Kodali, Manila; Kadiri, Sudarsana; Laaksonen, Laura; Alku, Paavo

Automatic classification of vocal intensity category from speech

Published in:
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'23)

DOI:
10.1109/ICASSP49357.2023.10097160

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Kodali, M., Kadiri, S., Laaksonen, L., & Alku, P. (2023). Automatic classification of vocal intensity category from speech. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'23) (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10097160

AUTOMATIC CLASSIFICATION OF VOCAL INTENSITY CATEGORY FROM SPEECH

Manila Kodali[1]*, Sudarsana Reddy Kadiri*[1]*, Laura Laaksonen*[2]*, and Paavo Alku*[1]

[1]Department of Information and Communications Engineering.
[2]Tampere Wireless Headset Audio Lab, Huawei Technologies Oy (Finland) Co., Ltd., Tampere, Finland.

ABSTRACT

Regulation of vocal intensity is a fundamental phenomenon in speech communication. Vocal intensity can be quantified using sound pressure level (SPL), which can be measured easily by recording a standard calibration signal with speech and by comparing the energy of the recorded speech signal with that of the calibration tone. Unfortunately, speech recordings are mostly conducted without the SPL calibration signal, and speech signals are saved to databases using arbitrary amplitude scales. Therefore, neither the SPL nor the intensity category (e.g. soft or loud phonation) of a saved speech signal can be determined afterwards. Even though the original level information of speech is lost when the signal is presented on arbitrary amplitude scales, the speech signal contains other acoustic cues of vocal intensity. In the current study, we study machine learning and deep learning -based methods in automatic classification of vocal intensity category when the input speech is expressed using an arbitrary amplitude scale. A new gender-balanced database consisting of speech produced in four vocal intensity categories (soft, normal, loud, and very loud) was first recorded. Support vector machine and deep neural network (DNN) models were used to develop automatic classification systems using spectrograms, mel-spectrograms, and mel-frequency cepstral coefficients as features. The DNN classifier using the mel-spectrogram showed the best classification accuracy of about 90%. The database is made publicly available at `https://bit.ly/3tLPGRx`.

Index Terms: Vocal intensity, SPL, MFCCs, SVM, DNN.

1. INTRODUCTION

In speech communication, the regulation of vocal intensity is a fundamental phenomenon. In our daily lives, we change the volume of our speech depending on the situation: to emphasize something, to be heard across a greater distance, in crowded areas, or, for example, to convey vocal emotions such as outrage [1–5]. In physics, intensity is defined as power per unit area [6]. In speech science, the word "vocal intensity" refers to the acoustic energy of speech, which is typically quantified using the sound pressure level (SPL) [7]. In this study, we use the term "vocal intensity" (widely used in speech acoustics and voice research (e.g. [7]) instead of the term "vocal effort" (used in phonetics [8]). We regard these two terms synonymous.

In addition to its linguistic content, speech carries significant paralinguistic information of, for example, vocal emotions, speaker identity and dialects [9–13]. Paralinguistic information can be divided into speaker traits and speaker states [14]. Speaker traits are long-term speaker-specific characteristics (e.g., gender, age). Speaker states refer to medium- and short-term speaker characteristics (e.g., emotions, state of health). In paralinguistic research, particularly in automatic classification of state of health from speech, knowing the $intensity\ category$ of a speech signal (e.g. soft phonation vs. loud phonation) is beneficial. This is because many disorders (e.g. vocal hyperfunction, dysphonia) affect the intensity regulation mechanism of speech and therefore automatic methods to detect changes in the intensity category could help in the speech-based classification of state of health [15–17]. In existing paralinguistic speech databases, however, the true intensity category or the true SPL that the speaker used in the production of the recorded signal remains unknown. This is because of the general recording practices that are followed in speech technology to collect speech to databases: speech recordings are mostly conducted with no calibration information and the recorded speech signals are expressed and saved in databases using arbitrary amplitude scales. Therefore, intensity category/SPL cannot be measured afterwards from the saved speech signals.

Audio devices change sound intensity solely by increasing or decreasing the gain of the output signal. However, the human speech production mechanism changes many acoustical properties (pitch, spectral tilt, phone duration, etc.) of the signal when intensity is altered [18]. Many speech science studies have been published on intensity regulation of speech investigating functioning of the human speech production mechanism [18–23]. Machine learning (ML) -based classification of vocal intensity category has been studied mainly as a binary classification problem either to detect whispering [24–26] or shouting [27–29]. However, studies on automatic multi-class classification of vocal intensity categories are sparse. In [30, 31], vocal intensity was categorized into five classes (whisper, soft, neutral, loud, and shout) and automatic classification systems were developed using the Mel-frequency cepstral coefficient (MFCC) features and the Gaussian mixture model (GMM), support vector machine (SVM) and Bayesian classifiers. However, the database used in [30] is small (12 male speakers and five sentences for each intensity class), and it is not publicly available. In addition, the database used in [31] contains only 13 male speakers (isolated vowels, numbers, words and four read sentences for each intensity class). It is also worth observing that there are no female speakers in either of these previous studies. Moreover, the experiments of both [30] and [31] were conducted by labeling each recorded speech signal using only one labeling approach. This labeling approach (which will be hereafter called as the $subjective$ labeling) corresponds to using a data collection scenario, where speakers are asked to produce speech in a few target intensity classes (e.g. soft, normal, loud) by allowing each talker to use his/her own subjective, habitual intensity levels, and by labeling each recorded speech signal using the corresponding target intensity class. In this study, we collected a larger gender-balanced database (25 male and 25 female speakers) consisting of four vocal intensity categories (soft, normal, loud, very loud) by including calibration information. Moreover, in addition to the subjective labeling approach, we also study the classification task based on the $objective$ labeling approach. This corresponds to labeling each produced speech signal objectively using the signal's

SPL.

The main contributions of this study are as follows:

1. We collected a large gender-balanced vocal intensity database consisting of speech of four different vocal intensity categories by including calibration information. The database is made publicly available at `https://bit.ly/3tLPGRx`.
2. We studied a 4-class (soft, normal, loud, and very loud) vocal intensity classification task from speech signals expressed using arbitrary amplitude scales by taking advantage of both ML-based and deep learning (DL) -based classifiers.

The paper is organized as follows. Section 2 describes the data collection process, SPL analysis and labeling. Section 3 explains the steps involved in the experimental setup. Section 4 reports the results. Finally, Section 5 concludes the study by summarizing the findings and suggesting future research areas.

2. DATA COLLECTION

Speech data was recorded in a listening room fulfilling the requirements of ITU-R BS.1116-1 [32]. A total of 25 male and 25 female speakers participated in the data collection. The age range was from 21 to 31 years for female speakers and 20 to 38 years for male speakers. The speech signals were produced in English and all the participants were proficient in English. For each speaker, both speech and electroglottography (EGG) signals were recorded, but only the former is used in the current study. For collecting speech, the DPA 4065-BL headset condenser microphone was used, and for collecting EGG, the EG2-PCX2 electroglottograph [33] was used. The other equipment used during the recordings were an Amprobe SM-CAL1 sound meter calibrator [34], a sound card, and a laptop. More details of the database can be found at `https://bit.ly/3tLPGRx`.

The data collection process was divided into two sessions (Session-1 and Session-2). Session-2 was a repetition of Session-1, and the tasks were identical in both sessions. Each session had two tasks. In the first task (referred to as Task-1), each speaker was asked to recite 25 isolated sentences in all four intensity categories. The orthographic transcriptions of the sentences were selected from the TIMIT database [35]. In the second task (referred to as Task-2), each speaker was asked to recite two different paragraphs in all four intensity categories. The first paragraph was taken from a weather forecast excerpt [36], and the second one was an excerpt from a novel [37]. Task-1 included in total 10000 sound files (25 sentences*50 speakers*4 intensity categories*2 sessions), with 2500 files per each intensity category. Task-2 included in total 800 sound files (2 paragraphs*50 speakers*4 intensity categories*2 sessions), with 20 files per each intensity category. In the remaining part of the study, only Task-1 data is considered due to the page limitation.

2.1. Computation of SPL

SPL was measured from the recorded speech signals using the calibration tone generated by the Amprobe SM-CAL1 calibrator. Utilizing the energy computation (i.e. the sum of squares) as well as the SPL value of the calibration tone (94 dB), the SPL values were determined on the dB-scale as follows:

$$SPL\{speech\} = 94 + 10\log_{10}\frac{Energy\{speech\}}{Energy\{calibration\}}. \quad (1)$$

The SPL values computed by Eq. (1) correspond to SPL measured at a distance of 5 cm from the speaker's lips using linear frequency weighting. The computed average SPL values in each intensity category were analyzed using box plots (see Fig. 1). The middle value

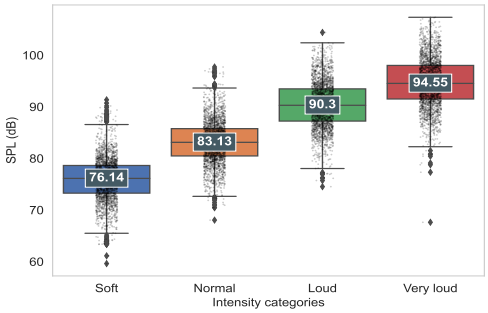

Fig. 1: Box plot of the average SPL (dB) values for the intensity categories (soft, normal, loud, and very loud).

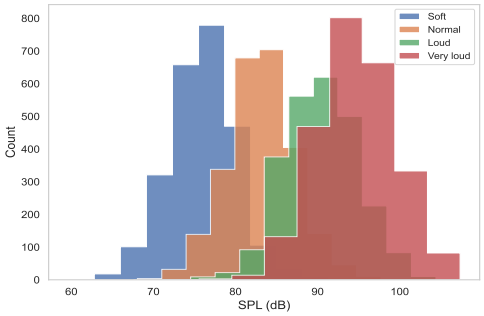

Fig. 2: Histogram of the average SPL (dB) values for the intensity categories (soft, normal, loud, and very loud).

represents the median, and the upper and lower margins represent the 25th and 75th percentiles, respectively. Whiskers illustrate data points that are within 1.5 times the interquartile range, and black dots represent all data points' distribution. As expected, it can be observed that the average SPL values indicate a rising trend when intensity category changes from soft to very loud. However, the histogram (Fig. 2) shows overlapping between the intensity categories. This is due to the data collection procedure used, that is, the speakers used their habitual vocal intensity in producing speech in each (target) intensity category and therefore the SPL values varied in each category between the speakers.

Table 1: Division of the recorded speech signals into four intensity categories based on SPL values (objective labeling).

Intensity category	SPL values	Number of files
Soft	SPL $<$79 dB	2348
Normal	79 dB $\leq$ SPL $<$86 dB	2525
Loud	86 dB $\leq$ SPL $<$93 dB	2814
Very loud	SPL $\geq$ 93 dB	2313
Total		10000

2.2. Labeling

Two labeling approaches were used in this study. The first one, called the *subjective labeling*, refers to labeling a recorded speech signal using the target intensity category adopted by the speaker in production of the corresponding signal in the recordings. The second labeling approach, called the *objective labeling*, refers to labeling a recorded speech signal based on an objective measure, the SPL of the signal. The SPL values used in the objective labeling are described in Table 1. As shown by the second column of the table, a speech sample was labeled as "soft" when its SPL $<$ 79 dB, as "normal" when its SPL was between 79–86 dB, as "loud" when its SPL

was between 86–93 dB, and as "very loud" when its SPL $>$ 93 dB. Both the subjective and objective labels were used in the automatic ML- and DL-based classification experiments of the study.

3. EXPERIMENTAL SETUP

A schematic diagram of the proposed vocal intensity classification system is shown in Fig. 3. The proposed system consists of three steps: pre-processing and normalization, feature extraction, and classification. The pre-processing and normalization steps are described in sub-section 3.1. The feature extraction step involves the extraction of three different features, namely, spectrograms, mel-spectrograms and MFCCs (described in sub-section 3.2). Finally, SVM and DNN classifiers are used for classification (described in sub-section 3.3).

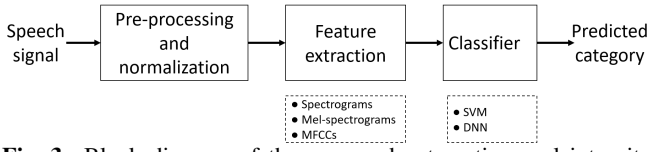

Fig. 3: Block diagram of the proposed automatic vocal intensity classification system.

3.1. Pre-processing and normalization

To remove silent regions, pre-processing was carried out using the sound exchange (SoX) tool [38]. After pre-processing, all signals were normalized by dividing the time-domain signal by its maximum amplitude value. The normalization was essential in order to model the scenario described in the study justification (see Section 1): by normalizing speech, the original intensity information that is embedded in the level/gain of the signal is deliberately removed and therefore cannot be used in the automatic classification.

3.2. Feature extraction

Three widely used frequency domain representations (spectrograms, mel-spectrograms and MFCCs) were used as features, and all of them were extracted from the normalized speech signals. Speech was windowed into frames using a Hamming window of 25 ms with 5 ms overlap. In computing the spectrogram, a FFT size of 2048 was used resulting in a 1025-dimensional feature vector. Two statistics (mean and standard deviation) were calculated over all the signal's feature vectors resulting in a 2050-dimensional spectrogram feature vector per utterance. In computing the Mel-spectrogram, the number of Mel filters was 128. Two statistics (mean and standard deviation) were calculated over all the frames, thus resulting in a 256-dimensional Mel-spectrogram feature vector per utterance. In computing the MFCCs, 39-dimensional vectors (with delta and delta-delta coefficients) were calculated. Two statistics (mean and standard deviation) were calculated over all the frames, resulting in a 78-dimensional MFCC feature vector per utterance.

3.3. Classifiers

This study uses SVM and DNN classifiers to classify speech signals into four intensity classes. SVM is one of the most widely used supervised ML algorithms in classification and regression tasks. On the other hand, DNN is the most widely used neural network model in classification and identification tasks. The main difference between the SVM and DNN is that SVM-based classification involves two separate stages (feature extraction and classification). In DNN, these two stages are merged into the same network, which implicitly conducts the two parts, yet the input can also be hand-crafted features.

In the SVM classifier model, the parameters were tuned using the GridSearchCV method [39]. The best-fitted parameters were the radial basis function (RBF) as the kernel, a cost function of 10, and a gamma of 0.001. The DNN model used in this study consists of two hidden dense layers. Cross-entropy was used as the loss function. The Adaptive Moment Estimation (ADAM) optimizer was used with a learning rate of 0.0001. The batch size is 32 and the number of training epochs was set to 100 [40]. Experiments were carried out using 5-fold cross-validation, where the data was partitioned into five equal folds between the speakers, which guarantees speaker-independency. One fold (10 speakers' data) was kept out for testing, and the other four folds (40 speakers' data) were used for training. Classification performance was measured by accuracy, and confusion matrices were used to visualize misclassifications. Evaluation metrics were stored per fold, and lastly, the mean and standard deviation of the metrics were calculated.

4. RESULTS

The results of the experiments are reported in Table 2. The performances are presented feature-wise (spectrogram, mel-spectrogram and MFCCs), classifier-wise (SVM and DNN), and label-wise (subjective and objective labeling). The results show that a better classification accuracy was obtained using the objective labels compared to using the subjective labels for all the features and for both classifiers. Among the features, the mel-spectrogram and spectrogram performed better than the MFCCs, especially for the objective labeling. Between the classifiers, DNN performed clearly better than the SVM classifier in both the subjective and objective labeling (except for the MFCC features with the DNN classifier). The best accuracy was obtained by using the mel-spectrogram features and the DNN classifier in the classification of the objective labels, for which an accuracy of 89.9% was obtained.

Table 2: Mean and standard deviation of classification accuracy (in %). The values are shown separately for the two labeling approaches, three feature sets (spectrogram, mel-spectrogram and MFCCs) and two classifiers (SVM and DNN).

Labeling	Features	SVM	DNN
Subjective	Spectrogram	65.0±0.5	64.0±0.5
	Mel-spectrogram	61.3±1.0	63.5±1.1
	MFCCs	60.5±0.4	64.4±1.7
Objective	Spectrogram	71.9±2.0	85.4±1.0
	Mel-spectrogram	71.3±2.0	89.9±0.4
	MFCCs	64.6±3.2	61.3±2.0

Figure 4 and 5 show the confusion matrices for the best performing systems with the subjective and objective labelings, respectively. From Fig. 4 it can be seen that there are less mis-classifications between the outermost categories (soft and very loud) than between normal and loud. In the case of the objective labeling (Fig. 5), the number of mis-classification in general is less compared to the subjective labeling. Overall, the results indicate that in automatic vocal intensity category classification, the use of labeling that is based on objective SPL-values yields a higher accuracy compared to the use of the subjective labeling approach.

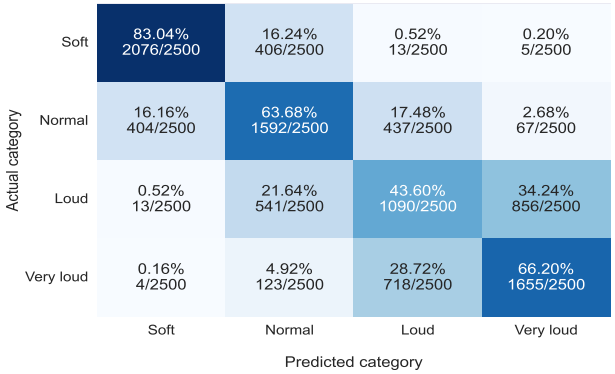

Fig. 4: Confusion matrix for the SVM classifier using the spectrogram features based on the subjective labeling approach. Row corresponds to the actual intensity category and column corresponds to the predicted intensity category.

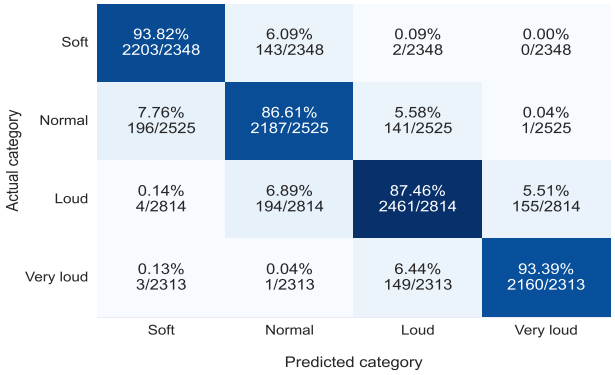

Fig. 5: Confusion matrix for the DNN classifier using the mel-spectrogram features based on the objective labeling approach. Row corresponds to the actual intensity category and column corresponds to the predicted intensity category.

5. CONCLUSIONS

In this study, we have investigated the automatic classification of intensity category of speech when the signal is expressed using an arbitrary amplitude scale without relevant SPL information. To the best of our knowledge, only two previous studies have been published previously on this topic [30, 31]. The topic is justified because most speech databases are recorded without calibration information and therefore SPL values cannot be computed from the recorded speech signals afterwards. However, by using the ML- and DL-based approaches investigated in the current study, it is possible to estimate the intensity category of speech despite the original level information of the signal is lost and the signal is expressed using an arbitrary amplitude scale. Given the importance of vocal intensity in paralinguistics and in speech communication in general, we argue that the automatic ML- and DL-based classification of vocal intensity category can be utilized in various areas, such as in speech-based biomarking of health and forensic applications.

In order to study the topic, we first collected a new speech database by recording speech signals of different intensity levels together with an SPL calibration signal. This data was then used to train and test ML-based and DL-based networks to classify intensity category of speech automatically into four classes (soft, normal, loud, very loud) when the original SPL information of the speech waveform was removed by normalization. Two classifiers (SVM and DNN), three feature sets (spectrogram, mel-spectrograms and MFCCs), and two labeling approaches (subjective and objective labeling) were compared. The results showed that by labeling the recorded speech data using the objective, SPL-based labels yielded better accuracy for the automatic systems compared to labeling the data using the target classes that were subjectively adopted in speech recordings by the speaker based on his/her habitual intensity regulation. The study showed that the combination of the mel-spectrogram feature and the DNN classifier gave the best accuracy (of about 90%) using the objective labels. The best system developed in the current study may serve as the reference system for future research.

Because laryngeal adjustments are essential in natural regulation of vocal intensity, exploiting features that are related to the voice source could be investigated in the future aiming at improved accuracy in the classification task. Potential voice source features are, for example, source spectrograms and the strength of excitation [41,42]. It was also found that intensity category information is not uniformly distributed in time throughout the speech signal; hence, advanced neural networks like attention-based models could also be used in the future to improve the classification performance [43].

6. ACKNOWLEDGEMENTS

This research has been funded by the Academy of Finland (project no. 330139) and Huawei Technologies Oy (Finland) Co.

7. REFERENCES

[1] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[2] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[3] K. Oatley, "The importance of being emotional," *New Scientist*, vol. 123, no. 1678, pp. 33–36, 1989.

[4] K. R. Scherer, "Speech and emotional states," *Speech Evaluation in Psychiatry*, pp. 189–220, 1981.

[5] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[6] H. D. Young, R. A. Freedman, T. Sandin, and A. L. Ford, *University Physics*. Addison-Wesley Reading, MA, 1996, vol. 9.

[7] I. Titze, *Principles of Voice Production*. Prentice-Hall, NJ, 1994.

[8] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[9] J. P. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech & Language*, vol. 28, no. 1, pp. 278–294, 2014.

[10] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, 2003, pp. 2417–2420.

[11] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech—a review," *Toward Robotic Socially Believable Behaving Systems-Volume I*, pp. 205–238, 2016.

[12] 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.

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

[14] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[15] J. P. Clark, S. G. Adams, A. D. Dykstra, S. Moodie, and M. Jog, "Loudness perception and speech intensity control in Parkinson's disease," *Journal of Communication Disorders*, vol. 51, pp. 1–12, 2014.

[16] M. Brockmann-Bauser, J. Bohlender, and D. Mehta, "Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders," *Journal of Voice*, vol. 32, no. 2, pp. 162–168, 2018.

[17] M. Brockmann-Bauser, J. H. Van Stan, M. C. Sampaio, J. E. Bohlender, R. E. Hillman, and D. D. Mehta, "Effects of vocal intensity and fundamental frequency on cepstral peak prominence in patients with voice disorders and vocally healthy controls," *Journal of Voice*, vol. 35, no. 3, pp. 411–417, 2021.

[18] P. Alku, M. Airas, E. Björkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 1052–1062, 2006.

[19] 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.

[20] E. M. Finnegan, E. S. Luschei, and H. T. Hoffman, "Modulations in respiratory and laryngeal activity associated with changes in vocal intensity during speech," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 4, pp. 934–950, 2000.

[21] N. Isshiki, "Regulatory mechanism of voice intensity variation," *Journal of Speech and Hearing Research*, vol. 7, no. 1, pp. 17–29, 1964.

[22] P. Ladefoged and N. P. McKinney, "Loudness, sound pressure, and subglottal pressure in speech," *The Journal of the Acoustical Society of America*, vol. 35, no. 4, pp. 454–460, 1963.

[23] S. Tanaka and M. Tanabe, "Glottal adjustment for regulating vocal intensity an experimental study," *Acta Otolaryngologica*, vol. 102, no. 3-4, pp. 315–324, 1986.

[24] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 883–894, 2011.

[25] N. Obin, "Cries and whispers-classification of vocal effort in expressive speech," in *Interspeech*, 2012.

[26] G. N. Meenakshi and P. K. Ghosh, "Robust whisper activity detection using long-term log energy variation of sub-band signal," *IEEE Signal Processing Letters*, vol. 22, pp. 1859–1863, 2015.

[27] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: human and machine," *Journal of the Acoustical Society of America*, vol. 133, pp. 2377–2389, 2013.

[28] S. Baghel, S. R. M. Prasanna, and P. Guha, "Exploration of excitation source information for shouted and normal speech classification," *Journal of the Acoustical Society of America*, vol. 147, pp. 1250–1261, 2020.

[29] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[30] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *The Eighth Annual Conference of the International Speech Communication Association*, 2007.

[31] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012.

[32] I. Rec, "Bs. 1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *Int. Telecomm. Union, Geneva Std*, 1997.

[33] "Eg2-pcx2 electroglottograph home page." http://www.glottal.com/ Electroglottographs.html, accessed: 2021-08-30.

[34] "Amprobe sound meter calibrator home page." https://www.amprobe.com/product/sm-cal-1/, accessed: 2021-10-30.

[35] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[36] "Weather forecasting excerpt," https://bit.ly/3iDF3K6, accessed: 2021-06-30.

[37] "The call of the wild by jack london," https://www.gutenberg.org/ebooks/215, 2008, [Online; Accessed: 2021-06-30].

[38] B. Barras, "Sox: Sound exchange," Tech. Rep., 2012.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[41] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3050–3061, 2013.

[42] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, and B. Yegnanarayana, "Excitation features of speech for emotion recognition using neutral speech as reference," *Circuits, Systems, and Signal Processing*, vol. 39, no. 9, pp. 4459–4481, 2020.

[43] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.