INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Voice Quality and Between-Frame Entropy for Sleepiness Estimation

Vijay Ravi, Soo Jin Park, Amber Afshan, Abeer Alwan

Department of Electrical and Computer Engineering, University of California Los Angeles, USA
vijaysumaravi@ucla.edu, sj.park@ucla.edu, amberafshan@g.ucla.edu, alwan@g.ucla.edu

Abstract

Sleepiness monitoring and prediction has many potential applications, such as being a safety feature in driver-assistance systems. In this study, we address the CompareE 2019 Continuous Sleepiness task of estimating the degree of sleepiness from voice data. The voice quality feature set was proposed to capture the acoustic characteristics related to the degree of sleepiness of a speaker, and between-frame entropy was proposed as an instantaneous measure of the speaking rate. An outlier elimination on the training data using between-frame entropy enhanced the system robustness in all conditions. This was followed by a regression system to predict the degree of sleepiness. Utterances were represented using i-vectors computed from voice quality features. Similar systems were also developed using mel-frequency cepstral coefficients and the ComParE16 feature set. These three systems were combined using score-level fusion. Results suggested complementarity between these feature sets. The complete system outperformed the baseline system which used the ComParE16 feature set. A relative improvement of 19.5% and 5.4% was achieved on the development and the test datasets, respectively.

Index Terms: voice quality, computational paralinguistics, sleepiness, entropy

1. Introduction

Assessing sleepiness is crucial in monitoring the level of alertness of a person in critical missions or life-threatening activities, such as in aviation or naval missions. Speech signals can be effective to assess degree of sleepiness in such situations because sleepiness is reflected in voice, and speech data can be collected unobtrusively [1]. This study aims to automatically assess the degree of sleepiness, as a participation in the Interspeech 2019 Continuous Sleepiness sub-challenge [2].

Several automatic sleepiness detection systems have been proposed. For example, [3] applied acoustic features such as F_0 contours for assessing sleepiness, and used the statistics of those features to represent an utterance. In [4], mel-frequency cepstral coefficients (MFCCs) were used with the hidden Markov model (HMM). [5] used prosodic and spectral features and suggested a feature selection method to obtain the vector representation. While these systems focused on binary classification between sleepy and not sleepy, the task in the current study is to estimate the degree of sleepiness.

When an individual is deprived of sleep and fatigued, the resulting cognitive-physiological changes can influence voice production. For example, sleepiness reduces the cognitive speech planning ability and speed, which might result in slowed speech; muscle tension decreases, which might lead to a lower fundamental frequency (F_0), lower formant frequency positions and broader formant bandwidths. Potential acoustic changes in speech induced by sleepiness were summarized in [6]. In that study, it was reported that F_0, and the first formant frequency were lower for the sleepy speech than for the alert speech sample. Another study on the effects of sustained wakefulness on speech found that various aspects of speech, including speaking rate, F_0 variation and the spectral tilt of the source spectrum were sensitive to sleepiness [7].

In this study, two feature sets are proposed to assess sleepiness: voice quality and between-frame entropy. The first feature set was inspired by a psycho-acoustic model of voice quality [8, 9]. This feature set effectively represented speakers' identity [10, 11, 12, 13] and emotional/psychological state [14, 15]. This set might also be effective in representing sleepiness; in that the acoustic features, such as F_0, formant frequencies, formant bandwidths, source spectral tilt, and inharmonic noise, often associated with sleepiness, overlap with this feature set. The second feature, between-frame entropy was used as an instantaneous measure of speech rate which is also often associated with sleepiness. Entropy is large when the spectral characteristics vary rapidly between frames. In this sense, between-frame entropy can be assumed to reflect instantaneous speech rate. Because it is computed at the frame level, it might provide information about the speaking rate with high time resolution.

We used the i-vector framework [16] for utterance representation. In this framework, each utterance is represented with a Gaussian mixture model (GMM) representing the feature distribution within an utterance. The GMM is often adapted from a universal background model (UBM), a statistical model for speech sounds, usually trained on a larger data corpus. The mean vectors of the adapted mixture model are concatenated and decomposed to a low dimensional representation. This low dimensional representation is called the i-vector. Because the i-vector effectively summarizes the feature distribution of an utterance, it has been widely used in various speech processing applications, including automatic speaker verification [17].

Additionally, an outlier elimination is applied prior to training the sleepiness prediction system. It has been noted that the effect of sleepiness on speech is highly idiosyncratic, and speech task-specific [18, 7]. High degree of fatigue might also result in an unexpected behavior that does not necessarily represent sleepiness. If some speech samples have considerably different characteristics from others, they might make it difficult for the system to learn a general pattern. In this context, detection and removal of outliers in the training dataset was applied to improve the system robustness.

The rest of the paper is organized as follows. In Section 2, the databases used in this study are described. The proposed acoustic features to represent sleepiness, and the sleepiness prediction system are presented in Section 3 and 4, respectively. In Section 5, the experimental results are discussed. The paper concludes with future work in Section 6.

http://dx.doi.org/10.21437/Interspeech.2019-2988

2. Database

2.1. The SLEEP Corpus

For the Continuous Sleepiness Sub-Challenge, a subset of the Duesseldorf Sleepy Language corpus collected from German speakers was used [2]. Audio recordings were obtained with a sampling rate of 44.1 kHz and were down-sampled to 16 kHz. The dataset consists of both read speech and spontaneous narrative speech.

Speakers reported their sleepiness on the Karolinska Sleepiness Scale (KSS, [19]) with a range of 1 (extremely alert) to 9 (very sleepy). Additionally, two observers assigned *post hoc* KSS ratings. The self-assessed ratings by the speakers and the ratings from the observers were averaged to form the reference degree of sleepiness.

2.2. Databases for Training the i-vector Extractor

Training a UBM and an i-vector extractor requires a database containing a large amount of recordings from multiple speakers. The NIST SRE 04, 05, 06, and 08 databases [20, 21, 22] and the Switchboard II corpus phase 2 data [23] were used. These databases provide more than 3,000 hours of speech samples in multiple languages from 3,408 female and 1,832 male speakers. The sampling rate of these recordings is 8 kHz.

3. Feature Extraction and Representation

3.1. Voice Quality Features

The voice quality feature set used in this study, denoted as *VQual*, includes the fundamental frequency (F_0); the first three formant frequencies (F_1, F_2, F_3) and their corresponding amplitudes (A_1, A_2, A_3); harmonic amplitude differences (H_1-H_2, H_2-H_4, H_4-$H_{2\mathrm{k}}$) that represent spectral tilt of source spectrum; and cepstral peak prominence (CPP, [24]) which is a measure for inharmonic noise. Here, H_1, H_2, H_4, and $H_{2\mathrm{k}}$ indicate the amplitudes (in dB) of first, second, fourth harmonics, and the harmonic nearest to 2 kHz, respectively. The *VQual* feature set includes acoustic features that are often associated with sleepiness. The first and the second derivatives of these features were also used. Features were extracted using the VoiceSauce toolkit [25].

3.2. Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) are widely used acoustic features in speech processing. MFCCs were extracted with a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with coefficient 0.97, and a sinusoidal lifter with coefficient 22. A filter bank with 23 filters was applied and 13 coefficients were extracted. The first and the second derivatives were also used.

3.3. Entropy

Between-frame entropy was calculated as described in [26]. MFCCs were obtained from the speech signal using a Hamming window of length 25 ms and a frame shift of 2.5 ms. A 30 ms rectangular window was applied to MFCCs and using the features in this window, the signal's local entropy was computed as:

$$H(v) = K \ln \sqrt{2\pi} + \ln \mathbf{Tr}(\Sigma), \tag{1}$$

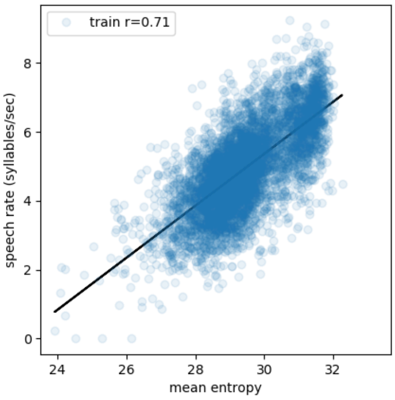

Figure 1: *Scatter plot of speech rate in terms of syllables per second vs. mean of frame level change in entropy within an utterance in the training dataset. A line was fitted using linear regression ($R^2 = 0.50$, $p < 0.001$).*

where, $H(v)$ is the entropy of the random variable v of dimension K and Σ is the $K \times K$ covariance matrix of the probability distribution function of the random variable v.

The spectral variability of the speech signal is represented by entropy. Rapid information gain in the speech spectrum corresponds to high entropy. Hence, we expected between-frame entropy to correlate with instantaneous speech rate. In order to verify this hypothesis, the correlation between the number of syllables per second, a conventional speech rate measure, and the mean between-frame entropy within an utterance was calculated. Syllables per second were computed using a Praat script as described in [27].

The computed speech rate was highly correlated with the mean frame-level change in entropy for the training data, as shown in Figure 1. The linear regression between the mean of speech rate and the mean of entropy resulted in $R^2 = 0.50$ and $p < 0.001$. The corresponding Pearson's correlation coefficient was 0.71. The high correlation provides evidence to the hypothesis that between-frame entropy can be used to represent speech rate. The unexplained variance in the linear regression might be due to the difference in information reflected in each methods. For example, the syllables per second measure only focuses on syllable nuclei, while the between-frame entropy changes over all frames in an utterance.

3.4. ComParE16 Feature Set

This feature set is provided as the baseline of the Interspeech 2019 Continuous Sleepiness sub-challenge. The ComParE16 feature set [28] consists of F_0, energy, spectral, cepstral and voicing related frame-level features. Additionally, the set includes the zero crossing rate, jitter, shimmer, harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. These are referred to as low-level descriptors (LLDs). The OpenSMILE toolkit was used to extract the ComParE16 features [29].

3.5. Utterance Representation

3.5.1. i-vector

i-vectors were used to represent the distribution of frame-level acoustic features. A universal background model (UBM, [30]) with 2,048 Gaussian mixtures was trained using the SRE and the Switchboard databases by applying the expectation maximization(EM) algorithm. Then, for each utterance, the UBM is adapted based on the eigenvoice adaptation technique [31]. The output of this is a 600-dimensional i-vector that is centered and length normalized. We followed the approach described in [16] to extract the i-vectors.

Speech signals in the SLEEP database were downsampled to a sampling rate of 8 kHz to match the bandwidth of the training databases prior to i-vector extraction.

3.5.2. Statistics Vector

Statistics vectors were used as a baseline utterance representation that maps the contours of the acoustic features onto a fixed dimensionality vector. Peaks, percentiles, moments, regression, temporal and modulation functionals were computed [28]. The OpenSmile toolkit [29] was used to compute the statistics vector.

4. Sleepiness Predictor

4.1. Implementation

As discussed in Section 1, unexpected behaviours of the speaker or the speaking style could both introduce high degree of undesired variability in the data resulting in outliers. It is well known that outliers can cause a degradation in regression tasks [32] which may considerably reduce system performance. To address this issue, the training data was pre-processed to eliminate outliers.

An elliptic envelope outlier detection approach using covariance estimation was used [33] in this paper. This method assumes Gaussian data and learns an ellipse. The covariance estimation is expected to be robust to outliers. This algorithm uses contamination factor to control the acceptable extent of variability between the outliers and the inliers of the data. The contamination factor is a representation of the proportion of outliers in the data. The outlier detector used between-frame entropy based i-vectors as features.

Support vector regression (SVR) [34] was used to model the features in order to predict the degree of sleepiness. Since, the feature sets used in this work have complementary information, we built separate models for each feature set. We then performed a system level fusion on the prediction scores obtained using individual feature sets. The final prediction was a linear combination of predictions obtained from individual features.

4.2. Experimental Setup

4.2.1. Evaluation Metric

The task in this challenge is to assess the sleepiness of a speaker as a regression problem. Spearman's correlation coefficient (ρ) was used for performance evaluation [2]. Spearman's correlation coefficient was selected among various correlation measures because of the robustness of this measure. For example, the Pearson correlation coefficient measures the strength of linear relationships between normally distributed variables [35], but the predictions in our experiments may neither be normally distributed nor have a linear relationship.

4.2.2. Predictor Performance with Individual Feature Sets

Four different configurations for predictors using between-frame entropy, VQual or MFCCs were made: with vs. without outlier elimination, and statistics vector vs. i-vector. In outlier elimination, 15% contamination was assumed based on a preliminary analysis on individual system performance. The complexity of the SVR predictor, C, was chosen from a range of values between 10^{-6} and 10^{5} so as to maximize the system performance on the development dataset. Then, the effects of outlier elimination and i-vector representation were analyzed to decide system configurations for individual feature sets.

A decision upon whether the outlier elimination should be used or not, and another decision between using the i-vector and statistics vector representations were made by selecting the best performing system configuration for each individual feature set.

4.2.3. Effects of System Fusion

Using the configurations decided for individual systems, the complementary effect among different features were tested by score level fusion. Individual systems that performed reasonably well ($\rho > 0.2$) were selected. Then, all possible combinations were made among selected systems along with the baseline system which used ComParE16 and statistics vector representation. The weights for the fusion were determined by a grid search.

5. Experimental Results

5.1. Individual System Performance

The effectiveness of eliminating the outliers in improving the robustness of the system was analyzed. A comparison between the use of all of the training data vs using the training data after removing outliers is shown in Table 1. The results in terms of Spearman's correlation coefficient for entropy, VQual and MFCCs are shown using both statistics vector and i-vector to represent the utterances. The results obtained after outlier elimination (ρ_2) are compared against the results using all of the training data (ρ_1).

There is a consistent improvement in system performance by eliminating outliers from the training data. For example, when statistics vectors were used as utterance representation, outlier elimination improved the ρ value from 0.158 to 0.192 for MFCCs and from 0.142 to 0.178 for VQual. Similarly, in the i-vector framework, outlier elimination improved the results for MFCCs from 0.248 to 0.252 and for VQual from 0.201 to 0.221. Thus, the configuration with outlier elimination was selected for all individual systems.

Table 1: *System performance on individual features for the development set. The best performing configuration for each individual feature set is boldfaced.*

Utt. representation	Feature set	ρ_1	ρ_2
Statistics vector	entropy	0.029	0.078
	VQual	0.142	0.179
	MFCC	0.158	0.192
i-vector	entropy	0.126	**0.131**
	VQual	0.201	**0.221**
	MFCC	0.248	**0.252**

The correlation between predicted degree and reference degree of sleepiness notably increased when the i-vector framework was used for utterance representation as compared to the statistics vector. This was observed for every acoustic feature. For MFCCs, for instance, the performance improved from 0.158 to 0.248, a relative improvement of approximately 56%. When VQual features were used the improvement was around 40% (from $\rho = 0.142$ to 0.201). Therefore, the i-vector representation was chosen instead of statistics vector for each individual system.

Although VQual and MFCC-based systems performed reasonably well ($\rho > 0.2$) with the selected configuration, the entropy-based system did not perform as well. This could potentially be due to the varying effects of sleepiness on speech rates between read and spontaneous speech. It was found that speakers' sleepiness was negatively correlated with speech rate for read speech, but an opposite pattern was observed for spontaneous speech [7]. Unfortunately, no further analysis on the type of speech task could be made because such metadata was not available for the dataset used in this study.

5.2. Fused System Performance

Based on the results of the individual systems, VQual and MFCC-based systems using the i-vector representation with an outlier elimination prior were selected for fusion. As mentioned earlier, the baseline system ($\rho = 0.251$) using ComParE16 and statistics vector was also used for fusion. Table 2 presents the results for score fusion among the three individual systems. Since outlier elimination provided better results, all experiments of score fusion were done with prior elimination of outliers in the training data.

Table 2: *Results for score-level fusion on the development dataset.*

Feature sets	ρ
ComParE16+VQual	0.283
ComParE16+MFCCs	0.296
MFCCs+VQual	0.265
ComParE16+MFCCs+VQual	**0.300**

Score fusion with VQual improved performance for all cases. For example, when VQual was fused with ComParE16, the ρ score increased from 0.251 for ComParE16 alone to 0.283 for the fused system (a relative improvement of 12%). Fusing with MFCCs also improved system performance ($\rho = 0.265$) when compared to the MFCC-alone system ($\rho = 0.252$). These results suggest that VQual provides complementary information about sleepiness to the MFCC and ComParE16 feature sets.

The best performance over all combinations of individual systems was when ComParE16, MFCCs and VQual features were fused altogether, resulting in the correlation score of 0.300. Compared to the baseline system, a relative improvement of 19.5% was achieved. This system was used as the complete system.

5.3. Performance on the Test Dataset

Degree of sleepiness was predicted on the test dataset using the complete system on the development dataset. For the test phase, training and development datasets were combined to train the predictor. Outlier elimination was performed on this combined data followed by a score level fusion of individual predictors trained using ComParE16, MFCCs and VQual feature sets. Statistics vector representation for ComParE16 and i-vector representation for MFCCs and VQual were used.

Table 3: *Complete system performance on the development and the test set.*

	Development	**Test**
ComParE16 (baseline)	0.251	0.314
ComParE16+MFCCs+VQual	0.300	0.331

The evaluation results of the complete system on development and test datasets are summarized in Table 3 in comparison to the baseline system. On the test data, the proposed system outperformed the baseline system with an improvement of 5.4% ($\rho = 0.314$ to 0.331).

6. Conclusion

This paper presents a systematic approach to estimate the degree of sleepiness from voice data. Eliminating outliers in the training samples prior to training the predictor provided a substantial performance gain regardless of the acoustic feature set and utterance representation used for the predictor. Between-frame entropy was introduced as an instantaneous measure of speech rate. It was used to detect outliers to develop a robust sleepiness prediction system. The i-vector framework effectively represented the feature distribution of an utterance. This is reflective in the improvement of the system performance over the baseline utterance representation using acoustic feature statistics. Voice quality features improved system performance when fused with any system based on other features, suggesting a complementary effect between features sets. The complete system, a fusion of individual systems based on VQual, MFCCs, and ComParE16 features, outperformed the baseline system both on the development and test datasets.

Although the proposed system in this paper outperformed the baseline system, there is room for improvement. For example, the UBM and i-vector extractor can be trained with a German speech database for a more reliable utterance representation. Using i-vector representation for ComParE16 might provide further performance gain. Feature selection or dimensionality reduction methods can be applied to analyze and compare the effect of individual features on system performance. An adaptive strategy to compensate for the effects of speaking style and speaker variability would be another promising approach, considering that the influence of sleepiness on speech varies based on those factors.

7. References

[1] C. A. de Vasconcelos, M. N. Vieira, G. Kecklund, and H. C. Yehia, "Speech analysis for fatigue and sleepiness detection of a pilot," *Aerospace Medicine and Human Performance*, vol. 90, no. 4, pp. 415–418, 2019.

[2] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder,

E. Bergelson10 *et al.*, “The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity,” *Interspeech*, 2019.

[3] T. Rahman, S. Mariooryad, S. Keshavamurthy, G. Liu, J. H. Hansen, and C. Busso, “Detecting sleepiness by fusing classifiers trained with novel acoustic features,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[4] R. Gajšek, S. Dobrišek, and F. Mihelič, “University of ljubljana system for interspeech 2011 speaker state challenge,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] J. Krajewski and B. Kröger, “Using Prosodic and Spectral Characteristics for Sleepiness Detection Work and Organizational Psychology , 42097 Wuppertal , Germany Aachen and Aachen University,” *Changes*, pp. 1841–1844, 2007.

[6] J. Krajewski, A. Batliner, and M. Golz, “Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach,” *Behavior research methods*, vol. 41, no. 3, pp. 795–804, 2009.

[7] A. P. Vogel, J. Fletcher, and P. Maruff, “Acoustic analysis of the effects of sustained wakefulness on speech,” *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3747–3756, 2011.

[8] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, “Toward a Unified Theory of Voice Production and Perception,” *Loquens*, vol. 1, no. 1, pp. 1–9, jun 2014.

[9] 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.

[10] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, “The relationship between acoustic and perceived intraspeaker variability in voice quality,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] S. J. Park, C. Sigouin, J. Kreiman, P. A. Keating, J. Guo, G. Yeung, F.-Y. Kuo, and A. Alwan, “Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition,” in *Proc. Interspeech*, San Francisco, USA, sep 2016, pp. 1044–1048.

[12] S. J. Park, G. Yeung, J. Kreiman, P. A. Keating, and A. Alwan, “Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems,” in *Proc. Interspeech*, Stockholm, Sweden, aug 2017, pp. 1522–1526.

[13] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, “Towards Understanding Speaker Discrimination Abilities in Humans and Machines for Text-Independent Short Utterances of Different Speech Styles,” *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 375–386, jul 2018. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.5045323

[14] S. J. Park, A. Afshan, Z. M. Chua, and A. Alwan, “Using Voice Quality Supervectors for Affect Identification,” in *Proc. Interspeech*. Hyderabad, India: ISCA, sep 2018, pp. 157–161. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/1401.html

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

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[17] J. H. L. Hansen and T. Hasan, “Speaker Recognition by Machines and Humans: A Tutorial Review,” *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[18] H. P. Greeley, J. Berg, E. Friets, J. Wilson, G. Greenough, J. Picone, J. Whitmore, and T. Nesthus, “Fatigue estimation using voice analysis,” *Behavior Research Methods*, vol. 39, no. 3, pp. 610–619, 2007.

[19] T. Åkerstedt and M. Gillberg, “Subjective and objective sleepiness in the active individual,” *International Journal of Neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.

[20] M. P. Alvin and A. Martin, “Nist speaker recognition evaluation chronicles,” in *Proc. Odyssey 2004, The Speaker and Language Recognition Workshop*. Citeseer, 2004.

[21] M. A. Przybocki, A. F. Martin, and A. N. Le, “Nist speaker recognition evaluation chronicles-part 2,” in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–6.

[22] A. F. Martin and C. S. Greenberg, “Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[23] D. Graff, K. Walker, and A. Canavan, “Switchboard-2 phase ii,” *LDC 99S79–http://www. ldc. upenn. edu/Catalog*, 1999.

[24] 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, aug 1994. [Online]. Available: http://pubs.asha.org/doi/10.1044/jshr.3704.769

[25] Y.-L. Shue, *The voice source in speech production: Data, analysis and models*. University of California, Los Angeles, 2010.

[26] H. You, Q. Zhu, and A. Alwan, “Entropy-based variable frame rate analysis of speech signals and its application to asr,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–549.

[27] N. H. de Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[28] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the acoustics of emotion in audio: what speech, music, and sound have in common,” *Frontiers in psychology*, vol. 4, p. 292, 2013.

[29] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[30] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, jan 2000. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1051200499903615

[31] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[32] K. A. Bollen and R. W. Jackman, “Regression diagnostics: An expository treatment of outliers and influential cases,” *Sociological Methods & Research*, vol. 13, no. 4, pp. 510–542, 1985.

[33] P. J. Rousseeuw and K. V. Driessen, “A fast algorithm for the minimum covariance determinant estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[34] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, “Support vector regression machines,” in *Advances in neural information processing systems*, 1997, pp. 155–161.

[35] J. Hauke and T. Kossowski, “Comparison of values of pearson’s and spearman’s correlation coefficients on the same sets of data,” *Quaestiones geographicae*, vol. 30, no. 2, pp. 87–93, 2011.