 285 are modal voices and 272 are pressed voices. The total duration of the database is around 52 minutes. As per the authors' knowledge, the database of [42] is the only voice quality repository which has simultaneous recordings of NSA and speech signals, and available for research purposes.

3. Classification Task Setup

A schematic block diagram of the proposed voice quality classification system is shown in Figure 1. The system converts the input (i.e., speech or NSA signal) into a time-frequency representation (i.e., spectrogram or mel-spectrogram) to be processed by a CNN classifier to predict the voice quality label (i.e. breathy vs. modal vs. pressed).

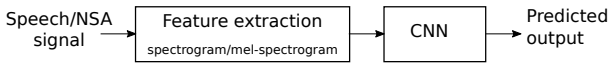

Figure 1: *A schematic block diagram of the proposed voice quality classification system with spectrogram/mel-spectrogram feature representation as input to a CNN classifier.*

3.1. Feature Extraction

The input speech/NSA signal is transformed into two popularly used spectro-temporal representations, namely, spectrogram and mel-spectrogram. Both of these representations are computed using the short-term spectral analysis, where the speech/NSA signal is split into overlapping time-frames (25-ms frames with a shift of 5 ms with the Hamming window) and the spectrum is computed with 1024-point discrete Fourier transform (DFT). The spectrogram is computed by taking the logarithm of the amplitude spectrum for all the time-frames. The mel-spectrogram is computed with a mel-filterbank of 80 filters. Figure 2 shows examples of spectrograms computed from NSA signals in breathy, modal and pressed phonation. It can be observed that there are large spectral variations in the strength of the harmonics (from low to high frequencies) between the three voice qualities.

3.2. Convolution neural network (CNN) classifier

CNNs are the most widely used deep learning architectures in speech, text, and image processing [44–46]. A CNN is usually formed by convolution layers, max−pooling and fully connected layers. The convolution layers extract the localized temporal features and translation-invariant features. The pooling layer compresses the frame-level information derived from the convolution layer to sentence-level information. Fully connected layers are trained to classify the three voice qualities. Table 1 shows the architecture of the CNN classifier used in this study. The hyper-parameters of the convolution layer are the number of filters, filter size, and stride. The max−pooling layer is defined by the stride and kernel size. Fully connected layers are defined according to input and output dimensions. The convolutional layers and max−pooling layers are frame-level layers, and the layers after max−pooling process sentence-level representations. The rectified linear unit activation function is commonly used in all the layers. Training was carried out using 100 epochs and Adam optimizer was used with a learning rate of 0.001. The length of the input signal was fixed to 2 sec, by truncating or appending based on the original length of the signal.

3.3. Reference classifiers based on a prior study

In order to compare the CNN classifier described in Sec. 3.2 with state of the art, we use the classification results published in [43]. The results reported in [43] were obtained by using different hand-crafted glottal source features along with the conventional MFCCs and by using SVM as classifier. Most importantly, the experiments reported in [43] were computed from the same data used for the CNN -based experiments of the current study, which enables direct comparisons. More specifically, five feature sets (four glottal source sets and one MFCC set) were studied in [43] using both speech and NSA signals. The first glottal source feature set consists of 12 glottal features derived using GIF, out of which 9 are time-domain features and 3 are frequency-domain features. This feature set is referred to as GIF-1D (GIF-based 1-dimensional features). The second feature set consists of four features derived using the ZFF method and this feature set is referred to as ZFF-1D. The third and fourth feature sets consist of MFCCs derived from the glottal source waveforms estimated by GIF and ZFF, and they are referred to as GIF-MFCC and ZFF-MFCC, respectively. The fifth feature set corresponds to the MFCCs derived directly from the input signal and this set is simply referred to as MFCCs. Ex-

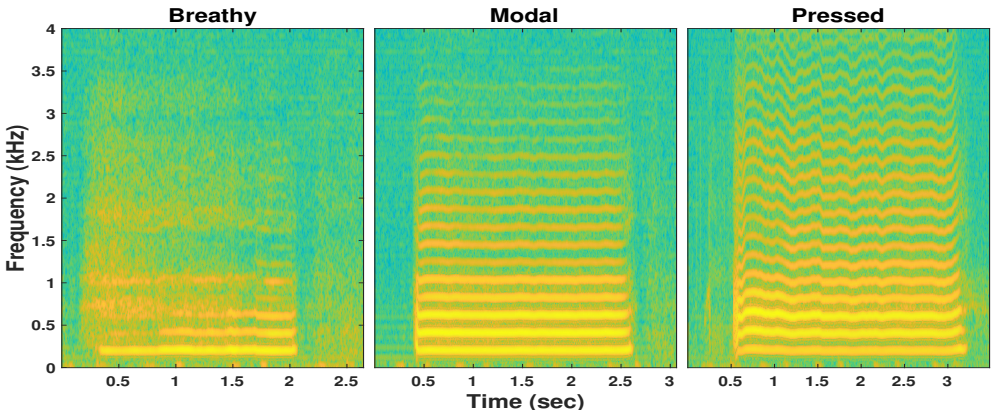

Figure 2: *Spectrograms of the NSA signal for the vowel [a] in breathy, modal and pressed voices.*

Table 1: *CNN architecture used for voice quality classification. Conv denotes convolution layer and FC denotes fully connected layer.*

Layers	Conv1	Max pooling1	Conv2	Max pooling2	Flatten	FC1	FC2	FC3
No. filters/output dim.	128	-	256	-	256*(#frames/4)*(#spectral bins/4)	256	64	3
Kernel size	(3, 3)	(3, 3)	(3, 3)	(3, 3)	-	-	-	-
Stride	(2, 2)	(2, 2)	(2, 2)	(2, 2)	-	-	-	-

periments were conducted with the individual feature sets and combinations of the feature sets to see complementary information. In total, the following 7 reference feature sets are used for comparison in the present study:

- GIF-1D
- ZFF-1D
- GIF-MFCC
- ZFF-MFCC
- MFCCs
- Combination of all glottal feature sets (GIF-1D, GIF-MFCC, ZFF-1D and ZFF-MFCC) and referred to as 'All-glottal'
- Combination of all-glottal and MFCCs (referred to as 'All-glottal+MFCCs').

3.4. Experimental setup

Experiments are carried out using a LOSO (leave-one-speaker-out) scheme. That is, out of the 31 speakers, the data of one speaker (consisting of all three voice qualities) is used for testing, and the data of the remaining 30 speakers (consisting also of all three voice qualities) is used for training and validation. Specifically out of the 30 speakers, 29 speakers are used for training and 1 speaker is used for validation. First, classification accuracies for each of the speaker are computed, and then the mean and standard deviation of these accuracies of all the speakers are computed. Apart from accuracy metric, confusion matrices are also used for assessing the performance.

4. Results

This section reports the results by first describing the classification accuracies obtained and then describing the confusion matrices for the speech and NSA signals. Table 2 shows the results in terms of the mean and standard deviation of accuracy. It can be observed that the performance of the CNN classifiers is clearly better than all the baseline systems that use the SVM classifier and different glottal features and/or MFCCs. The improvement given by the CNN classifier is also evident both when using the acoustic speech signal and the NSA signal as input. The use of the NSA input yielded larger accuracy compared to the use of the acoustic speech signal in all features compared. This observation corroborates previous findings reported using conventional hand-crafted features in [43] suggesting that the NSA signal indeed carries more voice quality -related information compared to the acoustic speech signal. The table also shows that for both the speech and NSA inputs, the CNN with the mel-spectrogram feature yielded a slightly larger accuracy compared to the CNN using the spectrogram feature. From the all classifiers, the best accuracy (of 93.8%) was obtained by the CNN classifier that used the mel-spectrogram feature representation computed from the NSA signal.

Table 2: *Voice quality classification accuracy (mean and standard deviation) for the speech signal and the NSA signal.*

Feature set	**Speech** [%]	**NSA** [%]
CNN classifier		
Spectrogram	89.2±4.5	93.4±4.2
Mel-spectrogram	**90.6**±4.8	**93.8**±4.1
SVM classifier (based on [43])		
GIF-1D	70.9±3.4	74.5±3.9
ZFF-1D	66.3±2.1	73.3±0.9
GIF-MFCC	63.4±2.0	71.6±5.7
ZFF-MFCC	67.4±3.9	76.7±2.5
MFCCs	75.8±1.3	84.0±3.4
All-glottal	76.9±4.6	84.9±2.2
All-glottal+MFCCs	80.6±2.2	86.9±2.7

Table 3 shows the class-wise accuracies in terms of confusion matrices for the CNN classifiers using the spectrogram and mel-spectrogram feature representation as well as for two

Table 3: *Confusion matrices in voice quality classification from speech and NSA signals using the CNN classifier with the spectrogram and mel-spectrogram features and using two best reference SVM classifiers with feature sets 'All-glottal' and 'All-glottal+MFCCs'. Here B, M and P refer to breathy, modal and pressed voices, respectively.*

Feature set	**Speech** [%]				**NSA** [%]			
CNN classifier								
Spectrogram		B	M	P		B	M	P
	B	**96.2**	2.8	1.0	B	**95.9**	3.8	0.3
	M	5.3	**88.4**	6.3	M	2.5	**91.2**	6.3
	P	2.9	14.3	**82.8**	P	1.4	5.2	**93.4**
Mel-Spectrogram		B	M	P		B	M	P
	B	**95.9**	3.1	1.0	B	**95.4**	3.3	1.3
	M	4.2	**87.0**	8.8	M	2.4	**91.6**	6.0
	P	2.6	8.1	**89.3**	P	1.8	3.3	**94.9**
SVM classifier (based on [43])								
All-glottal		B	M	P		B	M	P
	B	**87.8**	9.6	2.6	B	**90.6**	6.3	3.1
	M	28.1	**55.8**	16.1	M	14.8	**71.2**	14.0
	P	7.7	9.2	**83.1**	P	2.2	6.6	**91.2**
All-glottal+MFCCs		B	M	P		B	M	P
	B	**88.4**	8.6	3.0	B	**92.4**	5.3	2.3
	M	22.5	**63.2**	14.3	M	12.3	**75.4**	12.3
	P	6.6	5.9	**87.5**	P	4.0	4.8	**91.2**

best reference systems based on hand-crafted feature sets 'All-glottal' and 'All-glottal+MFCCs'. It can be clearly seen that for the reference systems, there are confusions between modal and pressed voices, and between modal and breathy voices. These confusions exist both for the speech signal input and for the NSA signal input despite that the mean classification accuracy is larger for the latter input. On the other hand, there are less confusions between modal and pressed voices, and between modal and breathy voices for the CNN classifier, and this is true both for the speech and NSA inputs.

5. Conclusions

In this paper, the automatic classification of three voice quality classes (breathy, modal, pressed) was studied. The study addressed the use of simultaneously recorded NSA and speech signals together with a deep learning -based CNN classifier that used spectrogram and mel-spectrogram as input feature representations. Classification experiments revealed that the NSA signal has better capability to classify the three voice qualities compared to the acoustical speech signal. In addition, our experiments revealed that the CNN classifier using either the spectrogram or the mel-spectrogram feature representation is clearly more effective in the classification of voice qualities compared to using a SVM classifier with hand-crafted glottal source features and MFCCs. The best voice quality classification performance was achieved with mel-spectrogram as input to the CNN classifier (93.8% for the NSA signal and 90.6% for the speech signal). As the results show that both the acoustical speech signal and the NSA signal carry valuable information about voice quality, it is possible to try to further improve the classification performance by merging the information from these two signals, which will be a topic of our future studies.

6. Acknowledgements

This work was supported by the Academy of Finland (grant number 313390). The computational resources were provided by Aalto ScienceIT.

7. References

[1] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press, 1980.

[2] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[3] M. Pietrowicz, M. Hasegawa-Johnson, and K. G. Karahalios, "Acoustic correlates for perceived effort levels in male and female acted voices," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 792–811, 2017.

[4] I. R. Titze, "Principles of voice production (second printing)," *Iowa City, IA: National Center for Voice and Speech*, 2000.

[5] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[6] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *Proc. ICPhS*, 2003, pp. 2417–2420.

[7] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, no. 3, pp. 414–429, 2012.

[8] S. J. Park, A. Afshan, Z. M. Chua, and A. Alwan, "Using voice quality supervectors for affect identification," in *Proc. INTERSPEECH*, 2018, pp. 157–161.

[9] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression," in *Proc. INTERSPEECH*, 2018, pp. 1676–1680.

[10] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger, and C. Neuschaefer-Rube, "The contribution of phonation type to the

perception of vocal emotions in German: an articulatory synthesis study," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1503–1512, 2015.

[11] M. Ito, "Politeness and voice quality-the alternative method to measure aspiration noise," in *Proc. Speech Prosody*, 2004.

[12] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Voice quality and f0 cues for affect expression: implications for synthesis," in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 1849–1852.

[13] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[14] P. Ladefoged, I. Maddieson, and M. Jackson, *Investigating Phonation Types in Different Languages*. Vocal Physiology: Voice Production, Mechanisms and Functions, New York: Raven Press, 1988.

[15] J. Kuang and P. Keating, "Vocal fold vibratory patterns in tense versus lax phonation contrasts," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2784–2797, 2014.

[16] C. M. Esposito, "The effects of linguistic experience on the perception of phonation," *Journal of Phonetics*, vol. 38, no. 2, pp. 306–316, 2010.

[17] S. ud Dowla Khan, "The phonetics of contrastive phonation in Gujarati," *Journal of Phonetics*, vol. 40, no. 6, pp. 780–795, 2012.

[18] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Trans. Audio, Speech & Lang. Process.*, vol. 21, no. 6, pp. 1170–1179, 2013.

[19] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. INTERSPEECH*, 2007, pp. 1410–1413.

[20] P. Alku, J. Vintturi, and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Communication*, vol. 38, no. 3-4, pp. 321–334, 2002.

[21] D. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density," in *Proc. INTERSPEECH*, 2013, pp. 3206–3210.

[22] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[23] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 12, pp. 2281–2291, 2017.

[24] P. Alku, "Glottal inverse filtering analysis of human voice production-A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[25] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech and Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[26] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter - A new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[27] M. Swerts and R. N. J. Veldhuis, "The effect of speech melody on voice quality," *Speech Communication*, vol. 33, no. 4, pp. 297–303, 2001.

[28] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, "Modeling the voice source in terms of spectral slopes," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1404–1410, 2016.

[29] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. R. Gerratt, J. Neubauer, and A. Alwan, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," *The Journal of the Acoustical Society of America*, vol. 132, pp. 2625–2632, 2012.

[30] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, "The relationship between acoustic and perceived intraspeaker variability in voice quality," in *Proc. INTERSPEECH*, 2015, pp. 2357–2360.

[31] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs)," in *Proc. INTERSPEECH*, 2018, pp. 232–236.

[32] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Communication*, vol. 118, pp. 33 – 47, 2020.

[33] S. R. Kadiri and P. Alku, "Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. EL418–EL423, 2019.

[34] K. N. Stevens, D. N. Kalikow, and T. R. Willemain, "A miniature accelerometer for detecting glottal waveforms and nasalization," *Journal of Speech and Hearing Research*, vol. 18, no. 3, pp. 594–599, 1975.

[35] D. B. Rendon, J. L. R. Ojeda, L. F. C. Foix, D. S. Morillo, and M. A. Fernández, "Mapping the human body for vibrations using an accelerometer," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 1671–1674.

[36] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. Cheyne, and R. E. Hillman, "Using ambulatory voice monitoring to investigate common voice disorders: Research update," *Frontiers in Bioengineering and Biotechnology*, vol. 3, p. 155, 2015.

[37] D. D. Mehta, J. H. Van Stan, and R. E. Hillman, "Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 659–668, 2016.

[38] I. R. Titze, J. G. Svec, and P. S. Popolo, "Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues," *Journal of Speech Language and Hearing Research*, vol. 46, no. 4, pp. 919–932, 2003.

[39] R. F. Coleman, "Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice," *Journal of Voice*, vol. 2, no. 3, pp. 200–205, 1988.

[40] M. Ghassemi, J. H. Van Stan, D. D. Mehta, M. Zañartu, H. A. Cheyne II, R. E. Hillman, and J. V. Guttag, "Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1668–1675, 2014.

[41] M. Borsky, M. Cocude, D. D. Mehta, M. Zañartu, and J. Gudnason, "Classification of voice modes using neck-surface accelerometer data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5060–5064.

[42] Z. Lei, E. Kennedy, L. Fasanella, N. Y.-K. Li-Jessen, and L. Mongeau, "Discrimination between modal, breathy and pressed voice for single vowels using neck-surface vibration signals," *Applied Sciences*, vol. 9, no. 7, p. 1505, 2019.

[43] S. R. Kadiri and P. Alku, "Glottal features for classification of phonation type from speech and neck surface accelerometer signals," *Computer Speech & Language*, vol. 70, p. 101232, 2021.

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[45] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*. MIT press Cambridge, MA, USA, 2017, vol. 1.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations*, 2015.