UNCERTAINTY-BASED ENSEMBLE LEARNING FOR SPEECH CLASSIFICATION

Bagus Tris Atmaja *

AIST
Tsukuba
Japan

Felix Burkhardt

audEERING GmbH
Gilching
Germany

ABSTRACT

Speech classification has attracted increasing attention due to its wide applications, particularly in classifying physical and mental states. However, these tasks are challenging due to the high variability in speech signals. Ensemble learning has shown promising results when multiple classifiers are combined to improve performance. With recent advancements in hardware development, combining several models is not a limitation in deep learning research and applications. In this paper, we propose an uncertainty-based ensemble learning approach for speech classification. Specifically, we train a set of base features on the same classifier and quantify the uncertainty of their predictions. The predictions are combined using variants of uncertainty calculation to produce the final prediction. The visualization of the effect of uncertainty and its ensemble learning results show potential improvements in speech classification tasks. The proposed method outperforms single models and conventional ensemble learning methods in terms of unweighted accuracy or weighted accuracy.

***Index Terms*—** speech classification, speech emotion recognition, ensemble learning, uncertainty quantification

1. INTRODUCTION

Speech classification is a task that classifies speech signals into a set of predefined classes. The task is challenging due to the variability of speech signals caused by various factors such as speaker, environment, and channel. Ensemble learning has been widely used in speech classification to address these challenges. Ensemble learning is a machine learning technique that combines multiple models to improve the system's performance. This technique is known to be effective in speech and audio classifications [1, 2, 3].

On the other hand, every scenario in life deals with uncertainties, including speech classification. Quantifying uncertainty becomes inseparable in machine learning and deep learning techniques [4]. Since the prediction of deep learning models includes uncertainty, the performance of the model measured by the accuracy might be affected by the uncertainty score. Given several models for the same task, the uncertainty of each model is different. Suppose that different models output different uncertainty values in a task. Ensembling several models is a potential solution to reduce uncertainty by choosing the lowest uncertainty value across models or weighting the prediction with uncertainty values.

We propose uncertainty-based ensemble learning methods for speech classification. The main contribution is the evaluation of four variants of uncertainties for ensemble learning. The first proposed method uses the lowest uncertainty estimates to select the most confident models for the final decision. We then add a threshold in the second method; higher uncertainties use the mean ensemble. In the third and fourth methods, we use inverse and complement of uncertainty to weigh model probabilities. We evaluate the proposed methods on a speech classification task and compare them with mean and max ensemble learning methods. We count the number of improvements (in terms of weighted and unweighted accuracies) for each method to justify which method is better than other methods.

2. METHODS

2.1. Uncertainty Quantification

Uncertainty quantification (UQ) is an important indicator for trustworthy deep learning. Previous studies have shown that uncertainty quantification can impact the performance of speech emotion recognition models [5]. Guided by that study, we expand the application of uncertainty quantification to ensemble learning for speech classification.

We approach uncertainty quantification by using the entropy of the probabilities (logits). The entropy is calculated as follows: $H(i) = -\sum_{i=1}^{N} p_i \log(p_i)$, where p_i is the result of the multiplication of logits with softmax function. We then normalized entropy values between 0 and 1 to obtain uncertainty values.

*Based on project JPNP20006 funded by NEDO Japan.

2.2. Uncertainty-based ensemble learning

We evaluated four variants of uncertainty-based ensemble learning described below.

Uncertainty lowest (ul). In this variant, we select the model with the lowest uncertainty value and infer the label based on that value.

Uncertainty threshold (ut). This variant is similar to **ul** by choosing the lowest value but below the threshold. For the uncertainty values above the threshold, labels are inferred based on the mean ensemble. The threshold is searched using a grid search from 0.11 to 0.9 with a step size of 0.01.

Uncertainty weighted (uw). We calculated the inverse of uncertainty ($1/\text{uncertainty}$) as the weights, normalized the weights per model, then multiplied each class model probability with their normalized weights and used the maximum one to infer the label.

Confidence weighted (cw). We calculated the confidence score as a complement to 1 of uncertainty ($1-\text{uncertainty}$) and then normalized them for all samples per model. Similar to **uw**, we multiply each class model probability with their normalized weights and use the maximum one to infer the label.

3. EXPERIMENTS

3.1. Tasks and Datasets

Five tasks and ten datasets are evaluated using speaker-independent criteria except for SER-EMNS and SR-RAVDESS. EMNS uses a single speaker while in SR-RAVDESS the goal is to predict the speaker ID.

Speech emotion recognition (SER) is a task to classify speech signals into a category of emotions. For this SER tasks, we evaluated IEMOCAP [6], EMNS [7], TurEV [8], KBES [9], Polish [10], and TTH [11] datasets. The choices of these datasets are based on the purpose of the ensemble learning to improve performance scores. We choose datasets with unweighted accuracies around 70% or lower from the previous study [1]. We evaluated four emotion categories except for the Polish dataset, which contains only three categories ('anger,' 'neutral,' and 'fear'). The four emotions are 'neutral,' 'happiness,' 'anger,' and 'sadness' for IEMOCAP, EMNS, KBES, and TTH, and 'angry,' 'calm,' 'happy,' and 'sad' for TurEV. EMNS dataset is balanced using SMOTE ([12]) during the training process, while others are kept in the original forms.

Non-verbal emotion recognition (NVER) is a task similar to SER, but instead of verbal speech signals, it contains non-verbal voice signals such as crying, laughing, screaming, or other non-verbal phrases. We evaluated two datasets for the NVER task, VIVAE [13] and JNV [14]. We evaluated four emotion categories, i.e., 'anger,' 'fear,' 'pleasure,' and 'surprise,' for VIVAE and 'angry,' 'disgust,' 'surprise,' and 'sad,' for JNV.

Speaker recognition (SR) determines who is speaking in an audio signal. We evaluated RAVDESS [15], which is a dataset of 24 actors intended to perform ten different emotions.

Gender prediction (GP) is a task to predict the gender of a speaker, male or female. We evaluated the RAVDESS (12 male and 12 female) dataset [15], which is already annotated with gender information.

Laughter classification (LC) is a task to classify laughter into different categories, happy and evil laughter data, distinguishing "laugh with" from "being laughed at." We evaluated the EvilLaughter database [16], sampled from Audioset [17] with examples of happy and evil laughter occurrences. The dataset contains 90 samples, each 45 per laughter category.

3.2. Acoustic Features

We evaluated three acoustic features (audmodel, hubert, and wavlm) for the SER and NVER experiments and two acoustic features (os, praat) for the SR, GP, and LC experiments. Description of the acoustic features are as follows:

audmodel [18] is a finetuned wav2vec2-large-robust model on the MSP-PODCAST dataset [19]. The feature is known to be useful for SER tasks [20], as well as vocal bursts emotion recognition [21]. We used audmodel as feature extractor to extract 1024-dimensional feature vector for each speech signal.

hubert [22] is self-supervised speech representation learning by masked prediction of hidden units. Specifically, we used the *hubert-large-ll60k* version (without finetuning) to extract a 1024-dimensional feature vector for each speech signal. This feature type achieved the highest SER score in the 2021 SUPERB benchmark [23].

wavlm is a large-scale self-supervised pre-training for full stack speech processing [24]. We evaluated the large version of WavLM for SER and NVER tasks. This feature has also achieved top performance on major SER datasets [25].

os (openSMILE) is the Munich versatile and fast open-source audio feature extractor [26]. Specifically, we used the eGeMAPSv0 version [27] to extract an 88-dimensional feature vector for each speech signal.

praat is a software package for doing phonetic analysis of speech [28]. We used Praat via parselmouth-praat [29] to extract pitch, the standard deviation of pitch, hnr, jitter, shimmer, formants, and other features of each speech signal. The feature dimension is 39.

All features are scaled using a standard scaler before being used for classification.

3.3. Classifiers

A Support Vector Machine (SVM) for classification has been used for all tasks with regularization is set to 1.0 for all tasks.

[1] https://github.com/felixbur/nkululeko/tree/main/data

3.4. Evaluation Metrics

All tasks are evaluated with unweighted accuracy (UA) and weighted accuracy (WA) scores. UA equals to unweighted average recall (UAR) or balanced accuracy while WA equals to overall accuracy. In the case of class-balanced distribution (of target class), like in SR and GP tasks, score of UA and WA are the same.

We implemented dataset processing, acoustic feature extraction, classifier, and evaluation metrics using the Nkululeko toolkit [30]. Samples of configuration files to perform the experiments in this study are openly accessible [2].

4. RESULTS AND DISCUSSION

4.1. Visualizing Effect of Uncertainty

We can observe the effect of using uncertainty calculation on the correctness of predictions to gain insight into the usefulness of uncertainty quantification in speech classification. We found on the evaluated datasets the Cohen's D effect could be categorized into three: no effect, middle effect, and large effect (Fig. 1), with most data having a large effect. Given the large effect of Cohen's D, we expect that utilizing uncertainty calculation could be useful in ensemble learning.

The size of Cohen's D indicates the correlation between the correctness of predictions and the uncertainty values. The larger the effect size, the higher the correlation, meaning the uncertainty values are more reliable indicators of prediction correctness. Datasets with large Cohen's D effect show promising potential for using uncertainty-aware ensemble methods to improve overall classification performance.

Fig. 1 for the large effect (bottom) can be used to explain the concept of uncertainty. If we only accept uncertainty less than 0.5, we would get 100% accuracy by choosing the lowest uncertainty score to infer the label, but the number of data will be small. This is why we propose to use a threshold. Based on our empirical study, the mean ensemble works better than other ensembles to infer labels for uncertainty values above the threshold.

4.2. Baseline: Single Models

We measure the performance of single models to compare them with ensemble learning. The performance results of a single model are shown in Table 1. The audmodel generally performs well across SER and NVER tasks. WavLM obtained the highest scores on SER-KBES and SER-Polish. Performance varies significantly across datasets, suggesting that some are more challenging than others. For NVER tasks, performance on the JNV dataset is notably higher than on the VIVAE dataset. The 'os' approach generally shows higher performance than 'praat' for the SR, GP, and LC tasks; for the

[2] https://github.com/bagustris/nkululeko_ensemble_speech_classficcation

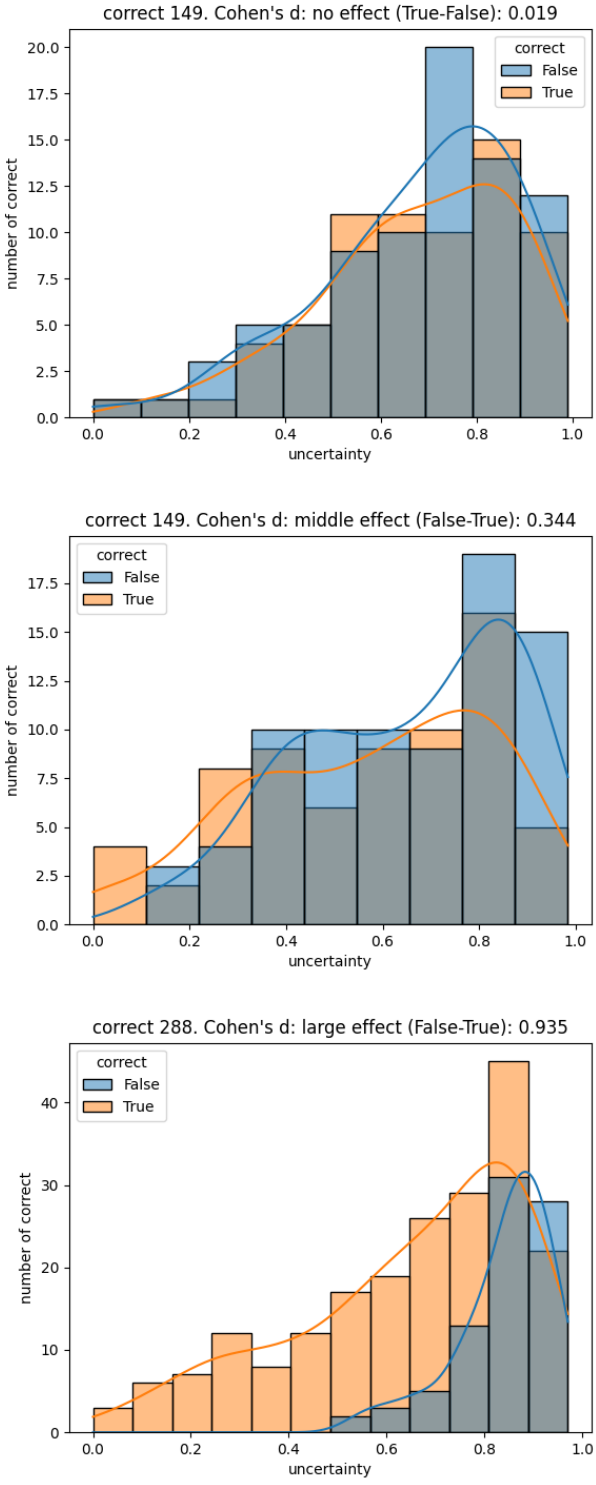

Fig. 1. Samples Cohen's D effect with no effect (top, EMNS with os), middle effect (middle, KBES with wavlm), and large effect (bottom, SR-RAVDESS with praat)

GP task, both 'os' and 'praat' approaches achieve comparable high performance. The LC task shows lower performance compared to other tasks, with significant differences between UA and WA scores, perhaps due to the low amount of data.

Table 1. Performance of single models; bold are the highest

Task-dataset	audmodel		hubert		wavlm	
	UA	WA	UA	WA	UA	WA
SER-						
IEMOCAP	**75.1**	**73.5**	69.9	69.1	73.1	73.1
EMNS	**53.7**	**58.3**	48.3	51.0	40.1	46.3
TurEV	**57.9**	**57.8**	36.8	36.8	50.9	50.9
KBES	77.5	80.0	75.8	78.0	**79.2**	**81.9**
Polish	65.5	65.5	57.6	56.7	**68.8**	**68.8**
TTH	**47.5**	**81.2**	43.9	79.4	45.1	80.2
NVER-						
VIVAE	**63.6**	**62.9**	57.5	56.9	59.6	58.9
JNV	**81.7**	**79.4**	74.1	70.5	74.2	67.6

	os		praat	
	UA	WA	UA	WA
SR	**92.7**	**92.7**	71.5	71.5
GP	**98.9**	**98.9**	97.2	97.2
LC	**50.0**	21.4	43.9	**50.0**

4.3. Ensemble Learning Results

Table 2 presents results for ensemble learning, both uncertainty-based and conventional methods (mean, max). The uncertainty-based approaches (ul, ut, uw, cw) generally perform better than single models, with **ut** and **uw** performing the best on average across tasks. Conventional ensembles (mean, max) also improve over single models but are often outperformed by uncertainty-based approaches. Based on the occurrences of high scores compared to single models, **ut** and **uw** achieved the most improvements of combined UA an WA scores (17 higher scores, in bold), followed by **ul** (15 high scores), **cw** and mean (14), and max (12).

In many cases, the combination of all three modalities (aud+hub+wav) tends to perform better than individual modality pairs, suggesting that multimodal approaches can be beneficial for these tasks. However, in such cases as SER-EMNS, SER-TurEV, and SER-KBES, the use of two models for fusion leads to better results compared to three models fusion. This indicates a need to explore different model combinations for optimal ensemble learning.

For each task, different results are obtained. KBES dataset obtained the stronger results for SER. NVER tasks are evaluated on VIVAE and JNV datasets, with JNV showing higher accuracies overall. The SR-RAVDESS and GP-RAVDESS tasks show very high accuracies, with GP achieving perfect scores across all metrics for all ensemble methods. The LC-Laughter task shows lower accuracies compared to other tasks, indicating it might be a more challenging classification problem.

Finally, although most datasets show large Cohen's D effects when comparing the distribution of correctness from uncertainty calculation, not all datasets show improvement when using uncertainty-based fusion methods, e.g., the TTH dataset. The calculation of UQ, which is based on entropy, can be extended to other methods, e.g., Monte Carlo dropout, to further observe uncertainty estimation and fusion performance. The ensemble learning could also be extended to regression tasks in addition to speech classification.

5. CONCLUSIONS

In this paper, we investigated uncertainty-based ensemble learning for speech classification. We evaluated four variants of uncertainty-based ensemble learning methods and compared them with single models and conventional ensemble learning methods. The experimental results show that the uncertainty threshold (**ut**) and uncertainty weighted (**uw**) methods obtained higher gain than other methods in terms of the quantity of improved UA/WA scores. The threshold in **ut**, however, is searched manually for each dataset, which is impractical for real applications. The uncertainty-weighted method, on the other hand, did not require manual threshold selection and achieves similar performance to the **ut** method.

6. REFERENCES

[1] Nicolae Cătălin Ristea and Radu Tudor Ionescu, "Self-paced ensemble learning for speech and audio classification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2, pp. 1276–1280, 2021.

[2] Po-Yuan Shih, Chia-Ping Chen, and Chung-Hsien Wu, "Speech Emotion Recognition With Ensemble Learning Methods," in *IEEE Int. Conf. Acoust. Speech, Signal Process. 2017*, 2017, pp. 2756–2760.

[3] Bagus Tris Atmaja and Akira Sasou, "Ensembling Multilingual Pre-Trained Models for Predicting Multi-Label Regression Emotion Share from Speech," in *2023 Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. oct 2023, pp. 1026–1029, IEEE.

[4] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, 2021.

[5] Oliver Schrüfer, Manuel Milling, Felix Burkhardt, Florian Eyben, and Björn Schuller, "Are you sure? Analysing Uncertainty Quantification Approaches for Real-world Speech Emotion Recognition," in *INTERSPEECH 2024*, 2024.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang,

Table 2. Performance of ensemble learning; ul: uncertainty-lowest, ut:uncertainty-threshold, uw: uncertainty-weighted, cw: confidence-weighted; aud:audmodel, hub:hubert, wav:wavlm; bold: higher than the best single model

Task-dataset	ul		ut		uw		cw		mean		max	
	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA
SER-IEMOCAP												
aud+hub	**75.3**	**75.8**	**75.5**	**74.4**	**75.4**	**75.2**	**75.5**	**74.3**	**75.4**	**74.3**	**75.2**	**74.0**
hub+wav	72.3	71.8	72.4	**74.4**	72.1	71.8	71.9	71.6	72.1	71.8	72.0	71.6
aud+hub+wav	**76.2**	**74.9**	**76.2**	**75.2**	**75.9**	**75.1**	**76.1**	**75.2**	**75.4**	**74.8**	**75.7**	**74.7**
SER-EMNS												
aud+hub	51.1	56.4	50.4	55.7	50.4	55.7	51.2	56.4	50.4	55.7	52.0	57.0
hub+wav	53.2	**58.4**	42.7	48.3	**57.0**	**62.4**	**55.8**	**61.1**	**57.4**	**62.4**	**55.8**	**56.1**
aud+hub+wav	49.4	55.0	50.6	56.4	50.6	56.4	50.2	55.7	49.8	55.0	52.4	57.7
SER-TurEV												
aud+hub	57.3	57.3	**58.8**	**58.8**	**58.2**	**58.2**	57.9	57.9	**58.2**	**58.2**	57.6	57.6
hub+wav	47.6	47.6	47.6	47.6	47.0	47.0	46.6	46.6	46.3	46.3	47.3	47.3
aud+hub+wav	57.3	57.3	57.9	57.9	57.6	57.6	57.6	57.6	56.1	56.1	57.3	57.3
SER-KBES												
aud+hub	78.3	81.9	79.2	**82.9**	79.2	**82.9**	79.2	**82.9**	79.2	**82.9**	78.3	81.9
hub+wav	79.2	81.8	79.2	81.9	79.2	81.9	79.2	81.9	79.2	81.9	79.2	81.9
aud+hub+wav	78.3	81.9	78.3	81.9	78.3	81.9	78.3	81.9	77.5	81.0	78.3	81.9
SER-Polish												
aud+hub	65.6	65.6	67.8	67.8	67.8	67.8	67.8	67.8	67.8	67.8	67.8	67.8
hub+wav	67.8	67.8	67.8	67.8	66.7	66.7	67.8	67.8	67.8	67.8	67.8	67.8
aud+hub+wav	67.8	67.8	**68.9**	**68.9**	**68.9**	**68.9**	67.8	67.8	67.8	67.8	67.8	67.8
SER-TTH												
aud+hub	45.6	80.5	45.6	80.5	45.6	80.6	45.6	80.6	45.6	80.7	45.5	80.6
hub+wav	44.4	79.8	44.6	79.9	44.1	79.9	44.1	79.9	44.1	79.9	44.1	79.9
aud+hub+wav	45.6	80.6	45.6	80.6	45.1	80.4	45.1	80.5	44.9	80.4	45.4	80.5
NVER-VIVAE												
aud+hub	**67.9**	**67.5**	**68.6**	**68.2**	**68.6**	**68.2**	**67.9**	**67.5**	**68.6**	**68.2**	**67.3**	**66.9**
hub+wav	**67.9**	**67.5**	**64.9**	**64.2**	**64.9**	**64.2**	63.1	62.3	**64.9**	**64.2**	63.1	62.3
aud+hub+wav	**68.2**	**67.5**	**68.8**	**68.2**	**68.2**	**67.5**	**69.5**	**68.9**	**68.2**	**67.5**	**68.1**	**67.5**
NVER-JNV												
aud+hub	**84.4**	**85.3**	**84.4**	**85.3**	**83.1**	**82.4**	**83.1**	**82.4**	**83.1**	**82.4**	**83.1**	**82.4**
hub+wav	75.6	70.6	79.5	79.4	75.6	70.6	75.6	70.6	75.6	70.6	75.6	70.6
aud+hub+wav	80.8	**82.4**	**81.8**	79.4	**81.8**	79.4	**83.1**	**82.4**	78.2	76.5	79.5	79.4
SR-RAVDESS												
os+praat	92.7	92.7	**94.4**	**94.4**	**94.1**	**94.1**	**94.1**	**94.1**	**94.4**	**94.4**	**93.1**	**93.1**
GP-RAVDESS												
os+praat	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
LC-Laughter												
os+praat	**54.5**	28.6	**59.1**	35.7	25.8	21.4	42.4	28.6	34.8	35.7	34.8	35.7

Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[7] Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang, "EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels," 2023.

[8] Salih Firat Canpolat, Zuhal Ormano\uglu, and Deniz Zeyrek, "Turkish Emotion Voice Database (TurEV-DB)," in *Proc. 1st Jt. Work. Spok. Lang. Technol. Under-resourced Lang. Collab. Comput. Under-Resourced Lang.*, 2020, number May, pp. 368–375.

[9] Md Masum Billah, Md Likhon Sarker, and M. A.H. Akhand, "KBES: A dataset for realistic Bangla speech emotion recognition with intensity level," *Data Br.*, vol. 51, pp. 109741, 2023.

[10] Marzena Mięsikowska and Dariusz Świsulski, "Emotions in polish speech recordings," 2020.

[11] Ngoc Anh Nguyen Thi, Bao Thang Ta, Nhat Minh Le, and Van Hai Do, "An Automatic Pipeline For Building Emotional Speech Dataset," *2023 Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2023*, pp. 1030–1035, 2023.

[12] Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, sep 2016.

[13] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel, "The variably intense vocalizations of affect and emotion (VIVAE) corpus prompts new perspective on nonspeech perception.," *Emotion*, vol. 22, no. 1, pp. 213–225, feb 2022.

[14] Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari, "JNV corpus: A corpus of Japanese nonverbal vocalizations with diverse phrases and emotions," *Speech Commun.*, vol. 156, no. October 2023, pp. 103004, 2024.

[15] S.R. Livingstone and F.A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS One*, pp. 1–35, 2018.

[16] Aljoscha Düsterhöft, Felix Burkhardt, and Björn W. Schuller, "Happy or Evil Laughter? Analysing a Database of Natural Audio Samples," 2023.

[17] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events,In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 776-780, doi: 10.1109/ICASSP.2017.7952261.," 2016.

[18] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, sep 2023.

[19] Reza Lotfian and Carlos Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, 2019.

[20] Bagus Tris Atmaja and Akira Sasou, "Multilingual, Cross-lingual, and Monolingual Speech Emotion Recognition on EmoFilm Dataset," in *2023 Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. oct 2023, pp. 1019–1025, IEEE.

[21] Bagus Tris Atmaja and Akira Sasou, "Evaluating Variants of wav2vec 2.0 on Affective Vocal Burst Tasks," in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* jun 2023, pp. 1–5, IEEE.

[22] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[23] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, ISCA, aug 2021, pp. 1194–1198, ISCA.

[24] Sanyuan Chen, Chengyi Wang, Zhuo Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Zhengyang Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Yao Qian, Yanmin Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2021.

[25] Bagus Tris Atmaja and Akira Sasou, "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," *IEEE Access*, vol. 10, pp. 124396–124407, 2022.

[26] Florian Eyben, Felix Weninger, Martin Wollmer, Björn Bjorn Schuller, Martin Wöllmer, and Björn Bjorn Schuller, *OpenSMILE - The Munich versatile and fast open-source audio feature extractor*, Number December. 2015.

[27] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[28] Paul Boersma and Vincent van Heuven, "Speak and unSpeak with Praat," *Glot Int.*, vol. 5, no. 9-10, pp. 341–347, 2001.

[29] Yannick Jadoul, Bill Thompson, and Bart de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, no. 2018, pp. 1–15, 2018.

[30] Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn W. Schuller, "Nkululeko: A Tool For Rapid Speaker Characteristics Detection," in *Proc. Lr.*, 2022, pp. 395–410.