INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The Contribution of Acoustic Features Analysis to Model Emotion Perceptual Process for Language Diversity

Xingfeng Li[1], *Masato Akagi*[2]

Japan Advanced Institute of Science and Technology
lixingfeng@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

The multi-layered perceptual process of emotion in human speech plays an essential role in the field of affective computing for underlying a speaker's state. However, a comprehensive process analysis of emotion perception is still challenging due to the lack of powerful acoustic features allowing accurate inference of emotion across speaker and language diversities. Most previous research works study acoustic features mostly using Fourier transform, short time Fourier transform or linear predictive coding. Even though these features may be useful for stationary signal within short frames, they may not capture the localized event adequately as speech transmits emotion information dynamically over time. This case introduces a set of acoustic features via wavelet transform analysis of the speech signal, and specifically, models the perceptual process of emotion for language diversity. For this aim, the proposed features are analyzed in a three-layer emotion perception model across multiple languages. Experiments show that the proposed acoustic features significantly enhance the perceptual process of emotion and render a better result in multilingual emotion recognition when compared it to the widely used prosodic and spectral features, as well as their combination in literature.

Index Terms: wavelet transform, speech emotion recognition, emotion dimension, three-layer model

1. Introduction

Speech emotional state plays a pivotal role in a daily conversation for delivering the ideas, thoughts, and moods. Researchers have recently agreed that people share primary emotions like anger, happiness, sadness, fear, surprise, and disgust independent of languages [1, 2]. For instance, emotion recognition in multilingualism communications has been shown to be cross-lingual, where multiple spoken languages may be used in one conversation or even in one sentence. Multilingual speech emotion recognition (mSER), considering human diversity for language in realistic conditions, is thus a growing area of focus within affective computing to enhance a natural human-machine interaction [3, 4]. Still, there exist two challenges to facilitate machines understanding emotion from multilingual speech, i.e., 1) design a computational emotion model generalizes across languages; and 2) extract powerful acoustic features with the ability to distinct emotional states.

This paper studies each of these two challenges to model the perceptual process of emotion for language diversity. Many computational models thoroughly used in machine learning and pattern recognition, have been constructed previously for acoustic SER, such as the Gaussian mixture model, support vector machine/regression, and hidden Markov model and so on [5, 6, 7]. All these models were found to be promising for training and testing on a single specific corpus. However, such models were limited to generalize in mSER tasks owing to the specific optimal patterns like the type of a kernel, changing significantly concerning the diversity for speakers and languages [8]. To solve this limitation, many researchers have reconsidered the perceptual process of emotion in human speech as a multi-layer scheme, and introduced other models like the extreme learning machine, deep neural network, and long short-term memory [9, 10, 11]. Despite the advances made in mimicking the perceptual process of emotion by multiple layers, this success usually requires a massive set of training data led to another challenge in data scarcity for mSER. Still, most deep learning models were performed by black-box testing using hidden layers, and many scientists argued these models might fail to clarify a comprehensive analysis of the perceptual process of emotion.

Emotion psychology studies have shown that, as an alternative to model the multi-layered perceptual process of emotion, Brunswik's lens principles of representative design seem to be too appealing a topic [12]. Scherer originally applied a Brunswik's lens model in three layers to infer the personality from voice [13]. It was assumed that the emotional state of a speaker is externalized by distinct distal cues, i.e., acoustic features; and proximally perceived as percepts, is the mechanism of a perceptual process of emotion. This model benefits greatly from a decomposition of the inference process, allows for assessing the particular cues of failures to improve the previous achievement and provide a more clear perceptual process. In [14], Huang and Akagi proposed to study the proximal representation of distal cues for describing a speaker's state in an expressive speech by adjectives. In [15], Elbarougy et al. replicated the earlier results after Huang by using the valence (pleasant and unpleasant) and arousal (relaxed and aroused) space to identify the speech emotional state dynamically. All the above mentioned models provide us the knowledge to understand the perceptual process of emotion for each single language. Still, the challenge remains to find the inference rules generalize across languages. Nonetheless, Li et al. reported four universal proximal percepts (semantic primitives), in a renewed three-layer perceptual process of emotion across languages by a feature selection approach, considering the combined effects of a judgment in human fuzzy-vague knowledge [16]. Even though the obtained results in that study may be highly representative on the understanding of semantic information, it still restricts to gain the nature of perceptual process of emotion since lacking robust acoustic features to distinct the emotional state. However, the latter is fairly vital of the inference of speaker's state from speech.

This study contributes to an important challenge in mSER, i.e., modeling the perceptual process of emotion for language diversity in three layers by acoustic features, semantic primitives, and emotion dimensions; and specifically, studying a set of acoustic features with the goal toward generalizing

http://dx.doi.org/10.21437/Interspeech.2019-2229

across languages. Most studies traditionally examined the Mel frequency cepstral coefficients (MFCCs) as one of the popularly used acoustic features for SER [6, 17]. Despite the progress made by the MFCCs, they are still restricted to externalize emotional states in speech due to the following two reasons. Firstly, MFCCs were computed by the short time Fourier transforms (STFT), failing to identify sudden burst in a slowly varying signal because of the windowed STFT holds uniform resolution over the time-frequency plane [18, 19]. Secondly, MFCCs inherently accept that a speech frame carries information for only one phoneme per time; however, there may exist language-dependent adjacent phonemes of both voice and unvoiced phonemes in nature which affect the low and high-frequency spectrum separately for SER [20]. Prosodic speech features, on the other hand, such as the fundamental frequency, energy, and timing, etc. have also been thoroughly studied for SER [13, 14]. Although prosodic features appeared to be profitable to predict aroused emotions; still, they failed to identify emotional states with similar arousals, like joy and anger which have similar properties in the prosodic domain with high fundamental frequency, high energy and so on [21]. In addition, prosodic features are typically obtained from each frame and then calculated via statistics of all features in one utterance, losing the temporal information carried in an emotional speech [21, 22].

The question of extraction of robust acoustic features generalize across languages is still a major concern of SER. Nevertheless, researchers largely accepted that the emotional state in speech has an impact on the speech production mechanism across the glottal source and vocal tract [23, 24]. In [25], Li et al. reported that glottal source information advanced the perception of emotions; besides, vocal tract cues affects contributed to the dimensional understanding across valence and arousal. Mokhtari et al. suggested that glottal amplitude quotient played a vital role in conveying paralinguistic information [26]. However, the glottal source and vocal tract are significantly affected by many factors, such as the gender, culture as well as the speaking style of a speaker and so on [27, 28]. This fact limits the relative contribution of acoustic features derived from the glottal source and vocal tract to model the perceptual process of emotion across languages.

These days, phonologists and phoneticians have popularly assumed that the expression and perception of emotion in speech is hierarchic and multi-functional. There exists relevant information at both short and long-term dependencies from micro-prosody information on the phonemes, to the prosody of words, phrases, and the whole sentence [22, 29]. Inspired by the distinct characteristics of emotional states across different prosodic levels and frequency structures of vocal fold and vocal tract, this study takes one step beyond current acoustic features extraction algorithms and proposes a method for robust feature extraction via the wavelet transform (WT) analysis of the speech and glottal signal. The WT benefits the analysis of acoustic features for mSER due to the following advantages: First, the WT decomposes a speech or glottal signal into subsequent sub-bands, allows for discerning the emotion-related oscillations in speech; and has the potential to form them separately according to each prosodic level that might approximately match the human hierarchic perception of emotion; Moreover, the WT is superior to the traditional signal processing methods such as the STFT and could process any non-stationary speech signals. It offers optimal time resolution for each frequency and can underlie dynamic variations associated with emotional state in speech.

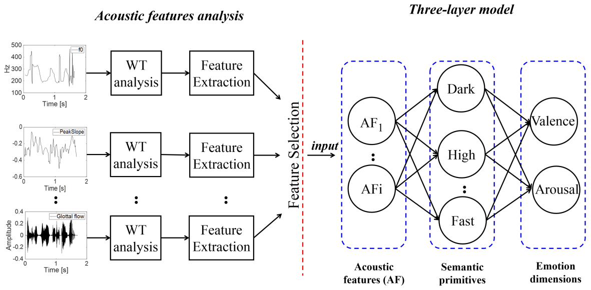

Figure 1: *Schematic diagram of the proposed multilingual emotion recognition system.*

This paper contains the following main parts: 1) we determined a computational emotion model in three layers to study the perceptual process of emotion for language diversity; 2) we proposed a set of robust acoustic features derived from the glottal source and vocal tract by the WT analysis; 3) we evaluate our proposed features by comparing them to the most widely used prosodic features, spectral features, as well as their combination. 4) we lastly strengthen how well the WT-based acoustic features are suited for characterizing the language independent cues in mSER.

2. Speech datasets

We chose three corpora of acted emotions across languages of Japanese, German, and Chinese. In addition, four similar emotions of neutral, happy, angry, and sad were selected to train the system and compare performance among corpora.

The Fujitsu database contains speech with acted emotions in Japanese, produced by one professional actress. The speaker uttered a sentence using five emotions: neutral, happiness, sadness, cold anger and hot anger. These recordings comprise 20 different sentences, each of them repeated once in neutral and twice in each of the other emotions. A total of 140 utterances were selected from this database: 20 neutral, 40 happiness, 40 hot anger, and 40 sadness.

The German corpus is the well-known Berlin Emo-DB. Ten professional actors (five males and five females) each uttered ten sentences in German to perform seven emotions. The number of utterances of each emotion was: 127 anger, 81 boredom, 46 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness. Finally, 200 utterances were selected from this corpus with 50 utterances in each of the four similar emotions as in the Fujitsu database.

The Chinese CASIA corpus contains 9600 utterances of six emotions: neutral, anger, fear, surprise, happiness, and sadness, produced by two male and two female professional actors. Each actor performed six emotions individually and produced 400 utterances in each category. This study selected 200 utterances of spontaneous content from four actors consists of 50 neutral, 50 happiness, 50 sadness, and 50 anger.

3. Research method

Figure 1 depicts a schematic diagram of the whole method. Acoustic features analysis was first given by the WT analysis of the low-dimensional features (LDFs) in speech across languages. The robust features based on energy and entropy were secondly calculated using the WT coefficients and selected by the sequential floating forward selection (SFSS).

The three-layer model incorporating fuzzy inference systems (FIS) took the optimal features as input and mapped them into valence and arousal dimensions through semantic primitives. The steps for modeling the perceptual process of emotion for language diversity were detailed as follows.

3.1. WT analysis of speech features

This study firstly extracted sixteen LDFs of the glottal source and vocal tract along with the speech waveform from the COVAREP toolbox(v1.4.2) [30]: fundamental frequency, normalized amplitude quotient, quasi-open quotient, difference in amplitude of the first two harmonics of the differentiated glottal source spectrum, parabolic spectral parameter, maximal dispersion quotient, spectral tilt/slope of wavelet responses, shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd), the confidence value of Rd, glottal flow, glottal flow derivative, and the first-fifth formants. Each of these LDFs was secondly decomposed into six resolution levels by the discrete wavelet transform (DWT) associated with the order ten Daubechies wavelets due to its better performance in SER [31, 32]. Finally, the energy- and entropy-related features derived from wavelet coefficients were calculated.

3.1.1. Wavelet energy-related features

Let $C_l(n)$ be the detail coefficients formed by DWT of one of the LDFs, where l=1,2,...,m, n=0,1,2,...,2^m-1, m=6 is the number of decomposition levels, and N is the length of the detailed coefficients at each node (l,n). Then, the energy at the decomposition level l will be given by

$$E_l = log_{10}\left(\frac{\sum \mid C_l(n) \mid^2}{N}\right) \tag{1}$$

Consequently, Equation 2 and 3 define the total energy and relative wavelet energy respectively.

$$E_{tot} = \sum_{l=1}^{m} E_l \tag{2}$$

$$E_{rel}^{(l)} = \frac{E_l}{E_{tot}} \tag{3}$$

Due to the fact that $\sum E_{rel}^{(l)} = 1$, the distribution of $E_{rel}^{(l)}$ reflects a degree of density in a time-scale, which may contribute some insights into discerning and characterizing distinct information of emotion in time-frequency planes [33].

3.1.2. Wavelet entropy-related features

The first wavelet entropy-related feature, called normalized total wavelet entropy (NTWE) performs as a measure of order/disorder of an emotional speech derived from the relative wavelet energy (c.f. Eq. 3) and given by

$$NTWE = -\sum E_{rel}^{(l)} * \log E_{rel}^{(l)} \tag{4}$$

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 (WE) is given by

$$WE_l = -\sum \mid C_l(n) \mid^2 * \log \mid C_l(n) \mid^2 \tag{5}$$

Then the total WE and relative WE are defined as

$$WE_{tot} = \sum_{l=1}^{m} WE_l \tag{6}$$

$$WE_{rel}^{(l)} = \frac{WE_l}{WE_{tot}} \tag{7}$$

This paper extracted a total of 459 acoustic features based on the WT analysis of 17 LDFs of the speech signal. Each of the 17 LDFs given 27 features consisting of 13 wavelet energy and 14 entropy-related features at a six level wavelet decomposition.

3.2. Primitives-based emotion evaluation

This study defined a computational emotion model to estimate emotions in speech by a three-layer process, assuming that human perception of emotion did not originate directly from a change in acoustic cues, but from an indirect route of a more subtle perception of semantic primitives. For instance, low arousal and negative valence speech often convey dark feelings, in contrast, high arousal and positive valence speech convey bright moods. We initially examined 17 semantic primitives in the three-layer model to describe emotional speech after [14], namely bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow. To construct the three-layer model, the three emotional corpora were first evaluated in terms of each semantic primitive via human listening tests. Emotional speech was evaluated 17 times by subjects: once for each semantic primitive for all utterances in one corpus. Each of the 17 semantic primitives was scored on a five-point scale: 1 Does not feel at all, 2 Seldom feels, 3 Feels a little, 4 Feels, 5 Feels very much. In addition, since this study describes emotions by a dimensional space spanned by valence and arousal, the corpora needed to be further annotated in terms of emotional dimensions. The same subjects were asked to evaluate these dimensions on a five-point scale (-2, -1, 0, 1, 2) for valence (-2 being very negative and +2 being very positive) and arousal (-2 being very relaxed and +2 being aroused).

Eleven native Japanese speakers (nine males and two females) were asked to evaluate the Fujitsu database, and ten native Chinese speakers (five males and five females) were asked to evaluate the CASIA corpus. Still, it was impractical for us to recruit enough German native speakers for the listening test. Nonetheless, psychology research has recently proven that human could recognize emotions in speech cross languages [14], so we asked nine Japanese native speakers (eight males and one females) to evaluate the Berlin-Emo DB instead. The basic theory of the semantic primitives and emotion dimensions was explained to the participants before they listened to a small set of demos involving different degrees of a certain emotion. The training test tried to enable listeners to understand the adjectives or dimensions. All stimuli were played randomly via binaural headphones at a comfortable sound pressure level in a soundproof room.

The averaged results of inter-evaluator correlation for the semantic primitives in terms of Fujitsu, Berlin Emo-DB, and CASIA were almost identical with values ranging from 0.84–0.93, 0.84–0.93, and 0.82–0.92, respectively. Besides, the average correlation between evaluators over valence and arousal was 0.96 and 0.96 for Fujitsu, 0.92 and 0.94 for Berlin Emo-DB, and 0.85 and 0.91 for CASIA. The inter-rater agreement was generally lower for valence than for arousal, indicating human evaluations were more poorly correlated with respect to valence compared to that of arousal.

4. Experiment

This paper studied mSER performance on the estimation of valence and arousal dimensions in multilingual scenarios. Obtained results were assessed by the correlation coefficients (CC) and mean absolute error (MAE) between the averaged human evaluators and the estimations by systems.

4.1. Experiment setup

Experiments were performed in mixed-corpus setting via leave-one-speaker-out cross-validation (LOSO), where the model was trained on all but one speaker's data of mixed-corpus of three different languages and then tested on the held data. The held-out speaker was rotated until all speakers were tested.

The training of the three-layer model was performed based on the adaptive neuro-fuzzy inference systems (ANFISs) owing to the ANFISs benefit from a lower root mean square error on the model of nonlinear input and output relations by fusing human knowledge [34]. And specifically, the nature of perception of emotion in speech is vague and fuzzy [6]; and the three-layer model fused human knowledge from evaluations of semantic primitives and emotion dimensions, which also included nonlinear processing to human perception of emotion. To avoid exorbitant costs in terms of time for system training and define the optimal features, we additionally used the SFFS in all selection from original sets of 459 acoustic features and 17 semantic primitives. SFFS is an iterative algorithm to evaluate the selected subset and combined effects of features and KNN classifier during the evaluation process. This work selected 21 acoustic features (c.f. Table 1) and four semantic primitives of $dark$, $strong$, $weak$, and $heavy$ to model the perceptual process of emotion for the language diversity.

To assess and compare the gain from using the proposed WT-based acoustic features, we built three baselines: the first set of features was 19 prosodic speech features (PSFs), contains four fundamental frequency-related features, four power-envelope related features, five power spectrum-related features, three duration-related features, and three voice-quality related features. These 19 prosodic features have been widely studied in a three-layer model in the literature [15, 35]. The second set consists of 196 statistical modulation spectral features (MSFs) that were derived from the acoustic frequency and modulation frequency domains. Superior to the standard MFCCs that carry a signal's short-term spectral properties only. The MSFs benefit both temporal and spectral properties of a speech signal as used by humans via an analysis of temporal envelope of multiple acoustic frequency bins [36, 37]. This has been proven by literature that MSFs outperformed MFCCs in SER [37]. The reader is suggested to refer to [36] for a detailed description of the MSFs. Moreover, the third baseline combined the PSFs and MSFs. Each of the three baselines was the same in training the three-layer model using the best four semantic primitives, but the best acoustic features chosen from each individual feature set.

4.2. Experiment results and discussion

Table 2 shows the CC and MAE of the estimation using the proposed acoustic features and the three baselines. As seen, the proposed WT-based acoustic features achieve the best performance in all estimations of emotion with respect to the valence and arousal dimensions, providing a CC of 0.82 and 0.93, while the MAE was 0.47 and 0.30 respectively. This result outperformed all another three baselines of the PSFs, MSFs, and the combination of PSFs and MSFs in predicting valence dimension, yielding a relative error reduction rate (RErs) of 50%, 58%, and 20% for CC, and 26.5%, 35.6%, and 7.8% for MAE. These differences are statistically significance ($p < 0.001$). Besides, the proposed features consistently outperformed other baselines on the estimation of arousal dimension. The individual values of the RErs are 22.2% and 46% for CC, and 16.7% and 31.8% for MAE in comparison with that of the PSFs and MSFs.

Table 1: *Selected 21 acoustic features in three-layer model*

Low dimensional features	**Functionals**
Time signal	$E_{rel}^{1/5/6}$, WE_6, E_2, $NTWE$
glottal flow	E_3, $WE_{2/5}$, WE_{rel}^4
2nd formants	E_{tot}, E_{rel}^6
parabolic spectral parameter	E_3, $WE_{rel}^{3/4}$, WE_{tot}
5th formants	E_{tot}
glottal flow derivative	E_2, WE_{rel}^5
1st formants	WE_{rel}^2
fundamental frequency	WE_{rel}^4

Relative wavelet energy at every decomposition level l ($E_{rel}^{(l)}$); wavelet entropy (WE_l); wavelet energy (E_l); normalized total WE ($NTWE$); relative WE ($WE_{rel}^{(l)}$); total energy (E_{tot}); total WE (WE_{tot})

Table 2: *The CC and MAE obtained for valence and arousal using different acoustic features. ** indicate that the features outperform other alternatives; * indicate the features outperform the baselines of PSFs and MSFs, but not significant different with PSFs+MSFs. ($p < 0.001$)*

Features	**Valence**		**Arousal**	
	CC	**MAE**	**CC**	**MAE**
PSFs	0.64	0.64	0.91	0.36
MSFs	0.57	0.73	0.87	0.44
PSFs+MSFs	0.75	0.51	0.93	0.29
Proposed	**0.82**	**0.47****	**0.93**	**0.30***

As an aside, the PSFs appeared to be suitable for the estimation of the arousal dimension; still, it limited to that of valence. The main reason was attributed to the existence of similarities between properties of some emotions, such as joy and anger have similar properties for the fundamental frequency [16, 21]. Moreover, the MSFs produced a comparatively low performance when compared it to other alternative features, were probably not immune to the diversity with respect to language and speaker [29, 36]. Further, the result of the fused features of PSFs and MSFs was observed to boost for arousal, nevertheless, rendered a low performance on the valence dimension. Clearly, the wavelet transform analysis of low-dimensional features in speech advances the mSER. The wavelet energy- and entropy-related features contributed to the study of the perceptual process of emotion and also led to a better understanding of their dynamics.

5. Conclusions

This paper studied the perceptual process of emotion for the language diversity in a three-layer model. A new set of acoustic features derived from the wavelet transform analysis of speech signal was proposed to capture the emotional information across languages better. The effectiveness of the proposed features was assessed under LOSO validations across three emotional corpora, yielded significant improvement over conventional prosodic and spectral features in the literature. The proposed method is potentially intriguing to underlie the dynamic process in emotional speech across languages and could be coped with affective speech-to-speech translation systems.

6. References

[1] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.

[3] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," in *Interspeech*, 2010.

[4] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE transactions on Affective computing*, vol. 4, no. 4, pp. 386–397, 2013.

[5] H. Hu, M.-X. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *ICASSP*, vol. 4. IEEE, 2007, pp. IV–413.

[6] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *ICASSP*, vol. 4. IEEE, 2007, pp. IV–1085.

[7] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[8] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[9] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech*, 2014.

[10] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *ICASSP*. IEEE, 2011, pp. 5688–5691.

[11] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[12] E. Brunswik, *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.

[13] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[14] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[15] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology*, vol. 35, no. 2, pp. 86–98, 2014.

[16] X. Li and M. Akagi, "A three-layer emotion perception model for valence and arousal-based detection from multilingual speech," in *Interspeech*, 2018.

[17] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[18] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE signal processing letters*, vol. 8, no. 7, pp. 196–198, 2001.

[19] E. Avci and Z. H. Akpolat, "Speech recognition using a wavelet packet adaptive network based fuzzy inference system," *Expert Systems with Applications*, vol. 31, no. 3, pp. 495–503, 2006.

[20] J. N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *ICASSP*, vol. 3. IEEE, 2000, pp. 1351–1354.

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

[22] G. Sanchez, H. Silen, J. Nurminen, and M. Gabbouj, "Hierarchical modeling of f0 contours for voice conversion," in *Interspeech*, 2014.

[23] J. Kane, M. Aylett, I. Yanushevskaya, and C. Gobl, "Phonetic feature extraction for context-sensitive glottal source processing," *Speech Communication*, vol. 59, pp. 10–21, 2014.

[24] S. Deb and S. Dandapat, "Emotion classification using dual-tree complex wavelet transform," in *2017 14th IEEE India Council International Conference (INDICON)*. IEEE, 2017, pp. 1–5.

[25] Y. Li, J. Li, and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 908–916, 2018.

[26] P. Mokhtari and N. Campbell, "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech," *IEICE TRANSACTIONS on Information and Systems*, vol. 86, no. 3, pp. 574–582, 2003.

[27] K. E. Cummings and M. A. Clements, "Application of the analysis of glottal excitation of stressed speech to speaking style modification," in *ICASSP*, vol. 2. IEEE, 1993, pp. 207–210.

[28] K. Cummings and M. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *The Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 88–98, 1995.

[29] H. Kruschke and M. Lenz, "Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[30] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[31] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[32] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *International journal of speech technology*, vol. 16, no. 2, pp. 143–160, 2013.

[33] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Başar, "Wavelet entropy: a new tool for analysis of short duration brain electrical signals," *Journal of neuroscience methods*, vol. 105, no. 1, pp. 65–75, 2001.

[34] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [book review]," *IEEE Transactions on automatic control*, vol. 42, no. 10, pp. 1482–1484, 1997.

[35] X. Li and M. Akagi, "Multilingual speech emotion recognition system based on a three-layer model." in *Interspeech*, 2016, pp. 3608–3612.

[36] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants." in *Interspeech*, 2016, pp. 262–266.

[37] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.