INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Classification of Vocal Intensity Category from Speech using the Wav2vec2 and Whisper Embeddings

Manila Kodali, Sudarsana Reddy Kadiri, Paavo Alku

Department of Information and Communications Engineering, Aalto University, Finland.
manila.kodali@aalto.fi, sudarsana.kadiri@aalto.fi, paavo.alku@aalto.fi

Abstract

In speech communication, talkers regulate vocal intensity resulting in speech signals of different intensity categories (e.g., soft, loud). Intensity category carries important information about the speaker's health and emotions. However, many speech databases lack calibration information, and therefore sound pressure level cannot be measured from the recorded data. Machine learning, however, can be used in intensity category classification even though calibration information is not available. This study investigates pre-trained model embeddings (Wav2vec2 and Whisper) in classification of vocal intensity category (soft, normal, loud, and very loud) from speech signals expressed using arbitrary amplitude scales. We use a new database consisting of two speaking tasks (sentence and paragraph). Support vector machine is used as a classifier. Our results show that the pre-trained model embeddings outperformed three baseline features, providing improvements of up to 7%(absolute) in accuracy.

Index Terms: Vocal intensity, sound pressure level, paralinguistics, Wav2vec2, Whisper.

1. Introduction

In speech communication, speakers frequently adjust their vocal intensity for various reasons, such as to emphasize something, to make spoken messages audible in noisy environments or when speaking over long distances, or to express emotions like anger or sadness. In contrast to audio equipment, which alter sound intensity by solely increasing or decreasing the gain of the signal, the human speech production mechanism changes several characteristics (e.g., pitch, spectral tilt, duration) of the produced acoustical signal in the regulation of vocal intensity [1]. Vocal intensity is typically quantified in sound pressure level (SPL) measured using a sound level meter [2]. In this study, we use the term "vocal intensity" (widely used in speech acoustics and voice research (e.g. [1]) instead of the term "vocal effort" (used in phonetics [3]). We regard these two terms synonymous.

Speech carries plenty of paralinguistic information, including vocal emotions, age, gender, and dialect [4, 5, 6]. Paralinguistic information can be divided into speaker traits (e.g., gender, age) and speaker states (e.g., emotions, state of health) [7]. Accurate classification of vocal intensity from speech signals is beneficial in paralinguistic research, particularly in biomarking the speaker's state of health [8, 9]. Many speech disorders, such as vocal hyperfunction and dysphonia, have a detrimental effect on the regulation of vocal intensity. Hence, vocal intensity category of speech carries valuable information about the speaker's state of health and this information could be used, for example, in studying speech-based biomarking of health. Current paralinguistic speech databases, however, lack information about the intensity category or SPL used by the speaker in the data recording. Since speech recordings are mainly collected without calibration information and the data is saved using arbitrary amplitude scales, the measurement of intensity category/SPL from the saved speech signals is not possible after the recordings. However, machine learning (ML) -based methods can in principle be used to automatically classify intensity category of speech despite the signal has been recorded without calibration information [10].

Most of the previous studies on automatic classification of vocal intensity category have addressed detection of a single vocal intensity class, particularly whispering and shouting, from speech of normal intensity (i.e., studying a binary classification problem) (e.g., [11, 12]). However, only a few studies have investigated automatic classification of $multiple$ vocal intensity categories (i.e., studying a multi-class classification problem). In [13], vocal intensity was classified into five different categories (whisper, soft, normal, loud, and shout). The authors developed an automatic classification system using the mel-frequency cepstral coefficient (MFCC) feature and the Gaussian mixture model (GMM) as a classifier. Classification of the same five intensity classes was also studied in [14] using the MFCC feature and support vector machine (SVM), Gaussian as well as Bayesian classifiers. However, the datasets used both in [13] and in [14] include small numbers of speakers (12 in [13]; 13 in [14]) and there were no female speakers in either studies. Moreover, in both of these previous studies, only MFCCs were used as the feature. Therefore, new research is needed in the study of automatic classification of intensity category of speech by including multiple intensity categories and more advanced neural net models, and by studying larger amounts of speech data produced by both female and male talkers.

Pre-trained models that have resulted from recent developments in deep learning are becoming popular in many areas of speech technology [15, 16, 17]. The use of pre-trained models constitute an attractive tool particularly in areas such as paralinguistics where speech dataset are typically small. In these areas, the use of pre-trained models enables utilising deep neural nets that are first trained in an area (such as ASR) where large datasets of speech exist and later used in an area (such as paralinguistics) where training data is less. Several approaches have been employed to utilise pre-trained models, including the use of them in feature extraction, fine-tuning, and in autoencoders [18, 19, 20]. Examples of application areas where pre-trained models have been used recently are emotion recognition [21] as well as detection of stuttering [17] and pathological speech [22] indicating that these models have potential uses in paralinguistics.

In this study, we investigate the use of two state-of-the-art

10.21437/Interspeech.2023-2038

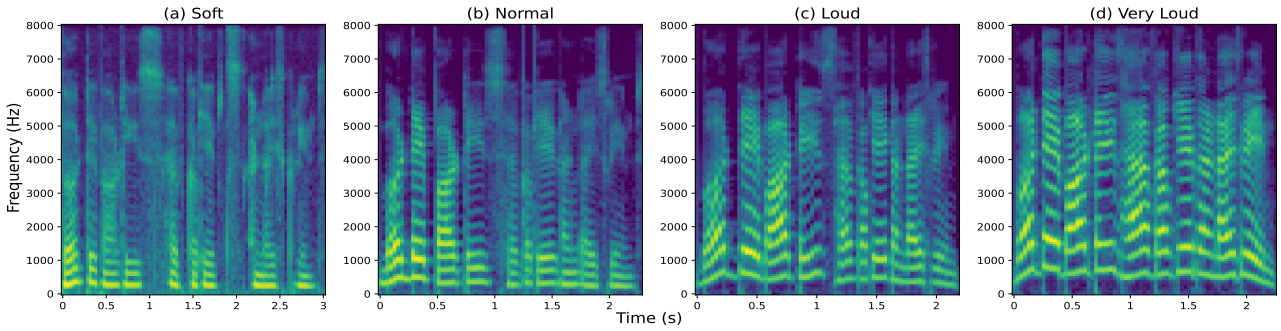

Figure 1: *Illustration of the mel-spectrogram for speech produced in the SENT task by a male speaker in (a) soft, (b) normal, (c) loud, and (d) very loud intensity category.*

pre-trained models namely, Wav2vec2 [15] and Whisper [16], as feature embeddings in automatic intensity category classification. To the best of our knowledge, pre-trained model embeddings have not been studied before for this task. We use a recently published, balanced speech corpus [10], which includes a large number of speakers (50), four intensity classes (soft, normal, loud, and very loud), and two speaking tasks (sentence and paragraph reading).

The main objectives of this study are as follows:

- To investigate the effectiveness of layer-wise Wav2vec2 and Whisper model embeddings in automatic classification of vocal intensity category (soft, normal, loud, and very loud) from speech signals that are expressed using an arbitrary amplitude scale with no SPL calibration information.
- To investigate the effect of the speaking task (sentence $vs.$ paragraph reading) in automatic classification of vocal intensity category.

The paper is organized as follows. Section 2 describes the dataset used in this study. 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 future work.

2. Dataset

In this study, we use a new publicly available dataset, which includes speech produced in English using four intensity categories (soft, normal, loud, and very loud) [10]. The dataset comprises recordings of 50 speakers (25 male and 25 female). For the male speakers, the age range is between 20 and 38 years, for the female speakers, the age range is between 21 and 31 years. The data was collected using two speaking tasks, the sentence reading task (denoted as SENT) and the paragraph reading task (denoted as PARA). In SENT, each speaker recited 25 isolated sentences in all four intensity categories. The orthographic transcriptions of the sentences were taken from the TIMIT database [23]. In PARA, the speakers recited two different paragraphs using the four intensity categories. The first paragraph was taken from a weather forecast excerpt [24] and the second paragraph was taken from a novel [25]. All the tasks were repeated two times by each speaker. For more details about the dataset, the reader is referred to `https://bit.ly/3tLPGRx`.

The SENT speaking task includes 10,000 sound files (25 sentences * 50 speakers * 4 intensity categories * 2 repetitions), with 2,500 files per each intensity category. The PARA speaking task comprises 800 sound files (2 paragraphs * 50 speakers * 4 intensity categories * 2 repetitions), with 200 files per each intensity category. Every sample of SENT and PARA was labeled using the target intensity category adopted by the speaker in the production of the corresponding signal in the recordings.

3. Experimental Setup

We use a vocal intensity classification pipeline system, which consists of three stages: pre-processing and normalization, feature extraction, and classification. The individual stages of the system are described in sub-sections 3.1, 3.2, and 3.3. A schematic diagram of the pipeline system studied is shown in Figure 2.

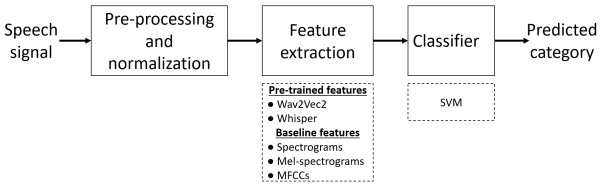

Figure 2: *Block diagram of the proposed automatic vocal intensity classification system.*

3.1. Pre-processing and normalization

In this stage, the entire speech signal (a sentence in the SENT task and a paragraph in the PARA task) is pre-processed to remove silence regions. The removal of silence is performed using the sound exchange (SoX) method [26]. After the silence removal, every signal is normalized by dividing the signal by its maximum amplitude value. This normalization is done in order to study the scenario described at the end of the 2^{nd} paragraph of Section 1. The original intensity information present in the level/gain of the signal is intentionally removed by this normalization procedure. Therefore, all the signals resulting from this stage are represented on arbitrary amplitude scales, and they can be used to test and train ML models for automatic classification of vocal intensity category in the studied scenario.

3.2. Features

In this stage, the Wav2vec2, Whisper and baseline features are extracted from the normalized speech signals that were computed in the previous stage of the pipeline. We use two state-of-the-art pre-trained models as feature embeddings: 'Wav2Vec2-Large-960h-Lv60 + Self-Training model' (Wav2vec2) and 'Whisper-large-v2 model' (Whisper) from Huggingface [27]. Wav2vec2 is trained as a self-supervised model that learns to predict masked portions of speech from unlabeled speech

data [15]. The data is taken from the Libri-Light and Librispeech datasets, which contain 960k hours of English speech. Wav2vec2 has a convolutional and transformer encoder architecture. It first transforms raw speech into feature vectors, then applies several convolutional neural network (CNN) layers and a transformer encoder. The transformer encoder has a stack of multiple self-attention layers that process the feature vectors and produce encoder hidden states.

Whisper is trained as a supervised model that learns to map speech to text from labeled speech data [16]. The data is taken from the web and contains about 680k hours of speech in 60 languages. Whisper has an encoder-decoder transformer architecture that takes 80-channel mel-spectrograms representation as input. The encoder consists of two CNN layers, a sinusoidal positional encoding, and a stack of transformer layers (with self-attention and feed-forward layers). It outputs the encoder's hidden states.

For speech classification tasks or any other downstream tasks, the outputs of the encoder layers are taken as features and fed to the classification head [21]. In this study, we used the encoder-hidden states of each encoder layer as feature for the classification of vocal intensity category. The encoder's hidden states are 3-D tensors (batch_size, encoder_sequence_length, hidden_size) that represent the output of each encoder layer. The Wav2vec2 model has a hidden unit size of 1024 and 24 transformer encoder layers while the Whisper model has a hidden unit size of 1280 and 32 transformer encoder layers. We converted these 3-D tensors to 1-D tensors by averaging over the sequence length. That is, the Wav2vec2 embeddings are 1024-D feature vectors per layer and per utterance, whereas the Whisper embeddings are 1280-D feature vectors per layer and per utterance.

3.2.1. Baseline features

To evaluate and compare the performance of the Wav2vec2 and Whisper embeddings, we also included three commonly used spectral features (spectrogram, mel-spectrogram, and MFCCs) as baseline features. Speech signals were windowed into frames using the Hamming window of 25 ms with a 5 ms overlap. Spectrograms were computed using the 1024-point fast Fourier transform (FFT) resulting in a 513-D vector. Mel-spectrogram was computed using the 1024-point FFT and the number of mel filters was 128, resulting in a 128-D vector. MFCCs were computed by calculating a 39-D vector that included the delta and delta-delta coefficients. Two statistics (mean and standard deviation) were computed for the baseline features over all the frames of an utterance to produce a 1026-D spectrogram feature vector, a 256-D mel-spectrogram feature vector, and a 78-D MFCC feature vector per each utterance. Figure 1 shows the mel-spectrogram in all four intensity categories for speech produced in the SENT task by a male speaker. It can be observed that as intensity increases from soft to very loud, harmonics become more prominent and energy in higher frequency bands also increases.

3.3. Classifier

The goal of this last stage is to first train a classifier using supervised learning based on the intensity class labels as well as the studied features, and then to classify speech signals into the four intensity categories. As a classifier, we used SVM, which is a popular supervised ML algorithm for classification and regression tasks. The dataset was divided into training, validation, and testing sets using the nested cross-validation, with the number of inner and outer loops set to 5 [28]. The GroupKFold method was implemented to split the inner and outer loops, which prevents the same speaker's data from being used simultaneously in the training, validation, and testing sets. To fine-tune the hyperparameters of the SVM, GridSearchCV was used by considering a subset of three kernel types ('rbf', 'linear', 'poly'), and the C and gamma values of 0.1, 1, and 10. Due to the large number of the best-fitted hyper-parameters per each inner loop and each setup, the resulting optimal parameter values are not reported in this paper.

3.4. Evaluation metrics

The performance of the classifier was evaluated using accuracy as the evaluation metric and using confusion matrices to visualize misclassifications. Evaluation metrics were computed for each outer loop and the mean and standard deviation were calculated across all the loops.

4. Results

The results of the vocal intensity category classification experiments are shown for the SENT and PARA speaking tasks in Table 1 and Table 2, respectively. These tables show both the overall classification accuracy and the class-wise accuracies (separately for all four intensity classes) for the baseline features and for the two best Wav2vec2 and Whisper features. According to Table 1, the mel-spectrogram baseline feature performed better than the other baseline features in the SENT task, with an overall accuracy of 64$\pm$2%. Similarly, Table 2 indicates that the spectrogram achieved the best classification performance among the baseline features for the PARA task, with an accuracy of 64$\pm$4%. Importantly, the Wav2vec2 and Whisper features showed better performance compared to the baseline features in both speaking tasks. This suggests that the Wav2vec2 and Whisper embeddings capture a wide range of speech characteristics, leading to better classification performance. For the SENT task, the Wav2vec2-3 and Whisper-18 features achieved an absolute improvement of 5% and 4%, respectively, compared to the best baseline feature. For the PARA task, the Wav2vec2-3 and Whisper-18 features provided an absolute improvement of 7% and 5%, respectively, compared to the best baseline feature.

Figure 3 shows the layer-wise performance of the Wav2vec2 and Whisper feature (represented by the dashed line with markers) and the best baseline feature (represented by the dashed line without markers) for both SENT (in green) and PARA (in blue). In most cases, the early and middle layers of Wav2vec2 (see Figure 3 (a)) outperform the best baseline feature for both tasks. A comparison of the two speaking tasks shows a consistently better performance for PARA compared to SENT. The trend shown by the Whisper feature (see Figure 3 (b)) suggests that this feature outperforms the baseline features. However, the performance of the Whisper layers varies considerably across the layers, which does not happen for Wav2vec2. In most cases, the PARA task shows again better performance than the SENT task. This trend is consistent for both the Wav2vec2 and Whisper features.

Figure 4 displays confusion matrices for the best-performing baseline feature, and for the Wav2vec2 and Whisper features in SENT (Figure 4 (a)) and PARA (Figure 4 (b)). In both tasks, it can be observed that Wav2vec2 showed fewer misclassifications between classes than the other features. In addition, all confusion matrices reveal that most misclassifications occur between loud and very loud speech, and the outermost

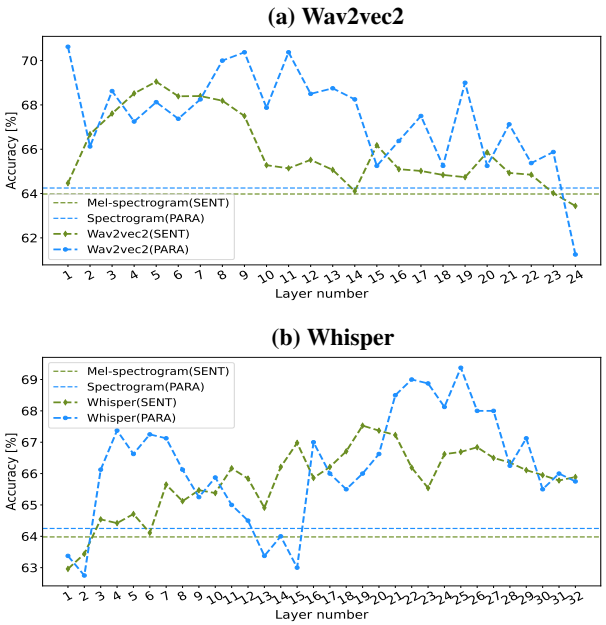

Figure 3: *Layer-wise Wav2vec2 (a) and Whisper (b) features for the SENT and PARA tasks with the best baseline feature. The blue and green lines represent SENT and PARA, respectively. The dashed lines without markers indicate the baseline feature, while the dashed lines with markers represent the Wav2vec2 and Whisper feature.*

categories have a lower rate of misclassifications.

5. Conclusions

In this study, we investigated the use of two pre-trained models (Wav2vec2 and Whisper) as feature embeddings in automatic classification of vocal intensity category (soft, normal, loud, very loud) from speech signals expressed on arbitrary amplitude scales. The experiments were carried using a new corpus consisting of speech of 50 talkers produced in four intensity categories using two speaking tasks (SENT and PARA). The experiments with the SVM classifier revealed that both the Wav2vec2 and Whisper features outperformed the baseline features (spectrogram, mel-spectrogram, and MFCC) in both speaking tasks. These findings suggest that pre-trained model embeddings are valuable features in classification of intensity class, and they can potentially be used in paralinguistic research (e.g., in biomarking the speaker's state of health) in scenarios where calibration information is not available and speech is expressed using arbitrary amplitude scales. Between the speaking tasks, PARA showed better performance than SENT in most cases, which may be due to different number of spoken words in the two tasks. However, further investigations are needed to study the effect of spoken words and the non-uniform distribution of intensity category information over time. Moreover, exploring using fusion techniques that fully leverage the effects of both pre-trained and baseline features, and fine-tuning of the models could be studied in the future. For example, it might be beneficial to fine-tune the last layers of the Wav2vec2 model, as they are more closely related to lexical contents, which is important in ASR, but which may not be useful in the classification of vocal intensity category. Advanced neural networks like time-based attention-based models can also be used to improve classification performance.

Table 1: *Classification accuracy results and class-wise accuracies for vocal intensity classification using three baseline features and two top-performing Wav2vec2 and Whisper features for the SENT task. ACC denotes accuracy and C denotes class.*

Feature	ACC [%]	C_{soft}	C_{normal}	C_{loud}	$C_{veryloud}$
Baseline features					
Spectrogram	62±2	82	60	42	64
Mel-spectrogram	64±2	82	62	47	65
MFCCs	61±2	81	56	45	64
Wav2vec2 feature					
Wav2vec2-3	**69±2**	80	74	58	62
Wav2vec2-4	**69±2**	80	74	58	64
Whisper feature					
Whisper-18	**68±2**	81	71	51	68
Whisper-19	**67±2**	83	68	51	68

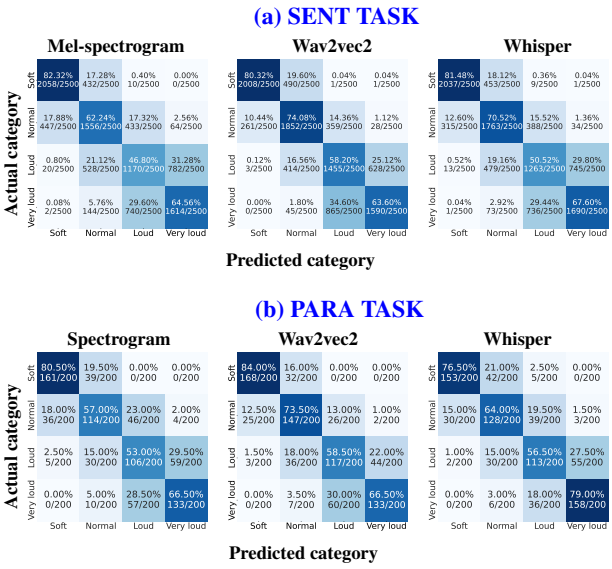

Figure 4: *Confusion matrices of the best baseline, Wav2vec2, and Whisper features for both SENT and PARA.*

Table 2: *Classification accuracy results and class-wise accuracies for vocal intensity classification using three baseline features and two top-performing Wav2vec2 and Whisper features for the PARA task. ACC denotes accuracy and C denotes class.*

Feature	ACC [%]	C_{soft}	C_{normal}	C_{loud}	$C_{veryloud}$
Baseline feature					
Spectrogram	64±4	81	57	53	67
Mel-spectrogram	63±5	77	58	52	66
MFCCs	63±3	78	61	44	71
Wav2vec2 feature					
Wav2vec2-0	**71±8**	84	74	59	67
Wav2vec2-10	**70±5**	84	74	59	67
Whisper feature					
Whisper-4	**69±5**	80	67	50	71
Whisper-21	**69±5**	77	65	57	79

6. Acknowledgements

This study was funded by the Academy of Finland (project no. 330139). The computational resources were provided by Aalto ScienceIT.

7. References

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

[2] J. G. Švec and S. Granqvist, "Tutorial and guidelines on measurement of sound pressure level in voice and speech," *Journal of Speech, Language and Hearing Research*, vol. 61, pp. 441–461, 2018.

[3] 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.

[4] 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.

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

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

[7] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

[8] 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.

[9] 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.

[10] M. Kodali, S. R. Kadiri, L. Laaksonen, and P. Alku, "Automatic classification of vocal intensity category from speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[17] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7026–7029.

[18] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3512–3516.

[19] J. Szep and S. Hariri, "Paralinguistic classification of mask wearing by image slassifiers and fusion," in *Proc. Interspeech 2020*, 2020, pp. 2087–2091.

[20] A. Haque, M. Guo, P. Verma, and L. Fei-Fei, "Audio-linguistic embeddings for spoken sentences," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7355–7359.

[21] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[22] S. Tirronen, S. R. Kadiri, and P. Alku, "Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features," *IEEE Open Journal of Signal Processing*, vol. 4, pp. 80–88, 2023.

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

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

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

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

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[28] 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.