Hierarchical Prosody Analysis Improves Categorical and Dimensional Emotion Recognition

Xingfeng Li[1], Taiyang Guo[2], Xinhui Hu[1],Xinkang Xu[1], Jianwu Dang[2,3], $MasatoAkagi$[2]
[1]Hithink RoyalFlush AI Research Institute, Zhejiang, China
[2]Japan Advanced Institute of Science and Technology, Nomi, Japan
[3]Tianjin University, Tianjin, China

***Abstract*—Extracting reliable speech features is one of the most fundamental difficulties in emotion recognition systems. The extraction of spectral features has drawn much research attention but the extraction of prosody features, studying emotional cues, was often done by calculating statistics at an utterance level. However, the detailed prosody of different linguistic units can contain a large amount of emotion-related information. In this paper, we propose a novel hierarchical prosody analysis strategy by wavelet decomposition that models multi-level emotion transition phenomena. Our approach was evaluated on the IEMOCAP corpus and performed the best compared with state-of-the-art alternatives for both categorical and dimensional emotion recognition tasks, enabling the advancement of capturing dynamics in emotion expressions.**

I. Introduction

Speech emotion recognition (SER) has attracted increasing attention in the human-machine interaction research field to identify a speaker's emotional state. It can improve many applications such as speaking assistance, well-being detection, and many others [1, 2]. Typically, researchers have treated SER problems as multiple classification tasks of emotional categories such as happiness, sadness, anger, and neutral [3, 4]. However, it is often difficult to use a single categorization scheme to describe the rich emotions that emerge in a human's conversation since the intensities of them may change over time [5, 6]. Recently, the most popular solution to this limitation is to extend the study of SER to regression tasks of valence and arousal dimensions [7–10].

This paper contributes to a non-trivial problem in SER, namely devising reliable acoustic features to guarantee a higher category classification performance, and specifically, to improve the dimension estimation accuracy of valence and arousal to extract gradual emotional intensities. Previous studies mostly resolved this problem by studying spectral features [11, 12]. Among these works, Mel-frequency cepstral coefficients (MFCCs) were commonly used and have been improved to perform rather robust SER [13, 14]. Other spectral features, such as Mel cepstral coefficients [15] and modulation spectral features [16], have also been proposed. It is widely thought that spectral features are useful in recognizing emotional states [17]. Although prosody is known to significantly contribute to the supra-segmental characteristics of emotional phenomena [18, 19], prior SER work has sparsely provided robust features from this domain.

Prosody is affected hierarchically by short- and long-term dependencies ranging from phonemes to utterance levels (Here, we define such prosody embodied indifferent terms as hierarchical prosody; hereafter *HiPros*) [20–23]. The importance of HiPros in speech was evident in many areas, including speech synthesis [24], emotional voice conversation [22], and intonation analysis and generation [20]. Unfortunately, limited attempts were working with HiPros analysis for SER tasks. Most prior studies investigating speech prosody in terms of statistics at an utterance-level failed to discern the HiPros phenomena [19, 21]. Such limitations were likely to hamper SER performance [17, 25].

These findings inspired numerous works to approach SER via hierarchically motivated prosodic features. To obtain these features, conducting a wavelet analysis is suggested to effectively decompose and model the different prosodic phenomena at every linguistic level of the speech [20–22]. Recently, numerous means to obtain HiPros features for SER were developed in [26–28], however, those features were obtained by discrete wavelet transform (DWT) decomposition solely on raw and glottal waveforms.

In this study, we demonstrate how our advanced SER system is superior: 1) we provide an understanding on the extraction of reliable speech features from the perspective of HiPros; 2) in contrast to prior work that used the wavelet coefficients of raw and glottal waveform to define HiPros, we first apply the wavelet coefficients of F0 and glottal flow (GF) parameterization to define essential information in prosody; and 3) we show how HiPros are well suited for gradual SER.

The rest of the paper is organized as follows. Section 2 introduces our proposed gradual SER architecture. Section 3 describes the HiPros analysis algorithm in the prosody domains. The results on gradual SER are presented and discussed in Section 4. Section 5 concludes this paper.

II. The Proposed Architecture

Fig. 1 shows a block diagram of the present study's SER system. The right block highlights this study's scope to introduce a novel algorithm on hierarchical analysis of speech prosody, with the aim to construct an improved emotion recognizer for identifying both the categories and

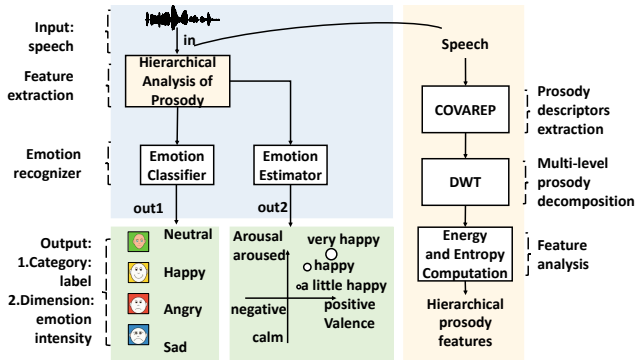

Fig. 1. Block diagram of emotion recognition system based on the proposed hierarchical prosodic features.

their location in valence-arousal space to define emotion intensity.

After receiving emotional speech as input, features are extracted. To this end, we initially used the COVAREP Toolkit [18] to extract prosody descriptors from the speech. DWT was next applied to decompose the prosody descriptors into multi-resolution levels. The energy and entropy of wavelet coefficients were then calculated to provide the robust features. The emotion recognizer, which incorporates support vector machines (SVMs), finally takes HiPros as input and maps them into an emotion categories and dimensions.

III. Hierarchical Prosody Analysis

A. Prosody descriptors extraction

The most expressive prosody, which is widely considered for dealing with emotion information extraction, is the F0 counter [20, 21]. However, research on speech production analysis has recently shown the importance of GF parameters to para-linguistic information characteristics [18, 25]. Therefore, we introduced HiPros based on the DWT decomposition of the F0 and GF parameterization counters in addition to the raw and glottal waveform. We extracted 12 commonly used prosody descriptors for SER, namely F0, normalized amplitude quotient, maximal dispersion quotient, quasi-open quotient, the difference in amplitude of the first two harmonics of the differentiated, parabolic spectral parameter, shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd), confidence value of Rd, spectral tilt/slope of wavelet responses, GF, GF derivative, and the speech waveform from the COVAREP Toolkit (v1.4.2) [18]. A detailed description of these features can be found in [29].

B. Multi-level prosody decomposition

The wavelet analysis is a method that relies on the introduction of an appropriate basis and characterization of a signal based on the distribution of amplitude within this basis [30]. This method of arranging successive wavelets in a hierarchical scheme is evident in modeling multi-level prosody phenomena [20–22]. In our approach, we perform a six-level multi-resolution decomposition based on DWT, giving parameters of translations by discrete values. These discrete values, known as wavelet coefficients, measure how much of the wavelet is at that resolution level, and the position is included in the speech prosody. Order ten Daubechies wavelets are used here as mother wavelets. Among several alternatives, the Daubechies functions were frequently used in SER and provide better results [31, 32].

The F0, referred to as one of the twelve prosody descriptors extracted by the COVAREP, was taken as an example. Fig. 2(a) shows the F0 counters of four emotional speech utterances in the neutral, happiness, anger, and sadness states uttered by a single speaker. Fig. 2(b) to (e) show the wavelet coefficients at different levels after a six-level multi-resolution analysis. It is clear from these figures that the wavelet coefficient distribution over the joint time-level plane is very distinct for all emotions, suggesting they could be well discriminated from each other. Fig. 2(f) shows four normalized mean energy (NME, c.f. Section 3.3.1) distributions corresponding to six wavelet resolution levels, further verifying this finding. The NME for each decomposition level can be calculated as the corresponding mean energy normalized by the max-min normalization over four emotional utterances. The figure indicates that the NME distribution of four emotions has different trends and varied with level-number, which confirmed that HiPros can improve emotion separability. We believe we can associate the decomposed values with emotions by energy and entropy analysis.

C. Feature analysis

Wavelet energy features In the following section, the wavelet coefficient formed by the DWT decomposition of one of the prosody descriptors is assumed to be given by $C_\xi(\kappa)$. The energy at each resolution level ξ=$-1,-2,...,-6$, will be the mean energy of the detailed signal

$$E_\xi = log_{10}\big(\Sigma|\ C_\xi(\kappa)\ |^2/N_\xi\big) \tag{1}$$

where N_ξ represents the number of wavelet coefficients at resolution level ξ. Thus, the total mean energy will be

$$E_\tau = \Sigma E_\xi \tag{2}$$

Then, the relative wavelet energy can be obtained by

$$p_\xi = E_\xi / E_\tau \tag{3}$$

Due to the fact that $\Sigma p_\xi = 1$, the distribution of p_ξ can be considered as a time-scale density. This enabled us to detect and characterize multi-level prosody phenomena in the time and frequency planes [22, 33].

Wavelet entropy features The Shannon entropy introduces a useful criterion for analyzing and comparing probability distribution, which provides a measure of the information of any distribution. We first defined the normalized total wavelet entropy, which provides a measure

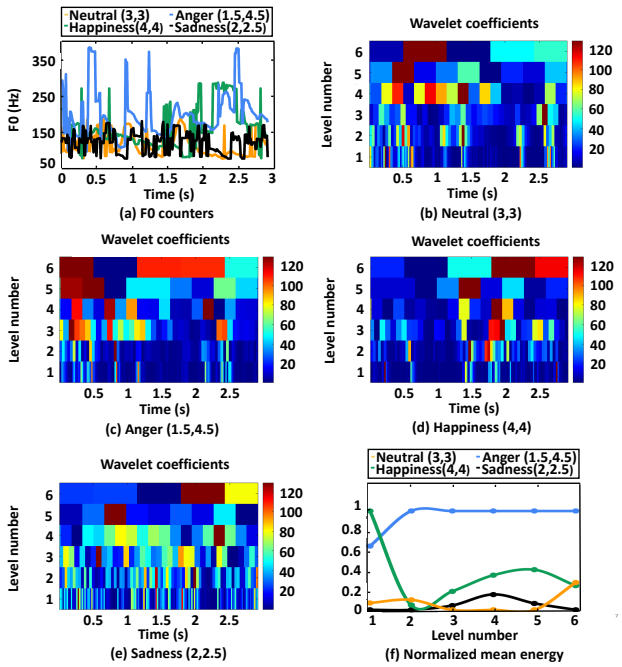

Fig. 2. F0 at a six-level wavelet decomposition by DWT for four emotions (valence, arousal) from a single speaker.

of order/disorder of an emotional speech derived from the relative wavelet energy (c.f. Eq. 3) and is given by

$$WE_{(\tau,nor)} = -\sum p_\xi * \log p_\xi \tag{4}$$

In particular, to capture the dynamical changes within the emotional state of speech, this study additionally defined three time-varying (TVR) entropy-related features. The TVR wavelet entropy is defined as

$$S_{(tvr,\xi)} = -\Sigma |\ C_\xi(\kappa)\ |^2 * \log |\ C_\xi(\kappa)\ |^2 \tag{5}$$

and the total WE and relative WE are given by

$$S_{(tvr,\tau)} = \Sigma S_{(tvr,\xi)} \tag{6}$$

$$p_{(tvr,\xi)} = S_{(tvr,\xi)} / S_{(tvr,\tau)} \tag{7}$$

Finally, we extract a total of 324 acoustic features on the basis of the wavelet analysis of 12 prosody descriptors of the speech signal. Each descriptor consists of 13 wavelet energy- and 14 entropy-related features at a six-level wavelet decomposition.

IV. Experimental Result

A. Experimental setup

Dataset We evaluated our method on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus [34]. This corpus contains five sessions, where each session has utterances from one male and one female speaker. Overall, there are 10 unique speakers. We selected four emotions of angry, happy, neutral and sad. To be consistent with previous work, we merged excitement with happiness and considered it as happy. Moreover, to study gradual emotion intensity, we included valence (1-negative, 5-positive) and arousal (1-calm, 5-excited) dimensions, which scaled from 1 to 5.

Baselines We conducted experiments on seven separate baselines to show the usefulness of the proposed HiPros from four domains. First, we included a typical prosody baseline by using the OpenSMILE prosodyShsViterbiLoudness configuration. Second, a spectral benchmark was given by 12 MFCCs, along with the delta, and delta-delta extracted from 25-ms frames with a 5-ms shift, in terms of statistics of mean, standard deviation, skewness, and kurtosis. Third, four commonly used OpenSMILE sets of IS11_speaker_state, IS12_speaker_trait, IS13_ComParE, and ComParE_2016 were referred in combined domains, which explored F0, energy, spectral, MFCC, duration, voice quality (the zero-crossing rate, jitter, shimmer, and harmonicsto-noise ratio), spectral harmonicity, and psychoacoustic spectral sharpness. Finally, we followed prior wavelet analysis schemes [26–28] on raw and glottal waveforms and presented a baseline in the wavelet domain.

Emotion recognizer and evaluation metrics The WEKA's SVM of C-SVC and nu-SVR algorithms with linear kernels were used to perform category classification and dimension estimation, respectively [35]. This was because the SVM is simple and confirmed to be well-validated after being tested on both classification and regression tasks. It is also one of the most commonly used choices in SER [16]. The features from training and testing data were scaled to [0,1] by max-min normalization before applying the SVM. All results obtained on the IEMOCAP corpus were presented by leave-one-speaker-out (LOSO) validation. The weighted and unweighted accuracies (WA and UA, respectively) were determined to assess the category classification performance. Moreover, the Pearson correlation coefficients (CC) and mean absolute error (MAE) were included to evaluate the estimation accuracy of emotion dimensions.

B. Experimental results and discussions

Table I details the results obtained by the proposed HiPros and seven baselines, and includes the recognition rate for categories; and the CC and MAE for valence and arousal dimensions. The RIR indicates the relative percentage improvement of the accuracy rate and the RRR refers to the relative percentage reduction of the error rate obtained by our study to the baseline work ($base$) with a ground truth equals to 1/0 and the metric sets to CC/MAE, which was calculated as:

$$\Re = \frac{proposed_{metric} - bases_{metric}}{groundTruth - bases_{metric}} * 100\% \tag{8}$$

where $\Re \in \{RRR, RIR\}$ and $metric \in \{CC, MAE\}$.

Our proposed HiPros achieved the best accuracy in overall classification, reaching up to 56.81% WA and 57.57% UA. It achieved an absolute increase of 8.6% WA and 9.85% UA

TABLE I
RECOGNITION RESULTS ON CATEGORICAL CLASSIFICATION AND DIMENSIONAL ESTIMATION FOR THE PROPOSED AND BASELINE FEATURES; BOLDFACE INDICATES THE BEST PERFORMANCE IN EACH CASE. * INDICATES THAT THE ESTIMATION RESULTS DIFFER SIGNIFICANTLY BETWEEN THE HIPROS AND BASELINES ($p < 0.05$)

Domain	Feature	**Categorical Emotion**		**Dimensional Emotion**							
		classification		valence estimation				arousal estimation			
		WA	UA	CC	RIR(%)	MAE	RRR(%)	CC	RIR(%)	MAE	RRR(%)
Prosody	prosodyShsViterbiLoudness	48.21	47.72	0.37	7.94	**0.16**	–	0.65	11.42	0.12	8.33
Spectral	MFCC	30.05	25.47	0.06	38.30	0.18	11.11	0.02	68.37	0.16	31.25
Combined	IS11_speaker_state	56.48	57.53	0.36	9.38	0.20	20.00	0.47	41.51	0.16	31.25
	IS12_speaker_trait	55.80	56.72	0.32	14.71	0.23	30.43	0.42	46.55	0.18	38.89
	IS13_ComParE	56.26	57.13	0.25	22.67	0.27	40.74	0.38	50.00	0.21	47.62
	ComParE_2016	55.75	56.63	0.29	18.31	0.26	38.46	0.35	52.31	0.22	50.00
Wavelet	Ref. [26–28]	50.52	50.87	0.34	12.12	**0.16**	–	0.64	13.89	0.12	8.33
	Proposed	**56.81**	**57.57**	**0.42**	–	**0.16***	–	**0.69**	–	**0.11***	–

higher than the prosody baseline of prosodyShsViterbiLoudness, suggesting that studying multi-level prosody phenomena by the wavelets indeed contributes to SER tasks. The HiPros also appeared to be superior compared with the previous wavelet-based features [26, 28], and improved the baseline accuracy from 50.52% to 56.81% WA and 50.87% to 57.57 % UA. This improvement further establishes the fact that HiPros are needed to explore in more robust prosody domains rather than simple raw and glottal waveforms. Also, by comparing the results by the HiPros with those by MFCCs, the SER performance gaps that appeared between the prosody and spectral features in prior work were narrowed by this study. Moreover, the HiPros offered a better performance even relative to the four widespread-use OpenSMILE sets in combined domains.

Most importantly, the advantage of this present study was most notable in estimating the valence and arousal dimensions, which can significantly improve the identification of gradual emotion transitions. As shown in Table I, the HiPros achieved the best accuracy in both estimation of arousal and valence, reaching CCs of up to 0.68 and 0.43, and MAEs of low to 0.11 and 0.16, respectively. In comparison with the seven baselines, HiPros provided valence RIRs of [7.94%,38.30%], arousal RIRs of [11.42%,68.37%], valence RRRs of [11.11%,40.74%], and arousal RRRs of [8.33%,50.00%]. For a fair comparison, we compared this approach with state-of-the-art deep learning-based SER systems to identify gradual emotion intensity using other works for general bench-marking. In [36], valence and arousal CC values of 0.36 and 0.49, respectively, were reported by using the MLP with GeMAPS features. Abdelwahab et al. [37] used the IS13_ComParE to obtaine valence and arousal CC values of 0.25 and 0.50, respectively. However, no information was provided regarding MAE values. The performance of the proposed SER system appears to be consistently superior to other alternatives; this clearly reflects the efficiency of the HiPros being more competent than existing features in performing emotion intensity tracking tasks, even by using a simple SVM algorithm.

Moreover, arousal commonly received higher correlations than valence for all attempts. This result was consistent with the evaluations conducted by humans because the valence dimension was labelled to be associated with not only the audio modal but also other modals such as visual expression and linguistics [8, 25]. With possible refinements in our future work, the performance of the HiPros can be further improved by combining other modals for SER.

V. CONCLUSION

We introduced advancements in devising robust prosodic features for gradual SER. The advancements were mainly based on DWT applications to decompose the multi-level prosodic phenomena present in an emotional speech. The results of our experiments demonstrated that the proposed features could perform better in classifying emotions and estimating emotion dimensions under LOSO conditions on the IEMOCAP corpus compared with state-of-the-art features. These findings suggest that the proposed method enables tracking dynamics of emotional states along the time plane. Further work may include automatic emotion estimation integration into man-machine interactions, such as affective speech-to-speech translation systems.

REFERENCES

[1] Yang Gao, Zhengyu Pan, Honghao Wang, and Guanling Chen, "Alexa, my love: Analyzing reviews of amazon echo," in *2018 IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, pp. 372–380.

[2] Felix Albu, Daniela Hagiescu, Liviu Vladutu, and Mihaela-Alexandra Puica, "Neural network approaches for children's emotion recognition in intelligent learning applications," in *EDULEARN15 7th Annu Int Conf Educ New Learn Technol Barcelona, Spain, 6th-8th*, 2015.

[3] Michael Neumann and Ngoc Thang Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019*. IEEE, 2019, pp. 7390–7394.

[4] Turgut Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320–326, 2019.

[5] Bagus Tris Atmaja and Masato Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP 2020*. IEEE, 2020, pp. 4482–4486.

[6] Klaus R. Scherer, "What are emotions? and how can they be measured?," *Soc. Sci. Info.*, vol. 44, no. 4, pp. 695–729, 2005.

[7] Roddy Cowie and Randolph R Cornelius, "Describing the emotional states that are expressed in speech," *Speech commun.*, vol. 40, no. 1-2, pp. 5–32, 2003.

[8] Michael Grimm, Emily Mower, Kristian Kroschel, and Shrikanth Narayanan, "Combining categorical and primitives-based emotion recognition," in *2006 14th EUSIPCO*. IEEE, 2006, pp. 1–5.

[9] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee, "Speech representation learning for emotion recognition using end-to-end asr with factorized adaptation.," in *INTERSPEECH*, 2020, pp. 536–540.

[10] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.
[11] Hao Hu, Ming-Xing Xu, and Wei Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *2007 ICASSP*. IEEE, 2007, vol. 4, pp. IV–413.
[12] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019*. IEEE, 2019, pp. 7405–7409.
[13] A Pramod Reddy and V Vijayarajan, "Audio compression with multi-algorithm fusion and its impact in speech emotion recognition," *Intl. J. of Speech Tech.*, pp. 1–9, 2020.
[14] Yan Wang and Weiping Hu, "Speech emotion recognition based on improved mfcc," in *Proc. of the 2nd CSAE*, 2018, pp. 1–7.
[15] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, vol. 1, pp. 137–140.
[16] Siqing Wu, Tiago H Falk, and Wai-Yip Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.
[17] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[18] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ICASSP*. IEEE, 2014, pp. 960–964.
[19] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 787–800, 2007.
[20] Hans Kruschke and Andreas Koch, "Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization," in *2003 ICASSP.* IEEE, 2003, vol. 1, pp. I–I.
[21] Gerard Sanchez, Hanna Silen, Jani Nurminen, and Moncef Gabbouj, "Hierarchical modeling of f0 contours for voice conversion," in *15th INTERSPEECH*, 2014.
[22] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki, "Emotional voice conversion with adaptive scales f0 based on wavelet transform using limited amount of emotional data.," in *INTERSPEECH*, 2017, pp. 3399–3403.
[23] Yunfeng Xu, Hua Xu, and Jiyun Zou, "Hgfm: A hierarchical grained and feature model for acoustic emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6499–6503.
[24] Martti Vainio, Antti Suni, Daniel Aalto, et al., "Continuous wavelet transform for analysis of speech prosody," *TRASP 2013, An Interspeech 2013 satellite event, August 30, 2013*.
[25] Yongwei Li, Junfeng Li, and Masato Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," *The J. of the Acoust. Soci. of America*, vol. 144, no. 2, pp. 908–916, 2018.
[26] Kunxia Wang, Guoxin Su, Li Liu, and Shu Wang, "Wavelet packet analysis for speaker-independent emotion recognition," *Neurocomputing*, 2020.
[27] Pankaj Shegokar and Pradip Sircar, "Continuous wavelet transform based speech emotion recognition," in *ICSPCS*. IEEE, 2016, pp. 1–8.
[28] Hariharan Muthusamy, Kemal Polat, and Sazali Yaacob, "Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals," *Mathematical Problems in Engineering*, vol. 2015, 2015.
[29] Maarten Brilman and Stefan Scherer, "A multimodal predictive model of successful debaters or how i learned to sway votes," in *Proc. of the 23rd ACM intl. conf. on Multimedia*, 2015, pp. 149–158.
[30] Stéphane Mallat, *A wavelet tour of signal processing*, Elsevier, 1999.
[31] Shashidhar G Koolagudi and K Sreenivasa Rao, "Emotion recognition from speech: a review," *Intl. J. of speech tech.*, vol. 15, no. 2, pp. 99–117, 2012.
[32] K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada, "Emotion recognition from speech using global and local prosodic features," *J. of speech tech.*, vol. 16, no. 2, pp. 143–160, 2013.
[33] Atsuko Schütt, Iori Ito, Osvaldo A Rosso, and Alejandra Figliola, "Wavelet analysis can sensitively describe dynamics of ethanol evoked local field potentials of the slug (limax marginatus) brain," *Journal of neuroscience methods*, vol. 129, no. 2, pp. 135–150, 2003.
[34] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.
[35] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, "The weka data mining software: an update," *ACM SIGKDD*, vol. 11, no. 1, pp. 10–18, 2009.
[36] Bagus Tris Atmaja and Masato Akagi, "Deep multilayer perceptrons for dimensional speech emotion recognition," *arXiv preprint arXiv:2004.02355*, 2020.
[37] Mohammed Abdelwahab and Carlos Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 26, no. 12, pp. 2423–2435, 2018.