Performance-weighted Ensemble Learning for Speech Classification

Bagus Tris Atmaja
AIST
Tsukuba, Japan
b-atmaja@aist.go.jp

Akira Sasou
AIST
Tsukuba, Japan
a-sasou@aist.go.jp

Felix Burkhardt
audEERING GmbH
Gilching, Germany
fburkhardt@audeering.com

***Abstract*—Ensemble learning is a useful technique to combine several models to improve classification performance. Previous research on speech classification shows the benefit of ensemble learning over single models; however, there is no systematic evaluation on the accommodating single model performance in ensemble learning for speech classification. There is also no detailed report on the usefulness of ensemble learning over task-specific acoustic feature. We evaluate performance-weighted ensemble learning by taking into account the previous single model performance for speech classification tasks, including speech emotion recognition, laughter type classification, gender prediction and age prediction. We compare different weighting schemes based on unweighted and weighted accuracies, in which we also reported our results using these metrics. Results on six tasks and eleven datasets show diverse findings on the effectiveness of performance-weighted ensemble learning over other ensemble methods and single models.**

***Index Terms*—speech classification, ensemble learning, performance-weighted ensemble learning, acoustic feature**

I. Introduction

Speech classification is an emerging task in speech processing with various applications such as speaker identification, gender prediction, intent classification, and speech emotion recognition. Speech classification is a subset of audio signal classification (ASC) or audio classification that focuses on classifying audio segments. The aim of speech classification is to predict the class label of an utterance based on acoustic features extracted from the audio signal.

Classical audio classification classically was performed using Hidden Markov Models (HMM), k-means clustering, and *neural nets* [1]. Among many aspects, which features will be relevant for the tasks is an important decision. Classifier is another important factor; in [2], the authors showed the benefit of CNN-based architecture for audio classification over a simple fully connected network.

Recent development of deep learning moved the way of speech classification to simpler approach using neural networks. An acoustic features are still required to represent the audio signal, or the system could directly extract information from raw acoustic signals. Self-supervised learning now is the *de-facto* acoustic feature extractor which leads to state-of-the-art (SOTA) performance in many tasks including speech emotion recognition [3], speaker identification/recognition, speaker verification, and intent classification [4]. In finetuning, the model usually takes the raw audio input to predict the class label.

Ensemble learning has been shown to improve the performance of speech classification tasks by combining predictions from multiple classifiers. Using classical feature extractor like zero crossing rate and MFCC, utilizing ensemble learning would significantly improve speech emotion recognition [5] and audio classification [6]. In this paper, we propose a performance-weighted ensemble learning method for speech classification where the predictions are weighted based on the performance of individual classifiers. The weights can be from accuracy, F1-score or other evaluation metrics. In this case, we evaluated unweighted and weighted accuracies (UA and WA). The weighted predictions are then summed to obtain the final prediction. This approach leverages the strengths of individual classifiers while mitigating their weaknesses.

The contribution of this paper is four-folds relative to the previous study [7]. First, we run the ensemble of audmodel and wavlm which is missing in the previous experiments. Second, we propose a new ensemble learning method called performance-weighted ensemble learning and evaluate two variants of performance-weighted values based on unweighted and weighted accuracies. Third, we re-run EvilLaughter experiments for unimodal by varying feature scaler and kernel type to further improve the baseline for unimodal prediction. Finally, we added age classification with the EmoDB dataset to evaluate the usefulness of task-specific acoustic feature over the ensemble methods.

II. Methods

Ensemble learning has been widely adopted as a way to improve predictive performance by combining multiple models, usually from different modalities. Although, combining several modalities from different inputs is also possible as will be shown in this paper. We introduce a performance-weighted ensemble that assigns weights to individual models based on their performance on the previous validation data. This fusion method is similar to [8] but for classification. The weights could be from unweighted accuracy or weighted accuracy in 0-1 scale. The illustration and pseudocode are provided in Figure 1 and Table I respectively.

Suppose we have three different models with predictions and performances (Fig. 1). The performances are designated

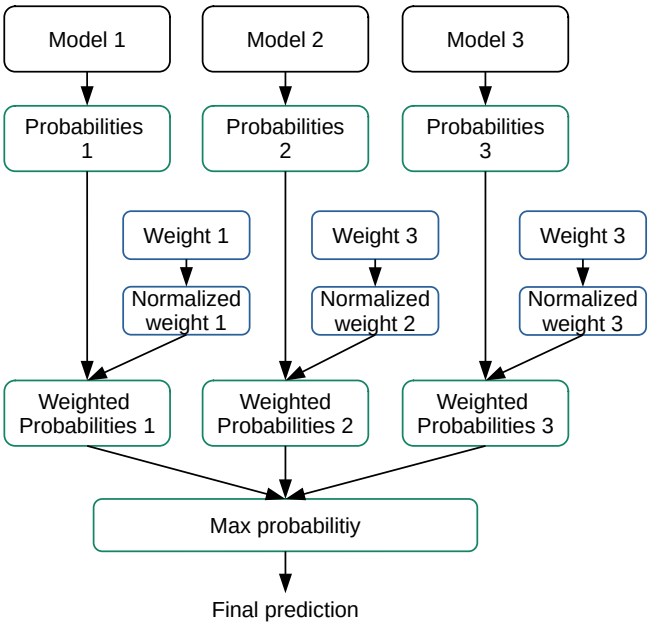

Fig. 1. Illustrative example of performance-weighted ensemble predictions with three models.

as weights. The weights are normalized according the total weights. This normalized weight then is multiplied to each class probability of each model and summed together to get the final ensemble prediction. The final prediction is the class with the highest probability.

III. Experiments

A. Tasks and Datasets

We evaluated the proposed performance-weighted ensemble learning on five tasks and ten datasets. The tasks are speech emotion recognition (SER), non-verbal emotion recognition (NVER), gender prediction (GP), speaker recognition (SR), and laughter classification (LC). The datasets are IEMOCAP [9], EMNS [10], TurEV [11], KBES [12], Polish [13], TTH [14], VIVAE [15], JNV [16], RAVDESS [17], and EvilLaughter [18] datasets. Details of the dataset configuration can be referred to the previous study [7]. We added EmoDB [19] for age classification task in this study. Based on the distribution of the age (Figure 2), we grouped the data into two categories: under 30 and 30es.

B. Acoustic Features

We evaluated ensemble of the same classifier with different features. The classifier is SVM for classification with C value of 1.0 and RBF kernel, except for laughter classification which uses linear kernel. We evaluated the following model fusions for performance-weighted ensemble:

- aud+hub: combination of two SVM models from audmodel [20] hubert-large-ll60k [21] features.
- hub+wav: combination of two SVM models from hubert-large-ll60k and wavlm-large [22] features.

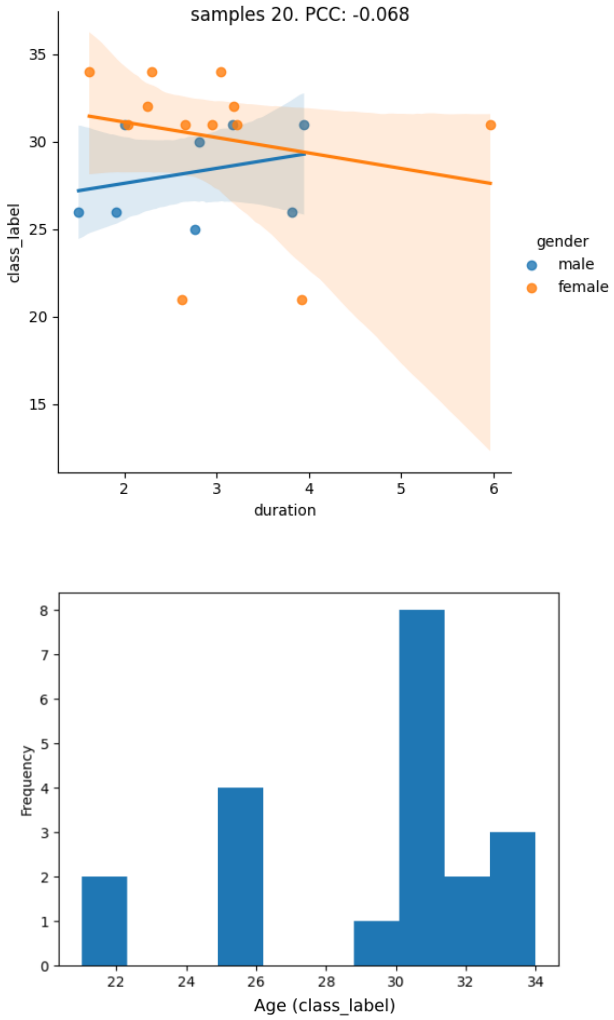

Fig. 2. Distribution of age in duration (top) and samples (bottom) in EmoDB

- aud+wav: combination of two SVM models from audmodel and wavlm features.
- aud+hub+wav: combination of three SVM models from audmodel, hubert-large-ll60k, and wavlm-large features.
- os+praat: combination of two SVM models from **O**pen**S**MILE [23] with eGeMAPSv02 feature set [24] and praat [25]–[27] features.
- agender+os/praat: combination two SVM models from agender feature [28] and os or praat features.
- os+praat+agender: combination of three SVM models from OpenSmile, praat, and agender [28] features.

The dimension (size) of aud, hub, wav, and agender features are 1024. For os and praat, the dimensions are 88 and 39 respectively.

C. Evaluation Metrics

We reported unweighted accuracy (UA) and weighted accuracy (WA) to measure the performance of ensemble learning, as well as individual model performance for comparison for LC and AC-EmoDB. Others individual performances can be

TABLE I
PSEUDOCODE CODE FOR PERFORMANCE-WEIGHTED ENSEMBLE METHOD

```
FUNCTION performance_weighted_ensemble(ensemble_preds_ls, labels, weights):
Initialize empty lists: final_predictions, final_confidences

ASSERT all weights are between 0 and 1
ASSERT number of weights equals number of models

Normalize weights:
    total_weight = sum of all weights
    FOR EACH weight:
        weight = weight / total_weight

FOR EACH idx in indices of first prediction dataframe:
    Initialize dictionary class_probabilities with labels as keys and 0 as values

    FOR EACH df, weight in zip(ensemble_preds_ls, weights):
        GET row from df at idx
        FOR EACH label in labels:
            class_probabilities[label] += row[label] * weight

    predicted_class = label with maximum value in class_probabilities
    APPEND predicted_class to final_predictions
    APPEND maximum value from class_probabilities to final_confidences

RETURN final_predictions (and optionally final_confidences if needed)
```

traced back in the previous study [7]. WA treats all classes equally while UA accounts for class imbalance by calculating average accuracy per class.

IV. RESULTS AND DISCUSSION

We present our result in Table II which summarize the performance of mean, uncertainty-based, and performance-weighted ensemble learning on the ten datasets. The most notables results are that the performance-weighted ensemble learning achieves new state-of-the-art results (underlined) on SER-IEMOCAP, SER-TurEV, and GP-RAVDESS datasets. The performance-weighted ensemble learning also achieves comparable results to the uncertainty-based ensemble learning on other datasets. All SOTA are achieved using ensemble learning of two feature sets, suggesting task-specific feature fusion is important for ensemble learning.

Note that for TurEV dataset the new state-of-the-result results are obtained by the ensemble of audmodel and hubert features. Although this result is lower than the reported UA (76%) in the reference paper [11], we argue that our results are more reliable since we use speaker-independent criteria. There is no information regarding the speaker independence in the previous study that leads to assumption that speaker-dependent evaluation might have been used.

We also conducted two paired sample tests to claim the significant different about two means between UA-weighted and WA-weighted ensemble learning. The results show that the $p - values$ for UA-weighted and WA-weighted are more than 0.05 (P for one tail are 0.41 and 0.10 for UA and WA), meaning the difference is not significant. However, the variance of WA-weighted is smaller than UA-weighted, indicating WA-weighted may produce more stable predictions across datasets.

For EvilLaughter dataset (LC), we re-run baseline models using os and praat with some modifications from the previous study [7]. First, both features use SVM model with linear kernel instead of RBF. Second, we did not scale the feature for os as it degrades the performance. The unimodal model with os and SVM achieves new state-of-the-result with UA of 76.5 and WA of 73.9%. However, the ensemble model still cannot outperform the unimodal model similar to the previous study [7].

The similar negative ensemble results were also observed for AC-EmoDB dataset. The single model result achieves higher score with 60% and 67% of UA and WA for agender features. Combination of os, praat, and agender could not surpass the UA of single agender model. This result suggests that for some datasets, specific feature tuned for specific task may more important than ensemble of multiple features.

V. CONCLUSIONS

In this paper, we presented performance-weighted ensemble learning for speech classification. Evaluation using unweighted accuracy and weighted accuracy as weights did not show significant different, meaning both can be used alternately. The results showed comparable performances to previous studies using uncertainty-based ensemble learning and achieved state-of-the-art results on IEMOCAP and TurEV datasets for unimodal speech emotion recognition and RAVDESS dataset for gender prediction. We also reported new results on laughter type classification, although this high score is achieved without ensemble learning. The best result

TABLE II

ENSEMBLE PERFORMANCE COMPARISON. BOLD: HIGHER THAN THE BEST SINGLE MODEL. UNDERLINE: NEW HIGHEST. UNCERTAINTY IS OBTAINED FROM THE HIGHEST SCORE AMONG FOUR VARIANTS IN THE PREVIOUS STUDY [7], EXCEPT FOR AUD+WAV WHICH USED **UW**

Task-dataset, features	Mean		Uncertainty		UA-weighted		WA-weighted	
	UA	WA	UA	WA	UA	WA	UA	WA
SER-IEMOCAP								
aud+hub	**75.4**	**74.3**	**75.5**	**75.8**	**75.6**	**74.5**	**75.6**	**74.5**
hub+wav	72.1	71.8	**72.4**	**74.4**	72.3	72.0	72.3	72.0
aud+wav	**76.9**	**75.8**	**76.9**	**75.8**	**77.0**	**75.9**	**76.8**	**75.8**
aud+hub+wav	**75.4**	**74.8**	**76.2**	**75.2**	**75.7**	**75.0**	**75.5**	**74.9**
SER-EMNS								
aud+hub	50.4	55.7	51.2	56.4	50.4	55.7	50.8	56.4
hub+wav	**57.4**	**62.4**	**57.0**	**62.4**	42.3	47.7	42.3	47.7
aud+wav	**57.4**	**62.4**	**57.0**	**62.4**	**56.9**	56.4	51.9	57.7
aud+hub+wav	49.8	55.0	**50.6**	56.4	**50.6**	56.4	**50.6**	56.4
SER-TurEV								
aud+hub	**58.2**	**58.2**	**58.8**	**58.8**	**60.7**	**60.7**	**60.7**	**60.7**
hub+wav	46.3	46.3	47.6	**47.6**	47.0	47.0	47.0	47.0
aud+wav	57.3	57.3	**58.5**	**58.5**	**57.9**	57.9	**57.9**	57.9
aud+hub+wav	56.1	56.1	**57.9**	**57.9**	56.7	56.7	56.7	56.7
SER-KBES								
aud+hub	79.2	**82.9**	**79.2**	**82.9**	75.8	78.1	**79.2**	**82.9**
hub+wav	79.2	81.9	79.2	81.9	79.2	75.8	79.2	81.9
aud+wav	79.2	81.9	79.2	81.9	79.2	81.9	79.2	81.9
aud+hub+wav	77.5	81.0	78.3	81.9	77.5	81.0	77.5	81.0
SER-Polish								
aud+hub	67.8	67.8	67.8	67.8	66.7	66.7	66.7	66.7
hub+wav	67.8	67.8	67.8	67.8	66.7	66.7	66.7	66.7
aud+wav	65.6	65.6	64.4	64.4	65.6	65.6	65.6	65.6
aud+hub+wav	67.8	67.8	**68.9**	**68.9**	**68.9**	**68.9**	**68.9**	**68.9**
SER-TTH								
aud+hub	45.6	80.7	45.6	80.6	45.6	80.7	45.6	80.7
hub+wav	44.1	79.9	44.6	79.9	44.1	79.9	44.1	79.9
aud+wav	46.5	81.2	46.4	81.1	46.5	**81.3**	46.5	81.2
aud+hub+wav	44.9	80.4	45.6	80.6	44.9	80.4	44.9	80.4
NVER-VIVAE								
aud+hub	**68.6**	**68.2**	**68.6**	**68.2**	**69.3**	**68.9**	**69.3**	**68.9**
hub+wav	**64.9**	**64.2**	**67.9**	**67.5**	**69.3**	**68.9**	64.9	**64.2**
aud+wav	**68.2**	**67.5**	**68.2**	**67.5**	**68.2**	**67.5**	**68.8**	**68.2**
aud+hub+wav	**68.2**	**67.5**	**69.5**	**68.9**	**68.2**	**67.5**	**68.2**	**67.5**
NVER-JNV								
aud+hub	**83.1**	**82.4**	**84.4**	**85.3**	**83.1**	**82.4**	**83.1**	**82.4**
hub+wav	75.6	70.6	**79.5**	79.4	78.2	76.5	78.2	76.5
aud+wav	78.2	76.5	78.2	76.5	78.2	76.	78.2	76.5
aud+hub+wav	78.2	76.5	**83.1**	**82.4**	78.2	76.5	78.2	76.5
GP-RAVDESS								
os+praat	**94.4**	**94.4**	**94.4**	**94.4**	**99.3**	**99.3**	**99.3**	**99.3**
SR-RAVDESS								
os+praat	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
LC-Laughter								
os+praat	69.4	65.2	69.4	65.2	69.4	65.2	73.0	69.6
AC-EmoDB								
os+praat	47.7	64.1	47.7	64.1	46.9	42.8	47.5	63.6
praat+agender	48.4	63.2	48.4	63.2	48.4	64.1	48.4	63.2
os+agender	47.0	60.2	46.7	59.7	46.7	59.7	47.0	60.2
os+praat+agender	49.1	67.1	46.0	62.3	48.6	65.4	49.1	67.1

for age classification is also obtained from single model with agender feature which is specially tuned for age and gender recognition.

Future research could be focused on the following aspects: (1) investigating the effectiveness of performance-weighted ensemble learning on other tasks and datasets, (2) exploring other weighting schemes for performance-weighted ensemble learning, and (3) evaluating the performance of ensemble learning on other modalities such as text and image.

ACKNOWLEDGMENTS

This paper is partly based on results obtained from projects, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan and JSPS KAKENHI Grant Number 24K0296.

REFERENCES

[1] D. Gerhard, "Audio Signal Classification: History and Current Techniques," 2003.

[2] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2017, pp. 131–135.

[3] B. T. Atmaja and A. Sasou, "Evaluating Variants of wav2vec 2.0 on Affective Vocal Burst Tasks," in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, jun 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096552/

[4] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*. ISCA: ISCA, aug 2021, pp. 1194–1198.

[5] P.-Y. Shih, C.-P. Chen, and C.-H. Wu, "Speech Emotion Recognition With Ensemble Learning Methods," in *IEEE Int. Conf. Acoust. Speech, Signal Process. 2017*, 2017, pp. 2756–2760.

[6] N. C. Ristea and R. T. Ionescu, "Self-paced ensemble learning for speech and audio classification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2, pp. 1276–1280, 2021.

[7] B. T. Atmaja, A. Sasou, and F. Burkhardt, "UNCERTAINTY-BASED ENSEMBLE LEARNING FOR SPEECH CLASSIFICATION," in *Proceedings of 2024 27rd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2024*, 2024.

[8] B. T. Atmaja, "Feature-wise Optimization and Performance-weighted Multimodal Fusion for Social Perception Recognition," in *Proc. 5th Multimodal Sentim. Anal. Chall. Work. Soc. Percept. Humor*. New York, NY, USA: ACM, oct 2024, pp. 28–35. [Online]. Available: https://dl.acm.org/doi/10.1145/3689062.3689082

[9] C. Busso, M. Bulut, C.-C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[10] K. A. Noriy, X. Yang, J. J. Zhang, N. Kari, Y. Xiaosong, and Z. Jian, "EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels," mar 2023. [Online]. Available: https://github.com/knoriy/EMNS-DCT http://arxiv.org/abs/2305.13137

[11] S. F. Canpolat, Z. Ormanoglu, and D. Zeyrek, "Turkish Emotion Voice Database (TurEV-DB)," in *Proc. 1st Jt. Work. Spok. Lang. Technol. Under-resourced Lang. Collab. Comput. Under-Resourced Lang.*, no. May, 2020, pp. 368–375. [Online]. Available: https://aclanthology.org/2020.sltu-1.52

[12] M. M. Billah, M. L. Sarker, and M. A. Akhand, "KBES: A dataset for realistic Bangla speech emotion recognition with intensity level," *Data Br.*, vol. 51, p. 109741, 2023. [Online]. Available: https://doi.org/10.1016/j.dib.2023.109741

[13] M. Mięsikowska and D. Świsulski, "Emotions in polish speech recordings," 2020. [Online]. Available: https://mostwiedzy.pl/en/open-research-data/emotions-in-polish-speech-recordings,111905234611461 69-0

[14] N. A. N. Thi, B. T. Ta, N. M. Le, and V. H. 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.

[15] N. Holz, P. Larrouy-Maestri, and D. 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. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/emo0001048

[16] D. Xin, S. Takamichi, and H. Saruwatari, "JNV corpus: A corpus of Japanese nonverbal vocalizations with diverse phrases and emotions," *Speech Commun.*, vol. 156, no. October 2023, p. 103004, 2024. [Online]. Available: https://doi.org/10.1016/j.specom.2023.103004

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

[18] A. Düsterhöft, F. Burkhardt, and B. W. Schuller, "Happy or Evil Laughter? Analysing a Database of Natural Audio Samples," 2023. [Online]. Available: http://arxiv.org/abs/2305.14023

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech 2005*, 2005, pp. 1517–1520. [Online]. Available: http://www.expressive-speech.net/emodb/

[20] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. 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. 10 745–10 759, sep 2023.

[21] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[22] S. Chen, C. Wang, Z. Z. Chen, Y. Wu, S. Liu, Z. Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Y. Qian, Y. Y. Qian, J. Wu, M. Zeng, X. Yu, and F. 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.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimed. - MM '13*. New York, New York, USA: ACM Press, 2013, pp. 835–838.

[24] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. 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.

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

[26] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, no. 2018, pp. 1–15, 2018. [Online]. Available: https://doi.org/10.1016/j.wocn.2018.07.001

[27] D. R. Feinberg, "Parselmouth Praat Scripts in Python," 2018.

[28] Felix Burkhardt and Johannes Wagner and Hagen Wierstorf and Florian Eyben and Björn Schuller, "Speech-based Age and Gender Prediction with Transformers," *15th ITG Conf. Speech Commun.*, pp. 3–7, 2023.