2020 IEEE REGION 10 CONFERENCE (TENCON)
Osaka, Japan, November 16-19, 2020

Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion

Bagus Tris Atmaja
Department of Engineering Physics
Sepuluh Nopember Institute of Technology
Surabaya, Indonesia
bagus@jaist.ac.jp

Yasuhiro Hamada
School of Information Science
Japan Adv. Inst. of Sci. & Tech.
Nomi, Japan
y-hamada@jaist.ac.jp

Masato Akagi
School of Information Science
Japan Adv. Inst. of Sci. & Tech.
Nomi, Japan
akagi@jaist.ac.jp

***Abstract*—This paper presents an evaluation of acoustic feature aggregation and acoustic-linguistic features combination for valence and arousal prediction within a speech. First, acoustic features were aggregated from chunk-based processing for story-based processing. We evaluated mean and maximum aggregation methods for those acoustic features and compared the results with the baseline, which used majority voting aggregation. Second, the extracted acoustic features are combined with linguistic features for predicting valence and arousal categories: low, medium, or high. The unimodal result using acoustic features aggregation showed an improvement over the baseline majority voting on development partition for the same acoustic feature set. The bimodal results (by combining acoustic and linguistic information at the feature level) improved both development and test scores over the official baseline. This combination of acoustic-linguistic information targeted speech-based applications where acoustic and linguistic features can be extracted from the sole speech modality.**

***Index Terms*—valence, arousal, affective computing, feature aggregation, feature fusion**

I. Introduction

Predicting valence and arousal is important since many affective analyses can be performed by analyzing these two attributes. Russel [1] showed categorical emotion could be mapped in valence-arousal space. Kensinger [2] argued that valence and arousal affect memory and attention. An emotional arousal event triggers the activation of endogenous memory modulating systems [3]. Costanzi et al. found that valence and arousal effect visuospatial working memory [4]. An interplay between valence and arousal enhances spatial memory performance.

A challenge for predicting valence and arousal categories was run to improve the unweighted average recall (UAR) from the official baselines [5]. In the elderly emotion sub-challenge (ESC), the participants were invited to predict both valence and arousal categories (low, medium, or high) given the speech chunks from story-based instances. This paper is a participation report on this elderly emotion sub-challenge. The manual and automatic speech transcriptions were also provided by the organizer. The official baselines are provided for test partition from three acoustic feature sets and one linguistic feature sets with different parameters. Although the organizer combined several features to fuse the different modalities, the highest UAR score came from an acoustic feature set for arousal and a linguistic feature set for valence prediction. Details of the challenge, including the dataset, can be found in [5].

The provided baseline method for predicting valence and arousal used majority voting from the output of chunk-based predictions for a story-based prediction. A story was divided into 5-second chunks for feature extraction. Predictions from all chunks were aggregated by majority voting to obtain a final prediction for a story/instance. Besides the UAR still low, this approach has difficulty when combining acoustic and linguistic information at feature fusion. We aim to overcome this limitation by aggregating acoustic feature at the input step for combination with linguistic features.

In this experiment, we instead proposed to aggregate acoustic input features for further acoustic-linguistic combination. Acoustic features from all chunks in a single story were aggregated by either mean or maximum value. This approach enables us to fuse acoustic and linguistic features for predicting valence and arousal categories from speech. Since linguistic features can be extracted from speech via a speech-to-text system, it is reasonable to combine those acoustic and linguistic features without the addition of information from other modalities. The target application of this approach is speech-based technology like voice assistant applications.

Although combining acoustic and linguistic information is not new, to best of our knowledge, no research reported on the use of acoustic feature aggregation for acoustic-linguistic feature fusion. While others have performed information fusion at network level [6], [7], this paper reports acoustic-linguistic information fusion at the feature level. Combining information at the feature level requires one-step processing only while combining information at the decision level requires two-step processing. Since the dataset used here is small, using one type of feature set (e.g., acoustics feature only) may limit information from which valence and arousal can be extracted. Using acoustic feature aggregation for acoustic-

978-1-7281-8455-5/20/$31.00 ©2020 IEEE

linguistic feature fusion, we improved the official baseline scores on test partition from a single feature prediction.

We briefly described the dataset and features used to produce the results in this paper. A modified method from the baseline is explained along with its results. Finally, we extend the discussion in this paper to observe the mismatch between development and test scores given in the baseline. A future study is needed to tackle the limitation of current reported results, including the results in this paper.

II. Related Work and Contribution

Recognizing valence from speech can be performed using acoustic features, linguistic features, or a combination of both. Combining acoustic and linguistic information for speech emotion recognition can be cast into two groups, early and late fusion. In the early fusion, either features or classifiers (e.g., networks) are fused at the input level. In the late fusion, the outputs of classifiers are combined at the decision level.

Yang and Hirscherberg [8] combined raw waveform and spectrogram of speech to recognize continuous degrees of valance and arousal via deep neural networks (DNN). The paper reported that combining both information from speech provides further improvement to the performance from a single input. Although obtained high concordance correlation coefficient (CCC) score for arousal, the CCC score on valence is low which is common on acoustic-only dimensional speech emotion recognition.

In [9], the authors used feature fusion to combine acoustic and linguistic information for real-time affect recognition in a mobile application. The authors concatenated 89 acoustic features with seven linguistic features and reported improvements in predicting six basic emotions over unimodal features. The highest reported performance was obtained using Logistic Model Tree.

In [6], [10], a combination of acoustic and linguistic information was performed at the networks/classifiers level. Both papers proposed recurrent-based networks concatenation with the attention layer. Both papers also reported performance improvements obtained by bimodal networks over unimodal networks. While paper [10] evaluated the combined acoustic-linguistic networks on a single dataset, the authors of [6] evaluated the bimodal networks on several datasets with a different number of emotion categories.

Griol et al. combined speech and linguistic emotion classification at decision level [11]. The paper reported an evaluation of three different decision fusion: majority voting, classification scores, and Borda count. The result revealed that Borda count provided the best results. A fusion strategy beyond the decision level was proposed by Tian et al. [12]. The paper proposed a hierarchical fusion to incorporate features at different levels of its knowledge-inspired structure. This model is close to the feature fusion approach with a multi-layer strategy. Different layers received inputs from different features and the previous layers. The authors reported improvement using the proposed method over early and late fusion methods on two emotional datasets.

TABLE I
Number of instances and chunks in each partition

Partition	# Stories (text)	# Chunks (audio)
Train	87	2496
Dev	87	2466
Test	87	2816
Total	261	7778

Our proposed approach differs from the previous ones by the nature of the dataset. Our target is predicting valence and arousal categories, while the previous research aimed at predicting emotion categories. For processing audio and text, it is sound to process text on story-based processing while the audio on chunk-based processing. Text data, with respect to text feature, is relatively smaller than audio data in size for the same utterance. Hence, processing acoustic feature on chunk-based processing, as given by the dataset, is more relevant than processing audio on a whole story.

In summary, the contributions of this paper can be divided into the following two parts:

1) aggregation of chunk-based acoustic features for story-based valence and arousal predictions and
2) acoustic-linguistic feature fusion from different levels of processing methods.

We evaluated the performance of the unimodal acoustic feature aggregation against the reference paper [5]. Our proposed solution by aggregating acoustic features at chunk-based processing enable us to concatenate acoustic and linguistic feature. This feature fusion, although already developed by others, differs from the previous research where audio and text are processed at the same level, i.e., per utterance, while we processed both at different levels (chunk vs. utterance/story).

III. Dataset and Features

A. Dataset

Although the baseline paper [5] provided an overview of the dataset, we briefly described it again to emphasize the problem we would like to solve.

Table I shows the number of instances/stories and chunks in all partition. The labels are given per each story. The baseline paper used majority vote method over predictions by acoustic features to aggregate the results, i.e., converting from chunks to a story. While this method is widely used to aggregate different classification methods, we proposed to aggregate at the feature extraction level. The goal of aggregating acoustic features after chunk-based feature extraction is to have the same (stories) number with the linguistic feature for feature concatenation. In other words, we want both acoustic and linguistic features to have the same number of n-dimensional vector for each story.

The label on the dataset is given on both alphabetic and numeric symbols, i.e., low ('L' or '0'), medium ('M' or '1'), and high ('H' or '2'). We used alphabetic labels as given in the baseline paper. Note that the number of chunks is different for each story; for instance, there are 34 chunks in the first story and 46 chunks in the second story.

B. Acoustic and Linguistic Features

We used acoustic and linguistic features provided by the organizer of The INTERSPEECH 2020 Computational Paralinguistic Challenge (ComParE). The evaluated acoustic features are ComParE, openXBOW, DeepSpectrum, and auDeep. The evaluated linguistic features (LinguistIc Feature Extractor, LIFE) are GMax, BLAtt, BLAtt + POS, and fused of those linguistic features. These features were extracted by using deep-learning based linguistic feature extractor. The resulted linguistic feature is 512-dimensional in size for each story. While we kept linguistic features as explained in the baseline paper [5], we added a set of acoustic features based on LibROSA toolkit [13] to list of acoustic features.

We extracted seven types of acoustic features from LibROSA features extractor: MFCCs (40 coefficients), chroma (12), mel-spectrogram (128), spectral contrast (7), tonal centroid (6), deltas of MFCCs (40), and deltas-deltas of MFCCs (40). This feature set is adopted from [7]. In total, there are 273 features on each frame. Paper [14] found that Mean+Std of GeMAPS [15] performed better than extended GeMAPS (eGeMAPS) and Bag-of-audio-words (BoAW) of GeMAPS. Hence, we extract Mean+Std from previous 273 LLDs resulting 546-dimensional functional features. Along with the aforementioned acoustic features, we evaluated both acoustic-only and acoustic-linguistic prediction of valence and arousal. We use simple aggregation methods, i.e., mean and maximum values of chunks' features, to determine the acoustic feature of story-based instances. Those features are concatenated with linguistic features in bimodal valence and arousal prediction.

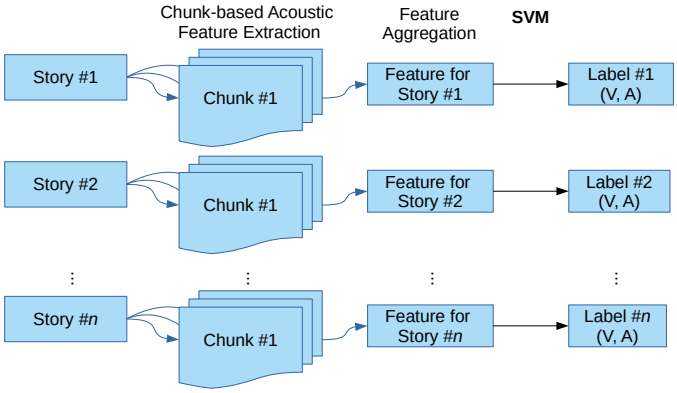

Fig. 1. Block diagram of acoustic features aggregation.

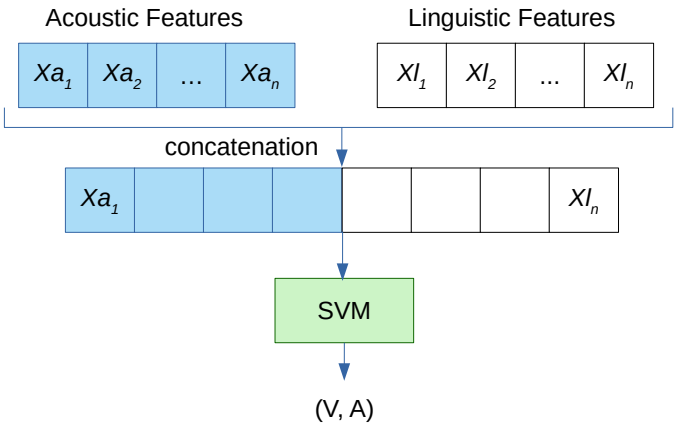

Fig. 2. Acoustic-linguistic feature concatenation with SVM.

IV. Methods

In this section, we describe our method to obtain the results. First, we introduce unimodal processing using acoustic features only with feature aggregation method. Second, we show how acoustic and linguistic features are fused and fed to the SVM classifier.

A. Unimodal Acoustic Features Aggregation

In this unimodal valence and arousal prediction, the input feature is acoustic. Acoustic features are extracted on frame-based processing as defined in [5]. For the LibROSA feature set, we applied the default settings, i.e., 2048 samples window size and 512 samples hop size. These features were standardized by removing mean and scaling to unit variance. The feature x is normalized according to

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ is the mean value of training samples and σ is standard deviation of training samples. The implementation of this normalization, along with SVM part, is performed using scikit-learn toolkit [16].

We evaluated two aggregation methods, mean and maximum values, over chunks to get feature for each story. Figure 1 shows our unimodal approach by aggregating acoustic features from chunks to stories. In mean aggregation, we averaged the values of each column vector of acoustic features for the same stories. In maximum aggregation, we chose the highest column vector value of acoustic features for each chunk on the same stories. By using those methods, each story has a same n-dimensional feature vector depend on extracted acoustic features, for instances, LibROSA with 546-dimensional (Mean+Std), ComParE with 6373-dimensional, and BoAW-125 with a 250-dimensional feature vector. Those feature vectors are fed into SVM to obtain valence and arousal categories.

B. Bimodal Acoustic-Linguistic Feature Fusion

Since the goal of the feature aggregation is to have the same dimension $(n \times 1)$ for both acoustic and linguistic features, it is easy to concatenate both features to improve valence and arousal prediction. Figure 2 shows our approach on acoustic-linguistic features fusion. Two feature sets are stacked horizontally to build the new feature vector for the input of the SVM classifier.

Given a set of acoustic-linguistic features pair (x_a and x_l) with valence and arousal category labels ('L', 'M', 'H'), the task of SVM is to classify whether a given feature set belongs to a category of valence and arousal. This classification task is performed using support vector classification (SVC) in scikit-learn toolkit [16] with a linear SVC kernel, 10^6 of

maximum iteration, and optimized complexities (C) values in the range $[10^{-6}, 10^{1}]$ with 10^{1} step size. For data balancing, imbalanced-learn toolkit was used [17]. The other parameters are left as default. The SVC classification is performed separately to predict valence and arousal categories for the same feature set.

V. Results and Discussion

We presented our results in three different analysis: unimodal acoustic features aggregation, bimodal acoustic-linguistic feature fusion, and development-test results mismatch.

A. Acoustic Features Aggregation Results

Since the modality of speech emotion recognition is the acoustic signal, the important feature is the acoustic features sets. We evaluated 13 acoustic feature sets derived from five feature extraction toolkits (listed in Section II). Table II shows the result of those feature sets with mean and maximum aggregation methods from chunks to stories in comparison with reference [5].

Although the results in Table II are from development partition, we can see that the UAR scores obtained by acoustic feature aggregation methods are higher than reference in overall feature sets. Hence, we believed a similar pattern would be observed in the test partition. Given limited trials for scoring test partition, we choose not to include this unimodal acoustic feature aggregation method for predicting final valence and prediction scores. Instead, we combined acoustic-linguistic feature concatenation, based on acoustic feature aggregation, to score the test partition for final valence and arousal prediction. This result suggests that the input feature aggregation method obtained higher performance (UAR) than the output aggregation method with majority voting.

In comparing mean vs. maximum aggregation methods, we found that mean aggregation lead to higher UAR score than maximum aggregation in development partition. This finding, however, is not incorporated for the combination of acoustic-linguistic feature concatenation. The choice of maximum acoustic feature aggregation for combination with linguistic features is intuited by the insight that in groups of audio chunks, i.e., in a single-story, the emotion within that story maybe represented by the highest value of the acoustic feature, for instance, the highest fundamental frequency in groups of chunks represent the actual the emotion behind it. Another reason for choosing maximum aggregation for acoustic-linguistic feature concatenation is development-test mismatch result explained in section 4.3.

El Ayadi et al. [18] listed some issues in feature extraction for speech emotion recognition systems. One of the issues is whether it is necessary to combine the acoustic feature with other types of features such as linguistic. Since linguistic features can be extracted from speech via speech-to-text technology, it is feasible to evaluate the contribution of linguistic features for future real-time implementation.

TABLE II
UAR results on development set: Unimodal acoustic features aggregation vs. reference [5]

Features	Baseline [5]		Mean Agg.		Max Agg.	
	V	A	V	A	V	A
LibROSA	-	-	**45.1**	38.3	42.7	39.7
ComParE	33.3	39.1	43.4	42.7	**45.3**	37.0
BoAW-125	38.9	42.0	44.6	**45.7**	44.6	40.1
BoAW-250	33.3	40.5	43.0	40.8	39.6	37.6
BoAW-500	38.9	41.0	42.6	41.0	42.9	37.9
BoAW-1000	38.7	30.5	43.5	41.5	40.2	39.8
BoAW-2000	**40.6**	39.7	41.9	44.8	43.4	40.1
ResNet50	31.6	35.0	36.5	36.7	37.1	39.0
AuDeep-30	35.4	36.2	38.4	42.1	42.8	35.6
AuDeep-45	36.7	34.9	39.5	40.5	39.3	33.3
AuDeep-60	35.1	**41.6**	43.4	42.1	40.7	41.4
AuDeep-75	32.7	40.4	41.9	44.4	40.9	**43.3**
AuDeep-fused	29.2	36.3	43.6	39.5	42.2	39.3

TABLE III
Result of bimodal valence and arousal prediction on development and test partition: official baselines vs. ours.

Features		Dev		Test	
Acoustic	Linguistic	V	A	V	A
ResNet50 [5]	-	31.6	35.0	40.3	**50.4**
-	BLAtt [5]	49.2	40.6	49.0	44.0
LibROSA	Gmax	**58.2**	34.6	40.5	34.8
ResNet50	Gmax	**58.2**	51.0	40.9	**50.4**
ResNet50	BLAtt	47.6	**52.5**	**56.3**	46.4
BoAW-250	BLAtt	**58.2**	44.4	49.0	47.4

Table III shows our results on using acoustic-linguistic feature concatenation for valence and arousal category prediction on development and test partitions. We improved the UAR score on development partition from 49.2 to 58.2 for valence and from 40.6 to 52.5 for arousal. On test partition, we improved the UAR scores from 49.8 to 56.3 for valence and from 49.0 to 50.4 for arousal. Although the gain was small, we showed that bimodal acoustic-linguistic feature concatenation improved the UAR scores of valence and arousal in most combinations of acoustic-linguistic feature pairs. Table III shows that evidence on both development and test partitions.

B. Development-Test Mismatch

We observed a mismatch between development and test results. In Table 2 of the reference paper [5], the best score for the feature set in the development partition is different from the best score in test partition. For instance, the best UAR score in development partition in majority voting (shown in Table II of this paper) is LIFE-fused for valence and BoAW-2000 for arousal. In test partition, LIFE-BLAtt obtained the highest UAR score for valence while the low-development-score ResNet50 obtained the highest UAR score for arousal. One of the possible reasons for this result is the data splitting portion for training/dev/test given in the baseline paper, i.e., 87/87/87. Since the portion of training/dev/test is also a parameter in machine learning, in this case SVM, we re-evaluated the development score with a splitting of 80%/20% for training/dev partition. We concatenated training and development partition provided by the ComParE Challenge organizer. We

used 80% of that concatenation data for training and the rest 20% for development score. However, the results still showed a mismatch between development and test scores although the gap is smaller than in reference development score. Another possible reason for that mismatch is the small size of data that cannot be solved except by adding more data.

Given this mismatch, we did not use the highest score of the development partition for testing our method and features. Instead, we relied on the previous test results for testing our acoustic-linguistic feature combination. The same reason also applies to choose maximum aggregation over mean aggregation.

VI. Conclusions

We presented an evaluation of acoustic features aggregation for acoustic-linguistic feature concatenation. Two aggregation methods are evaluated, i.e., mean and maximum aggregation. Our results suggest that using input feature aggregation in unimodal acoustic valence and arousal prediction is better than using output-based aggregation using majority voting. This result may be intuitively explained that the processing inputs is more informative than processing outputs. Apart from that benefit, using acoustic feature aggregation enables us to combine acoustic and linguistic features on a story or instance-based processing. The results by combining acoustic-linguistic features are higher than the results by using unimodal acoustic features, both from the official baseline and our acoustic features aggregation.

We observed a mismatch between development and test scores. This problem is a merit study for future research. Current method shows difficulties to determine which method performs better (from the development scores). There is a need for an approach to guarantee that the best method in development partition is also the best method in the test partition.

References

[1] J. A. Russell, "Affective space is bipolar," *J. Pers. Soc. Psychol.*, 1979.

[2] E. A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal," pp. 241–251, 2004.

[3] L. Cahill and J. L. McGaugh, "Mechanisms of emotional arousal and lasting declarative memory," *Trends Neurosci.*, vol. 21, no. 7, pp. 294–299, 1998.

[4] M. Costanzi, B. Cianfanelli, D. Saraulli, S. Lasaponara, F. Doricchi, V. Cestari, and C. Rossi-Arnaud, "The Effect of Emotional Valence and Arousal on Visuo-Spatial Working Memory: Incidental Emotional Learning and Memory for Object-Location," *Front. Psychol.*, vol. 10, no. November, pp. 1–13, nov 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2019.02587/full

[5] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. Macintyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *INTERSPEECH*, 2020.

[6] N.-h. Ho, H.-j. Yang, S.-h. Kim, and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," *IEEE Access*, vol. 8, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9050806/

[7] B. T. Atmaja and M. Akagi, "On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers," 2020. [Online]. Available: http://arxiv.org/abs/2004.00200

[8] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, 2018, pp. 3092—-3096.

[9] S. Gievska, K. Koroveshovski, and N. Tagasovska, "Bimodal feature-based fusion for real-time emotion recognition in a mobile context," *2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015*, pp. 401—-407, 2015.

[10] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," in *2019 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Lanzhou, 2019, pp. 519—-523.

[11] D. Griol, J. M. Molina, and Z. Callejas, "Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances," *Neurocomputing*, vol. 326-327, pp. 132–140, jan 2019.

[12] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *2016 IEEE Spok. Lang. Technol. Work.* IEEE, dec 2016, pp. 565–572.

[13] B. McFee and Others, "librosa/librosa: 0.7.1," oct 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3478579

[14] M. Schmitt and B. Schuller, "Deep Recurrent Neural Networks for Emotion Recognition in Speech," in *DAGA*, 2018, pp. 1537–1540.

[15] F. Eyben, K. Scherer, J. Sundberg, E. And, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. 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, 2016.

[16] 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," *J. Mach. Learn. Res.*, vol. 12, pp. 2825—-2830, 2011.

[17] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, sep 2016.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.