CLASSIFICATION OF VOICE MODES USING NECK-SURFACE ACCELEROMETER DATA

Michal Borsky[1], *Marion Cocude*[1], *Daryush D. Mehta*[2], *Matías Zañartu*[3], *Jon Gudnason*[1]

[1]Center for Analysis and Design of Intelligent Agents, Reykjavik University, Reykjavik, Iceland
[2]Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital, Boston, MA
[3]Department of Electronic Engineering, Universidad Tecnica Federico Santa Marıa, Valparaıso, Chile

ABSTRACT

This study analyzes signals recorded using a neck-surface accelerometer from subjects producing speech with different voice modes. The purpose is to explore if the recorded waveforms can capture the glottal vibratory patterns which can be related to the movement of the vocal folds and thus voice quality. The accelerometer waveforms do not contain the supraglottal resonances, and these characteristics make the proposed method suitable for real-life voice quality assessment and monitoring as it does not breach patient privacy. The experiments with a Gaussian mexture model classifier demonstrate that different voice qualities produce distinctly different accelerometer waveforms. The system achieved 80.2% and 89.5% for frame- and utterance-level accuracy, respectively, for classifying among modal, breathy, pressed, and rough voice modes using a speaker-dependent classifier. Finally, the article presents characteristic waveforms for each modality and discusses their attributes.

Index Terms— voice mode classification, laryngeal accelerometer, GMM, MFCC

1. INTRODUCTION

Voice quality assessment (VQA) is defined subjectively through listening tests using one of several auditory-perceptual protocols [1, 2]. For example, the Consensus Auditory-Perceptual Evaluation of Voice CAPE-V protocol seeks to document an individual's voice quality along several dimensions, including roughness, breathiness, strain, pitch, and loudness deviation. Clinical voice specialists evaluate subjects using a visual analog scale. Automatic VQA systems are often designed to infer the same distortion parameters [3, 4]. Although speech is the most readily available signal, alternative measurement methods may be more suitable for objective voice quality assessment if they were to more directly capture voice source characteristics. The electroglottogram captures an indirect estimate of the vocal fold contact area, which is related to the voice source. Inverse filtering the oral airflow or acoustic microphone speech signal is more closely related to the voice source signal that enters the vocal tract, where it is modulated to carry the linguistic articulatory content and radiates from the lips as the acoustic speech signal.

Better insight into clinical problems can be gained by analyzing signals that can be more directly associated with vocal fold dynamics. Some physiological parameters which contribute to the overall quality of the source signal are stiffness and thickness of the vocal folds, their abduction or adduction, elevation of the larynx and the constriction of supraglottal structures. These structural differences cause distinct movement patterns in the vocal folds, which in turn produce distinct phonation types. However, the approach also requires a sophisticated measuring device, professional manipulation, and laboratory settings in order to obtain proper waveforms.

The surface accelerometer [5] is an indirect measurement of vocal fold aerodynamics [6]. The mechanical vibrations which naturally occur during phonation are transmitted as sound waves through the trachea to the neck surface. Previous works have demonstrated that accelerometer signals contain both glottal and subglottal vibratory patters [7], but no supraglottal resonances when positioned appropriately. Accelerometer is also robust against background acoustic noises. These attributes makes a neck-surface accelerometer ideal for real-life voice quality assessment and monitoring.

This study follows on our work with EGG [8] that showed that distinct phonation modes produce distinct waveforms. This work studies the mechanical vibrations which occur during phonation and also, the number of participants within this study was greater. A similar approach was presented in [9, 10], but all of these studies worked with speech or inverse-filtered speech. The authors in [11] have also demonstrated that accelerometer signals can be used to differentiate vocal hyperfunction from normal patterns of vocal behavior. This is one of the few studies that analyzes the use of neck-surface accelerometer in the task of voice modality classification. The study provides examples of prototype waveforms and compares them with EGG waveforms, which are much more researched.

The article is structured as follows. Section 2 describes the data acquisition protocol. Section 3 describes the performed classification task and experimental setup. Section 4 presents the results and relates the reached conclusion to findings reached on other speech-related signals. The conclusion is summarized in Section 5.

2. DATABASE

The presented experiments were performed on the "InLab" part of Ambulatory Voice Monitoring database, which contains recordings collected in acoustically controlled conditions [12]. The signals were acquired for normal and pathological subjects, but only normal participants were used for the purpose of this study. The database contains the electroglottograph, accelerometer, acoustic, oral air pressure, and oral air flow signals. All signals were sampled at f_s = 20 kHz, time-synchronized and amplitude-normalized. The neck-surface accelerometer [13] was attached under the thyroid prominence and above the collarbone. The accelerometer was an off-the-shelf single-axis sensor (BU-27135, Knowles Electronics, Itasca, IL).

All subjects were verified to have normal vocal status using laryngeal endoscopy and auditory-perceptual judgment by a speech-language pathologist. Each subject recorded sustained vowels (/a/, /e/, /i/, /o/, /u/). The vowels were recorded at a comfortable pitch and loudness (modal voice), and in three non-modal voice qualities: breathy, pressed, and rough [14]. Four male and 24 female subjects were selected.

3. CLASSIFICATION TASK SETUP

The accelerometer signals can be initially characterized by their short-time discrete Fourier transform. As a consequence, our feature extraction setup used standard Mel-frequency cepstral coefficients (MFCCs) that have been proven to model voice pathology using speech signals [15, 16, 17, 18]. The frame length was 32 ms with a frame shift of 16 ms. The Mel-filter bank contained 32 filters and f_{min} was set to 50 Hz and f_{max} to 4000 Hz. The feature vector contained only static MFCCs with 0^{th} coefficient. We did not employ any voice activity detection (VAD) prior to feature extraction to cut out the silence frames. Table 1 summarizes the total number of signals and extracted frames for each voice mode. The classification task was performed using a Gaussian mixture model (GMM) classifier with two mixtures. The GMM parameters $\Theta = \{\pi, \mu, \Sigma\}$ were initialized for each class separately using the maximum-likelihood method and then re-estimated using the expectation-maximization algorithm. The choice to use only two mixtures was motivated by our previous decision of not using VAD. One mixture modeled the distribution of voice frames while the second mixture modeled the outlying data, which were identified to belong to the silence frames. The GMMs were described by its full covariance matrices Σ.

The evaluation was carried out with a four-class classification approach. The speaker-dependent classification was always done by holding out one utterance while training the models on the rest and then repeating the whole process for all signals. The results were evaluated in terms of classification accuracy [%] at either frame or utterance level. The utterance-level accuracy was based on performing hard classification at the frame level and selecting the most frequently occurring class. This approach gave insight into the distribution of misclassified frames within each token.

Table 1. *Number of signals and frames for each voice mode*

	Modal	Breathy	Pressed	Rough
Signals	351	172	196	213
Frames	78 298	31 166	28 093	30 620

4. RESULTS AND DISCUSSION

The initial analysis was focused on determining the optimal number of MFCCs for the given number of mixtures. Figure 1 summarizes the frame- and utterance-level results starting with just a single, 0^{th} coefficient and ending with the the total number of 30 MFCCs. The initial values with just the 0^{th} coefficient were relatively good at 60.7% and 68.4% for the frame- and utterance-level respectively. The accuracy curves begin to rise sharply and then plateau for approximately 6-10 MFCCs. The optimal number of MFCCs for the voice mode classification task using the neck-surface accelerometer was found to be 16, for which the frame- and utterance-level accuracy reached 80.2% and 89.5%, respectively. Finally, the curves begin to fall for more than than 16-20 MFCCs.

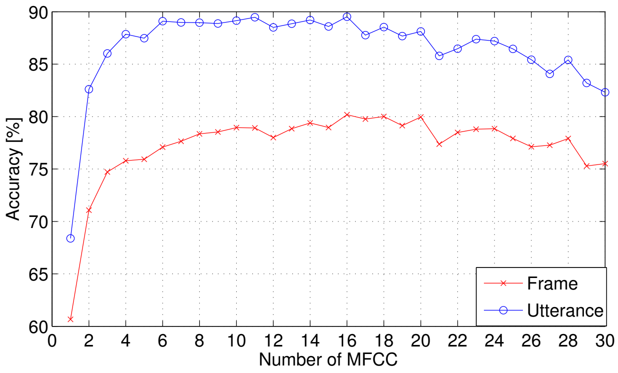

Fig. 1. *Accuracy [%] for different number of MFCC.*

A closer look at the nature of errors for misclassified utterances revealed that the posterior values of frames belonging to the (incorrectly) assigned class was ≥ 0.75 for roughly one half of them. This behavior shows that the classifiers displayed a trend to misassign the frames, and thus whole utterances, predominantly into a single class. This trend might indicate that the subjects did not produce a token with a pure voice mode (in practice, speakers typically exhibit degrees of all non-modal voice types when attempting to produce a vowel with only one voice mode). On the other hand, about a quarter of frames had a uniform distribution of posteriors across the classes, which demonstrated the limitations of the proposed classification method and accelerometer data.

The current analysis would indicate than just the waveform energy, which is incorporated into the 0^{th} coefficient, carried most of the information about the voice mode. However, in our subsequent experiment, we removed this coefficient out of the feature vector and performed the classification task once again. Table 2 summarizes the best achieved results for feature vectors with and without the 0^th coefficient. It can be noted that this feature vector achieved practically the same frame- and utterance-level accuracy. The only difference was that more coefficients were needed to reach the optimal performance. As a results, all further analysis will be presented for feature vectors which contained c[0].

Table 2. *Influence of c[0] coefficient on accuracy*

	Opt. num.	Acc. [%]	
	of MFCC	Frame	Utterance
with c[0]	16	80.2	89.5
without c[0]	18	80.1	89.9

The confusion matrix for the frame-level accuracy is summarized in Table 3 and there are several interesting things that can be taken from it. First, the highest classification accuracy of 88.3% was achieved for the breathy voice mode. It was then followed by the modal, rough, and pressed voice modes. Taking a closer look at the confusion rates between voice mode revealed that the pressed and rough types were far more often confused with each other than with either breathy or modal. The misclassification rates reached 14.9% for pressed being classified as rough and 13.2% the other way around. This observation leads to the conclusion that the physiological processes which generated rough and pressed voice modes produced similar vocal fold vibratory properties captured by the accelerometer as similar waveforms. A similarly strong trend was not observed between the breathy and modal types, when only the modal voice showed statistically significant preference for breathy type. The utterance-level confusion matrix is summarized in Table 4, and the results there follow the same trends as from the utterance-level classification.

Table 3. *Frame-level accuracy [%] matrix.*

		Recognized			
		B	M	P	R
Actual	B	**88.3**	4	4.3	3.4
	M	7.6	**86.2**	3.6	2.6
	P	6.3	8.2	**70.6**	14.9
	R	6.5	4.6	13.2	**75.7**

The prototype waveforms were estimated as a likelihood weighted average from all frames for the best performing speaker in the database, using the following formula:

Table 4. *Utterance-level accuracy [%] matrix.*

		Recognized			
		B	M	P	R
Actual	B	**95.9**	0.6	3	0.5
	M	3.7	**93.7**	2	0.6
	P	3.6	5.1	**80.6**	10.7
	R	1.4	1.4	9.4	**87.8**

$$\mathbf{O}_{proto(j)} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(\mathbf{O}_i|j)}{\sum_k p(\mathbf{O}_i|k)} \mathbf{O}_i, \tag{1}$$

where $\mathbf{O}_{proto(j)}$ is the prototype observation vector for class j, and $p(\mathbf{O}_i|j)$ is the probability of vector $\mathbf{O}_i$ being generated by class j. The first problem was a difference in pitch across frames and the signals. A closer look revealed that the analyzed subjects produced vowels which had a fairly similar average fundamental frequency (f_o). The relative variance reached 2.1±1.7, 3.3±2.9, 5±7.3 and 11.7±9.2 [%] for the breathy, modal, press and rough voice mode respectively. We chose to report relative values rather than the absolutes in *Hz* in order to penalize subjects with low f_o, as the error in f_o estimation was more severe for them. The only notable difference was the rough voice mode, which displayed a 11.7% relative difference. This observation was in line with the fact that rough voices often lack a periodic structure.

Fundamental frequency analysis also indicates that f_o varied significantly across frames for which signals had to be synchronized in order to obtain prototype waveforms. The subharmonic-to-harmonic pitch detection algorithm [19] was used to estimate the f_o within each frame. The frames had their time axis stretched or compressed to accommodate for the difference between the reference and current frame. A secondary problem was the difference in phase, which was effectively solved using the autocorrelation function. The estimate was done using frames from utterances which were labeled *a priori* as belonging to that particular class. The frames were also averaged in relation to their number so that the resulting waveforms could be compared in terms of their amplitude. Figure 2 illustrates the estimated prototype waveforms for each voice mode for the best performing speaker, and Figure 3 illustrates the respective power spectra.

The breathy, modal, and pressed voice styles displayed a clearly periodic structure. The period for modal voice was approximately 5 ms and approximately 6 ms for the breathy and pressed speech. The rough voice mode displayed the highest period of about 7.5 ms. Voice modes also differed in their amplitudes as the rough and breathy types, especially, had a relatively small amplitude. The low-amplitude behavior for breathy style was somewhat expected. Its physiological process is characterized by a minimal vibration of the vocal folds, which translates into a low-energy vibration radiated through the neck tissue. The measured waveform also lacked any distinct shape aside from the onset and offset slopes. The modal and pressed types were characterized by significant peaks which separated the glottal periods. The spectrum of the pressed waveforms was much richer in higher harmonics than other waveforms. This observation was consistent with general characteristics of the pressed voice mode. The rough waveform was similar in shape to the pressed type but also contained multiple significant peaks in a single period.

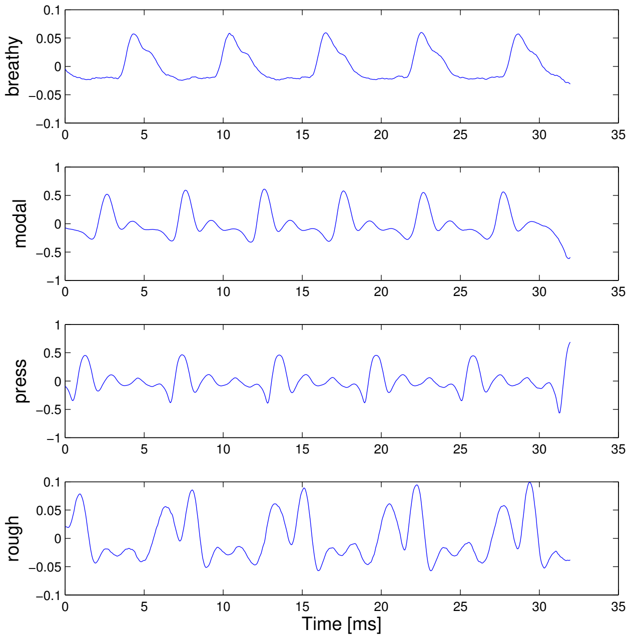

Fig. 2. *Likelihood-averaged accelerometer waveforms for each voice mode.*

It is acknowledged that participants potentially produced varying degrees of the prompted voice qualities. Recordings were thus screened by an expert listener with no prior knowledge of prompted quality using the CAPE-V protocol to obtain dichotomous perception labels. Cohen's κ was 0.6 (good-to-strong agreement) between the prompted and perception labels. Future work warrants a formal auditory-perceptual evaluation of the tokens.

5. CONCLUSION

This article presented a analysis of subglottal neck-surface accelerometer signals during voicing. Results demonstrated that vocal fold vibratory patterns were transferred to the neck surface and adequately captured by the accelerometer for the purpose of voice modality classification. The speaker-specific system achieved 80.2% and 89.5% accuracy at the frame and utterance level, respectively. Pressed and rough voice modes were predominantly confused with each other. No such trend was found between breathy and modal voice modes. The article also presented prototype waveforms for each modality, which were computed as the likelihood-weighted waveforms. Some of the estimated accelerometer waveforms displayed characteristics that have been previously described for acoustic or electroglottograph waveforms. The breathy waveform was characterized by its low amplitude and steep spectral rolloff, and the spectrum of the pressed waveform contained significant higher harmonics. These findings demonstrate the potential of using accelerometer sensors for voice quality assessment in naturalistic environments.

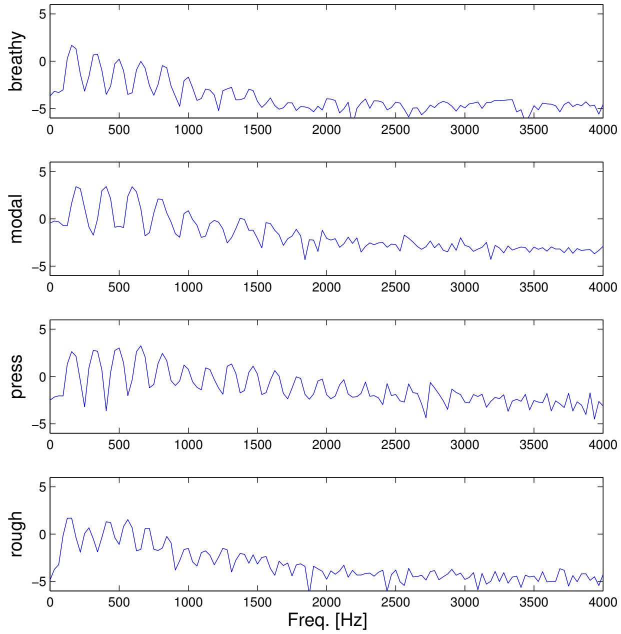

Fig. 3. *Power spectra of the prototype waveforms for each voice mode.*

6. ACKNOWLEDGMENT

This work is sponsored by The Icelandic Centre for Research (RANNIS) under the project *Model-based speech production analysis and voice quality assessment*, Grant No 152705-051. The work was performed during Marion Cocude's Erasmus+ traineeship at Reykjavik University. The authors thank J. H. Van Stan for data collection and pereptual evaluation. Additional support received from the Voice Health Institute and the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders under Grant R33 DC011588 (PI: R. E. Hillman). The paper's contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

7. REFERENCES

[1] Hirano Minoru and Karen R. McCormick, “Clinical examination of voice,” *The Journal of the Acoustical Society of America*, vol. 80, no. 4, October 1986.

[2] Gail B. Kempster, Bruce R. Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, , and Robert E. Hillman, “Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol,” *American Journal of Speech Language Pathology*, vol. 18, no. 2, pp. 124–132, May 2009.

[3] T. Villa-Cañas, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and J. I. Godino-Llorente, “Automatic assessment of voice signals according to the GRBAS scale using modulation spectra, mel frequency cepstral coefficients and noise parameters,” in *Symposium of Signals, Images and Artificial Vision - 2013: STSIVA - 2013*, Sept 2013, pp. 1–5.

[4] Zhijian Wang, Ping Yu, Nan Yan, Lan Wang, and Manwa L. Ng, “Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the GRBAS scale,” *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 241–251, 2016.

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

[6] 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, April 2016.

[7] M. Nolan, B. Madden, and E. Burke, “Accelerometer based measurement for the mapping of neck surface vibrations during vocalized speech,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sept 2009, pp. 4453–4456.

[8] M. Borsky, D. D. Mehta, J. P. Gudjohnsen, and J. Gudnason, “Classification of voice modality using electroglottogram waveforms,” in *Proceedings of Interspeech*, Sept 2016, pp. 1–5.

[9] M. Lugger, F. Stimm, and B. Yang, “Extracting voice quality contours using discrete hidden markov models.,” in *Proceedings of Speech Prosody*. 2008, pp. 29–32, ISCA.

[10] John Jon Kane and Christer Gobl, “Identifying regions of non-modal phonation using features of the wavelet transform.,” in *Proceedings of Interspeech*. 2011, pp. 177–180, ISCA.

[11] M. Ghassemi, J. H. Van Stan, D. D. Mehta, M. Za nartu, H. A. Cheyne, 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 Bio-Medical Engineering*, vol. 61, no. 6, pp. 1668–1675, 2014.

[12] Daryush D. Mehta, J. H. Van Stan, Matias Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Corté, A. H. Cheyne II, and Robert E. Hillman, “Using ambulatory voice monitoring to investigate common voice disorders: Research update,” *Frontiers in Bioengineering and Biotechnology*, vol. 3, no. 155, pp. 1–14, 2015.

[13] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, “Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 11, pp. 3090–3096, Nov 2012.

[14] Bruce R. Gerratt and Jody Kreiman, “Toward a taxonomy of nonmodal phonation,” *Journal of Phonetics*, vol. 29, no. 4, pp. 365 – 381, 2001.

[15] J.I. Godino-Llorente, Rubén Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, “Automatic detection of voice impairments from text-dependent running speech,” *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 176 – 182, 2009.

[16] R. Fraile, N. Sáenz-Lechón, J. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, “Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex,” *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 146–152, 2009.

[17] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam, “Vocal fold disorder detection based on continuous speech by using mfcc and gmm,” in *GCC Conference and Exhibition (GCC), 2013 7th IEEE*, Nov 2013, pp. 292–297.

[18] Shaheen N. Awan et al., “Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the cape-v,” *Clinical Linguistics & Phonetics*, vol. 24, no. 9, pp. 742–758, 2010.

[19] Xuejing Sun, “A pitch determination algorithm based on subharmonic-to-harmonic ratio,” in *the 6th International Conference of Spoken Language Processing*, 2000, pp. 676–679.