Interspeech 2024
1-5 September 2024, Kos, Greece

Fine-tuning of Pre-trained Models for Classification of Vocal Intensity Category from Speech Signals

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

[1]Department of Information and Communications Engineering, Aalto University, Finland.
[2]Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, USA.
manila.kodali@aalto.fi; skadiri@usc.edu; paavo.alku@aalto.fi

Abstract

Speakers regulate vocal intensity on many occasions for example to be heard over a long distance or to express vocal emotions. Humans can regulate vocal intensity over a wide sound pressure level (SPL) range and therefore speech can be categorized into different vocal intensity categories. Recent machine learning experiments have studied classification of vocal intensity category from speech signals which have been recorded without SPL information and which are represented on arbitrary amplitude scales. By fine-tuning four pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, HuBERT, audio speech transformers), this paper studies classification of speech into four intensity categories (soft, normal, loud, very loud), when speech is presented on such arbitrary amplitude scale. The fine-tuned model embeddings showed absolute improvements of 5% and 10-12% in accuracy compared to baselines for the target intensity category label and the SPL-based intensity category label, respectively.

Index Terms: Vocal intensity, wav2vec2, HuBERT, sound pressure level, audio speech transformers

1. Introduction

Regulation of vocal intensity affects acoustic and prosodic characteristics of speech, including the speech signal's amplitude, fundamental frequency, spectral tilt, and phone durations [1, 2, 3]. These changes are due to various mechanisms of the human speech production mechanism below, within, and above the larynx [4]. Understanding of the vocal intensity regulation mechanisms is important in the study of speaker's state in topics such as emotion recognition and particularly in automatic biomarking of the speaker's state of health. Speech disorders such as vocal hyperfunction and dysphonia affect intensity regulation, thus making vocal intensity a valuable indicator of speaker's state of health [5]. This information can be harnessed to develop automatic speech-based biomarking technologies of human health.

Vocal intensity is quantified in dB using sound pressure level (SPL). SPL can be measured using a sound level meter or by recording a calibration tone prior to the speech recording and by calculating SPL as the energy ratio between the recorded speech signal and the calibration tone [6]. Unfortunately, most current speech databases do not support either of these procedures, preventing the measurement of SPL from the recorded speech data. However, recent studies have shown that machine learning (ML) models can be used in the classification of intensity category and in the prediction of SPL from speech signals recorded in non-calibrated conditions [7, 8, 9].

Previous studies on the automatic classification of intensity category have primarily focused on binary classification, such as distinguishing between whispered and normal speech [10, 11] or between shouting and normal speech [12, 13, 14]. However, only a few studies have investigated multi-class classification of vocal intensity category. In [7] and [8], the authors explored the automatic classification of five intensity categories (whisper, soft, normal, loud, and shout) using mel-frequency cepstral coefficients (MFCCs) as features and using various classifiers including the Gaussian mixture model (GMM), support vector machine (SVM), and Bayesian classifiers. However, these studies used non-public databases with a limited number of speakers and did not consider non-calibrated recording scenarios in which speech signals are presented on arbitrary amplitude scales.

Recently, a large open database, named Aalto Vocal Intensity Database (AVID), was released containing speech data recorded from 50 speakers in four intensity categories (soft, normal, loud, and very loud) [9]. The database includes a rich set of SPL labels and the target intensity category label corresponding to the intensity class adopted by the speaker during the recording. In [9], the authors investigated multi-class classification of the AVID's four intensity categories using MFCCs, mel-spectrogram, and spectrogram as features, and SVM and convolutional neural network (CNN) as classifiers. They specifically addressed a non-calibrated recording setup, in which speech signals were expressed using arbitrary amplitude scales. In [15], the authors explored the use of state-of-the-art pre-trained model embeddings (wav2vec2-LARGE and whisper) in the automatic classification of the AVID's four intensity categories. However, the effect of fine-tuning the pre-trained models in the classification of vocal intensity category has not yet been explored. Therefore, the aim of this study is to fine-tune state-of-the-art pre-trained models on the AVID data using two different labeling approaches (the target intensity category label and the SPL-based intensity category label). The classification task is studied by simulating a non-calibrated recording scenario where the original level information of speech is absent and speech signals are expressed on an arbitrary amplitude scale.

The main contributions of this study are as follows:

- To investigate four different fine-tuned models (wav2vec2-BASE, wav2vec2-LARGE, HuBERT, AST (audio speech transformers)) in the multi-class classification of vocal intensity category (soft, normal, loud, and very loud) using two different labeling approaches (the target intensity category label and the SPL-based intensity category label) from speech signals that are expressed on an arbitrary amplitude scale without SPL calibration information.
- To investigate the layer-wise performance of both the fine-tuned and non-fine-tuned embeddings of all four models for the classification of vocal intensity categories.

10.21437/Interspeech.2024-2237

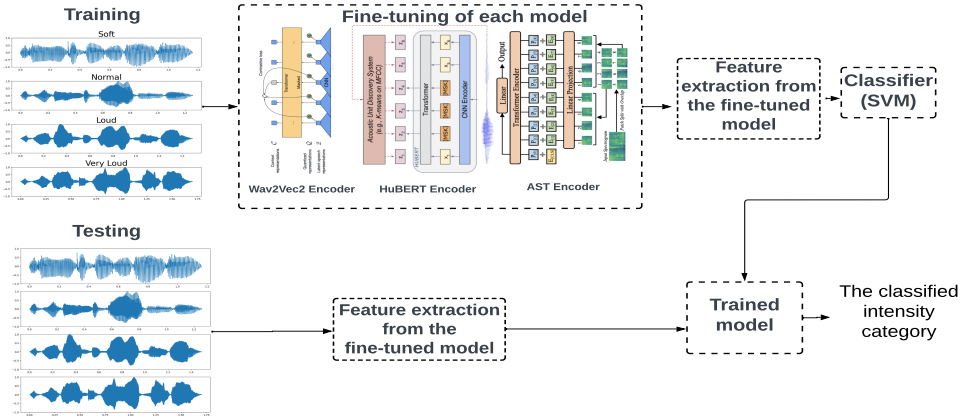

Figure 1: *A schematic diagram of the proposed automatic vocal intensity classification system.*

2. Database

In this study, we utilize the new open AVID database [9]. AVID contains gender-balanced English speech and electroglottography (EGG) recordings from 50 speakers (25 male and 25 female), who recited speech in four intensity categories: soft, normal, loud, and very loud. The male speakers age ranges from 20 to 38 years, while the female speakers age ranges from 21 to 31 years. The database consists of two speaking tasks: sentence reading and paragraph reading. In the sentence reading task, each speaker recited 25 isolated sentences in all four intensity categories. In the paragraph reading task, the speakers recited two different paragraphs in all four intensity categories. In this study, we used the sentence reading task, consisting of 10,000 speech utterances (25 sentences * 50 speakers * 4 intensity categories * 2 repetitions). Further details of the database can be found in [9, 16].

2.1. Labeling

The AVID data was labeled sentence-wise using two different labeling approaches. The first approach, termed as the 'target intensity category label', refers to labeling a recorded speech signal using the target intensity category adopted by the speaker in production of the signal. The second labeling approach, referred to as the 'SPL-based intensity category label', corresponds to labeling a recorded speech signal based on an objective measure, the SPL of the signal. In the SPL-based labeling approach, a signal was labeled as "soft" if its SPL was < 79 dB, "normal" if the SPL ranged between $79-86$ dB, "loud" if the SPL ranged between $86-93$ dB, and "very loud" if the SPL was > 93 dB.

3. Experimental setup

A schematic diagram illustrating the proposed intensity category classification pipeline is presented in Figure 1. The pipeline comprises three main steps: fine-tuning, feature (embeddings) extraction from the fine-tuned model, and classification. The fine-tuning step involves fine-tuning four state-of-the-art pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, HuBERT-LARGE, and AST) as detailed in Subsection 3.1. Next, features are extracted from the fine-tuned models (as described in Subsection 3.2). Finally, the SVM classifier is utilized for classification, with its parameters and evaluation metrics outlined in Subsection 3.4.

Before fine-tuning, each speech signal is normalized sentence wise by dividing the signal waveform by its maximum amplitude. This is carried out to investigate the non-calibrated recording scenario, where speech signals are stored without calibration information and are therefore presented on an arbitrary amplitude scale [9]. The normalization procedure is used to deliberately remove the key intensity feature of speech, the amplitude level of the time-domain signal waveform. Therefore, all speech sentence inputs at the system training and testing stages share the same amplitude range as shown in Figure 1.

3.1. Fine-tuning pre-trained models

To fine-tune the models, the normalized AVID data was divided into training (60%), validation (20%), and testing (20%) sets using the GroupK 5-fold cross-validation. During the fine-tuning stage, the models were provided with both the data and the labels of the intensity categories according to the supervised learning paradigm. A total of four pre-trained models were separately fine-tuned and used in this study: wav2vec2 (BASE and LARGE), HuBERT (LARGE), and AST model. The wav2vec2 and HuBERT models have been pre-trained in a self-supervised manner with a large amount of speech data, and fine-tuned using a smaller dataset for ASR. Thus, the last layers contain information about phonemes, while the early layers represent information related to phones. These pre-trained models can be used for various downstream tasks, such as voice pathology detection and emotion detection [17, 18, 19]. The AST model has been specifically designed for audio classification and trained on AudioSet [20], a weakly-labeled audio event classification dataset. The AST model has achieved state-of-the-art results on various audio classification tasks [21]. In addition, the AST model is the most downloaded model for audio classification on Huggingface. Therefore, we chose these four models to be used for fine-tuning in the current study. A brief overview of the four models architectures is provided below.

Wav2vec2: The wav2vec2 architecture includes a CNN encoder, a context network, and a quantization module. For wav2vec2-BASE, the context network comprises 12 transformer encoder layers, while for wav2vec2-LARGE, it consists of 24 transformer encoder layers [22].

HuBERT: We used the HuBERT-LARGE model, whose architecture looks similar to wav2vec2, but the training process differs. HuBERT builds targets via a seperate clustering process, while wav2vec2 learns its targets simultaneously while

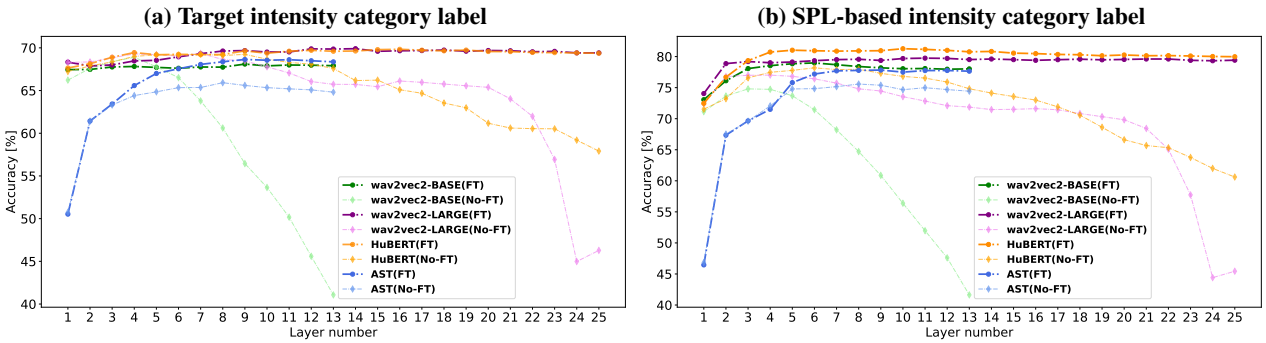

Figure 2: *Layer-wise performance for the fine-tuned models and for the non-fine-tuned models for the (a) target intensity category label and (b) SPL-based intensity category label. The dashed lines with the circle markers in green, purple, orange, and blue refer to the fine-tuned wav2vec2-BASE, wav2vec2-LARGE, HuBERT, and AST models, respectively. The lighter versions of the corresponding colors with the diamond markers refer to the non-fine-tuned models.*

training the model. HuBERT also consists of a CNN encoder, a context network (with 24 transformer encoder layers), followed by a projection and a code embedding layer [23].

AST: The AST model is the first convolutional-free and purely self-attention-based model. Initially, AST sequentially splits the spectrogram obtained from speech and feeds them to a linear projection layer, positional embedding layer, followed by 13 transformer encoder layers [24].

For fine-tuning, two fully-connected (FC) layers are added to all these models, where the last FC layer performs the multi-class classification task. Both FC layers are randomly intialized and the model weights are intialized from the original models. The categorical cross-entropy loss function is used to optimise the model. For all the models, a batch size of 8, an epoch size of 8, and learning rate of 3e-5 is set for the ADAM optimizer. All these models are fine-tuned separately for each fold of the training (a total of 5 folds) for both labeling approaches.

3.2. Feature extraction from the fine-tuned models

In this stage, features/embeddings are extracted for each layer from the fine-tuned models for the unseen test data. The encoder's hidden states, which are 3-D tensors representing the output of each encoder layer, are taken as features. Wav2vec2-LARGE and HuBERT have 1024 hidden units with 24 encoder layers, while wav2vec2-BASE and AST models have 768 hidden units and 12 encoder layers. The 3-D tensors are averaged over the sequence length, resulting in 1024-D feature vectors for wav2vec2-LARGE and HuBERT per layer and per sentence, and 768-D feature vectors for wav2vec2-BASE and AST per layer and per sentence. Additionally, the temporal average of inputs to the first transformer layer is also included as a feature vector.

3.3. Baseline features for comparison

Three spectral features (spectrogram, mel-spectrogram, and MFCCs) and eGeMAPS [25], which includes a 88-D feature set, were used as baselines. The spectral features were computed using a 25 ms Hamming window with a 5 ms overlap. Spectrograms and mel-spectrograms (with 128 mel filters) were computed with 1024 FFT points, resulting in 513-D and 128-D feature vectors, respectively. Both static and dynamic (delta and double-delta) MFCCs were computed, resulting in a 39-D feature vector. For all the spectral features, mean and standard deviation were computed over all the frames for each sentence.

3.4. Classifier

This study uses SVM as a classifier. The experiments were conducted using the GroupK 5-fold cross-validation. This means that the speakers were divided into five equal groups, and no speaker is included in both the testing and training at the same time (i.e., unseen speakers' data for testing/fine-tuning the models). We used the radial basis function as kernel and set the regularization parameter to 1. Gamma was set as $\gamma = 1/(D \cdot \widehat{Var}(X))$, where D represents the dimensionality of the features and $\widehat{Var}(X)$ is the estimated variance of the features in the training data (X).

3.4.1. Evaluation metrics

Accuracy and confusion metrics were employed to assess the models' performance and visualizing mis-classifications. The evaluation metrics were calculated for each fold, and subsequently, the mean and standard deviation were determined across all five folds.

4. Results

Table 1: *Classification accuracy (in %) for both the baseline features and for the best fine-tuned features of all the four models, evaluated for the target intensity category label and the SPL-based intensity category label.*

Features	Target intensity category label	SPL-based intensity category label
Baseline features		
Spectrogram	66.08±2.77	81.00±2.36
Mel-spectrogram	65.41±2.11	68.65±4.00
MFCCs	63.19±2.63	66.62±4.8
eGeMAPS	65.07±2.26	69.56±4.08
Fine-tuned features		
Wav2vec2-BASE	68.10±2.10	78.98±4.63
Wav2vec2-LARGE	69.90±2.79	79.71±4.76
HuBERT	69.7±3.40	81.27±3.82
AST	68.10±2.10	77.98±3.24

Table 1 shows the classification accuracy for the baseline features and for the best fine-tuned features from four models, evaluated for both the target intensity category label and SPL-based intensity category label. From the table, it can be seen that the fine-tuned model features performed better than the baseline features for both labeling categories (except for the spectrogram feature for the SPL-based intensity category label,

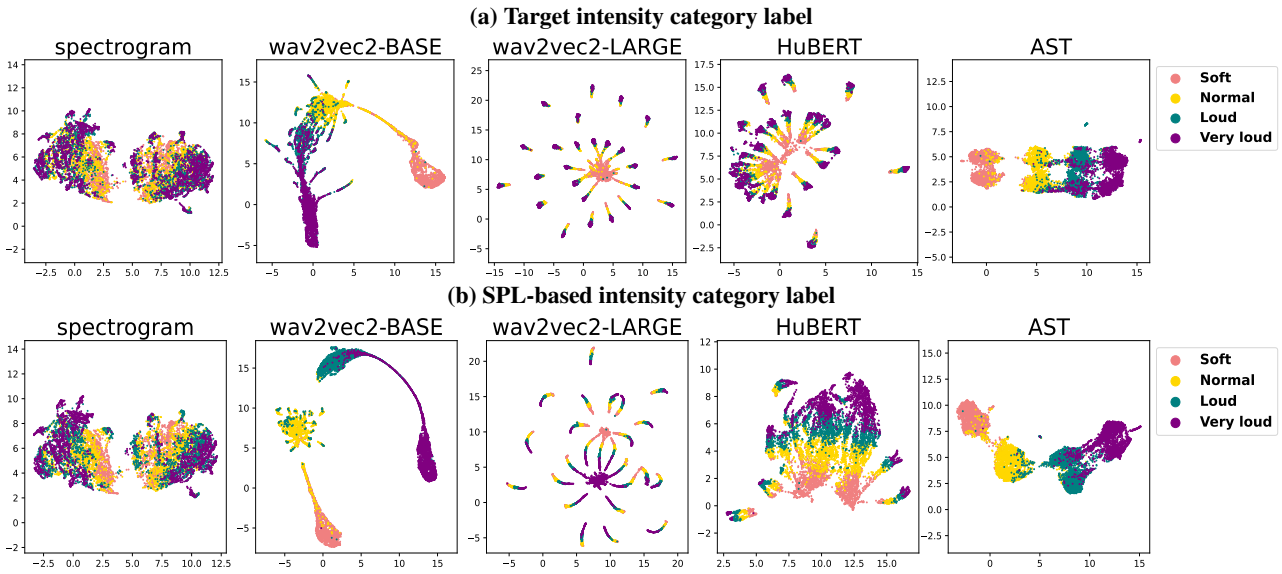

Figure 3: *Visualization using the UMAP algorithm for the best performing baseline feature (spectrogram) and for all the best-performing fine-tuned model embeddings based on the (a) target intensity category label, and (b) SPL-based intensity category label.*

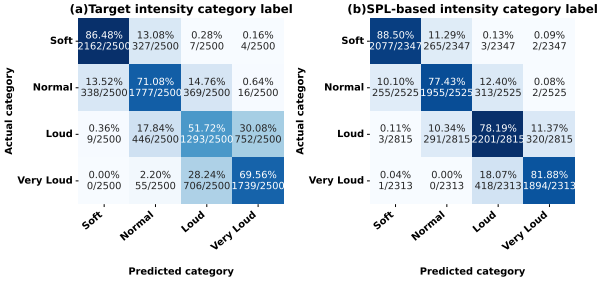

Figure 4: *Confusion matrices for the best performing HuBERT model features for the (a) target intensity category label and (b) SPL-based intensity category label.*

where it performed similar or slightly better to the fine-tuned model features). Between the labeling categories, the SPL-based label showed better performance than the target intensity category label. Among the baseline features, the spectrograms performed better than the other baseline features. Among the fine-tuned features, the wav2vec2-LARGE and HuBERT model features performed better than the wav2vec2-BASE and AST model features. This suggests that models with more layers with large amount of pre-training data are capable of capturing a wider range of speech characteristics, resulting in improved classification of vocal intensity category.

Figure 2 depicts the layer-wise performance of the non-fine-tuned (No-FT) and fine-tuned (FT) models for the wav2vec2-BASE, wav2vec2-LARGE, HuBERT, and AST models. It can be seen that the performance of the non-fine-tuned models (i.e., wav2vec2-BASE-No-FT, wav2vec2-LARGE-No-FT and HuBERT-No-FT) decreased as the layer numbers increased, but for AST (i.e., AST-No-FT), the performance remained stable or increased with an increase in layer number. However, upon fine-tuning, the performance of all models improved across all layers compared to the non-fine-tuned models. These improvements indicate an enhanced capability to learn speech representations that are relevant to the studied task of intensity category classification.

Figure 4 displays confusion matrices for the best performing HuBERT model features for both the target intensity category label and the SPL-based intensity category label. For the target intensity category label, it can be observed that there are more misclassifications between loud and very loud speech, while soft speech is more accurately classified. For the SPL-based intensity category label, only a few misclassifications between the categories can be observed.

Furthermore, the best baseline feature (spectrogram) and the fine-tuned model features are visualised by using the UMAP (Uniform Manifold Approximation and Projection) algorithm in Figure 3. From the figure, it can be observed that UMAP shows better clustering of the intensity categories for the SPL-based intensity category label. Interestingly, the clusters from the AST model appear to be more distinct from each other compared to the clusters of the other models. This could be due to the fact that the AST model was trained especially for audio classification tasks.

5. Conclusion

This study investigated the impact of fine-tuning four pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, HuBERT and AST) in the multi-class classification of vocal intensity category (soft, normal, loud, very loud) using two labeling approaches (the target intensity category label and the SPL-based intensity category label). The original amplitude level information of speech was deliberately removed using a normalization approach in order to simulate the non-calibrated recording scenario, which is widely used in collecting speech data in speech technology. The speech signals were then used for fine-tuning each model separately. From the fine-tuned models, embeddings were extracted and used in the multi-class classification task. Compared to the baseline features, the results indicated absolute improvements of roughly 4-5% for the target intensity category labels and 10-12% improvements for the SPL-based intensity category lables (except for the spectrogram feature). In addition, the experiments showed better performance for the SPL-based intensity category labels. Among the models, wav2vec2-LARGE and HuBERT gave slight improvements of 1-2% compared to the other models.

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] S. Hodge, R. Colton, and R. Kelley, "Vocal intensity characteristics in normal and elderly speakers," *Journal of Voice*, vol. 7, pp. 503–511, 2001.

[2] R. Schulman, "Articulatory dynamics of loud and normal speech," *Journal of the Acoustical Society of America*, vol. 85, pp. 295–312, 1989.

[3] J.-S. Lienard and M.-G. D. Benedetto, "Effect of vocal effort on spectral properties of vowels," *Journal of the Acoustical Society of America*, vol. 106, pp. 411–422, 1999.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] P. Alku, M. Kodali, L. Laaksonen, and S. R. Kadiri, "AVID: A speech database for machine learning studies on vocal intensity," *Speech Communication*, vol. 157, p. 103039, 2024.

[10] M. Sarria-Paja and T. H. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Letters*, vol. 20, pp. 783–786, 2013.

[11] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," in *Eighth Annual Conference of the International Speech Communication Association*, 2007, pp. 2396–2399.

[12] S. Baghel, M. Bhattacharjee, S. M. Prasanna, and P. Guha, "Automatic detection of shouted speech segments in indian news debates." in *Interspeech*, 2021, pp. 4179–4183.

[13] K. Phapatanaburi, L. Wang, M. Liu, S. Nakagawa, T. Jumphoo, and P. Uthansakul, "Significance of relative phase features for shouted and normal speech classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 2, 2024.

[14] 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.

[15] M. Kodali, S. Kadiri, and P. Alku, "Classification of vocal intensity category from speech using the Wav2vec2 and Whisper embeddings," in *Proc. Interspeech*, 2023, pp. 4134–4138.

[16] M. Kodali, P. Alku, and S. R. Kadiri, "AVID: Aalto vocal intensity database," May 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7948300

[17] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[18] 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.

[19] F. Javanmardi, S. R. Kadiri, and P. Alku, "Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech," *IEEE Journal of Biomedical and Health Informatics*, 2024, (in press).

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[21] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[22] A. M. A. Baevski, Y. Zhou and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[24] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.