Interspeech 2018
2-6 September 2018, Hyderabad

A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech

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

Japan Advanced Institute of Science and Technology

lixingfeng@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

Automated emotion detection from speech has recently shifted from monolingual to multilingual tasks for human-like interaction in real-life where a system can handle more than a single input language. However, most work on monolingual emotion detection is difficult to generalize in multiple languages, because the optimal feature sets of the work differ from one language to another. Our study proposes a framework to design, implement, and validate an emotion detection system using multiple corpora. A continuous dimensional space of valence and arousal is first used to describe the emotions. A three-layer model incorporated with fuzzy inference systems is then used to estimate two dimensions. Speech features derived from prosodic, spectral, and glottal waveform are examined and selected to capture emotional cues. The results of this new system outperformed the existing state-of-the-art system by yielding a smaller mean absolute error and higher correlation between estimates and human evaluators. Moreover, results for speaker independent validation are comparable to human evaluators.

Index Terms: emotion recognition, emotion dimension, three-layer model, prosodic feature, spectrogram, glottal waveform

1. Introduction

Identifying an emotional state from human voices based on speech emotion recognition (SER) has been an increasing area of focus within affective computing for interpreting the semantics of a spoken utterance. The purpose is to create a natural human-machine interaction in a real-world-context [1]. These days, psychology has proven that the human ability to perceive emotions is cross-lingual with no utterance required in his/her native or foreign language [2–4] . Despite great progress made in SER, natural human-behavior interaction is still an obstacle, on the grounds that there is a strong dependence on the language being spoken, i.e. the best vocal features usually differ from one language to another, which in turn are harder to generalize in multilingual tasks [5] [6]. Automated multilingual SER, however, still faces several problems that are not yet solved, such as the definition of human emotions, the appropriate ability to recognize and to predict emotion, and the extraction of representative and generalizable vocal features. This study addresses each of these issues with our SER system that can handle multiple input languages.

The first important issue in the design of the SER system is defining the human emotions to be predicted. Other approaches used a small set of discrete emotional classes to define happiness, anger, sadness, and so on [7]. However, emotions are not constant, which may change the intensities in the course of time and the surroundings [8]. Hence, a description using just one categorical label is not sufficient. Our study characterized emotions by using a two-dimensional emotional space spanned by arousal (relaxed vs. aroused) and valence (pleasant and unpleasant) after Russell's study [9], which provided a framework for detecting the dynamics in gradual emotion transitions in day-to-day life.

The second important issue in SER is the need to furnish a recognition model to predict valence and arousal. A number of effort has been done to be able to predict emotion dimensions from acoustic correlates from various estimators such as a fuzzy inference system (FIS) and support vector regression [10] [11]. However, the limitation of these works lies in the fact that performance has been poor in terms of valence. Scherer [12] adopted a version of Brunswik's lens model, originally developed in 1956, [13] to perceive human-emotions by a multi-layer process. A three-layer model consists of acoustic features, semantic primitives, and emotion dimensions, which was later adopted for accurate estimation of emotion dimensions [14]. This model reported reasonable results on both the valence and arousal dimensions. Our study was inspired by this human-perceptual-based strategy. We originally examined this three-layer model for multilingual SER tasks, and confirmed that it was well suited for mimicking human emotion perception processing across languages [15].

The third important issue to be considered is the extraction of the best features that can efficiently work for SER in multiple languages. In the past, a great amount of work has been published on developing the best features to characterize different emotions in monolingual speech [16]. Unfortunately, collected sets from different works were not consistent. Differences in the features for different languages made multilingual SER tasks quite difficult. Our earlier work [15] and [17] analyzed a set of features including F0, power-envelope, voice-quality, power-spectrum and duration in a three-layer model, and yielded a comparable performance to human evaluators among three languages [15]. Despite the substantial performance reported, there were two restrictions for practical applications in those studies.

First, the approach to feature extraction relied on manually segmenting speech signal to underlying phrases and phonemes, and then calculated one feature vector for each segmented phrase and phoneme, which was not realistic for automated SER in real-life scenarios. Second, speech emotional content required comprehensive characterization, which may not be given yet. When characterizing emotion, examined features are mainly from a prosodic domain that is well suited for describing the arousal dimension; however, are slightly limited for valence. As a result, that study reported a comparatively poor performance in the case of valence.

To address these problems of feature extraction, our study focuses on extracting feature vectors from the utterance level, as it can be adopted easier into real-life settings automatically. Moreover, aside from the prosodic features, two more analysis

10.21437/Interspeech.2018-1820

domains referred to as spectral and glottal waveform were further explored toward a comprehensive characterization of emotional speech. Prosodic features were first decided on the grounds that these features were advantageous for distinguishing low and high arousal emotions in accordance with human perceptions [18] [19]. Additionally, we explored the spectral features in light of the fact that these features are generally treated as strong correlates of the varying shapes of the vocal tract and rate of change in articulator movements [20]. It has been further reported that the emotion dimension of valence was also reflected in the acoustic parameters of the spectral features [21] [22]. Finally, the glottal waveform was highly significant perceptually [23] [24], and was greatly modified by the emotional state and general speaking manner of the speaker [25] [26]. Parametric analysis of glottal source components in speech further offers rich information for capturing emotional cues. We hypothesized that the combination of features from the prosodic, spectral, and glottal waveform domains can improve the estimation performance of valence and arousal.

In line with these findings, the main contribution of this study toward multilingual SER is to define a robust set of combined features to characterize emotional states among languages. Extensive evaluations were performed from different aspects: i) exploring the impact of the proposed features to compare the performance of our system with literature; ii) assessing the generalization ability of the proposed system by conducting cross-speaker validation; iii) enhancing the developed set of features by comparing it to a language-dependent set of features in a monolingual scenario.

2. Database

We experimented with three corpora of acted emotions in different languages: Japanese, German, and Chinese. In addition, to train the system and compare performances using the three corpora, four similar emotions including neutral, happiness, anger and sadness were selected.

Fujitsu The Japanese corpus was the Fujitsu Database recorded by Fujitsu Laboratory. In this corpus, a professional actress was asked to utter a sentence using five emotions: neutral, happiness, cold anger, sadness, and hot anger. There were 20 different sentences. Each sentence had one neutral utterance and two utterances 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.

Berlin The German corpus was the well-known Berlin Emo-DB. Ten professional actors (five males and five females) each uttered ten sentences in German to simulate seven different 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 data.

Casia The Chinese corpus was released by the Institute of Automation, Chinese Academy of Sciences (Casia). It was composed of 9600 utterances including six emotions: neutral, anger, fear, surprise, happiness, and sadness. Four professional actors (two males and two females) individually simulated each of these emotions and produced 400 utterances in six categories of different emotions. Ultimately, 200 utterances of spontaneous content from four actors covering four emotions (neutral, happiness, sadness, and anger) were selected, i.e. 50 utterances in each emotion.

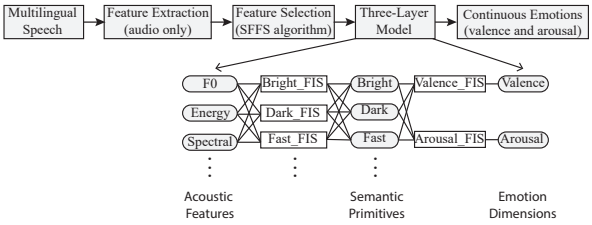

Figure 1: *Schematic diagram of the three-layer model based multilingual emotion recognition system.*

3. Methodology

Figure 1 shows the structure of our multilingual SER system. Feature extraction was first performed on multilingual emotional speech. Selection of the most relevant features was then done by sequential floating forward selection (SFFS). The three-layer model incorporating fuzzy inference systems took the best features as input and mapped them into valence and arousal dimensions through semantic primitives.

3.1. Acoustic Features

To automate SER from multilingual speech, identifying effective features is a significant task. We developed a robust set of combined features from three analysis domains including prosodic, spectral, and glottal waveform; and produced a total of 260 acoustic correlates as follows.

Prosodic related features: We first extracted three low level descriptors of fundamental frequency, short-term energy, and energy entropy using STRAIGHT [27]. Eight statistical functions per utterance were then calculated for each low level descriptor: mean, median, maximum, minimum, standard deviation, difference between maximum and minimum, 25% and 75% quantiles. All together, we extracted 24 prosodic features from speech signals.

Spectral features: Modulation spectral features (MSFs) were examined in the spectral domain of emotional speech. In contrast to conventional spectral features, which use mel-frequency cepstral coefficients (MFCC) that convey a signal's short-term spectral properties only. MSFs are based on a frequency analysis of the temporal envelope of multiple acoustic frequency bins, capturing both spectral and temporal properties of speech signals used by human listeners. This has been supported by reports that MSFs outperformed MFCC in SER [28]. We calculated five statistical functions of spectral flatness, spectral centroid, and 2nd to 4th central moments of the modulation spectrogram on the domains of 32 acoustic-frequency bands and six modulation-frequency bands. In addition, the modulation spectral tilt was calculated on the modulation-frequency domain, providing 196 MSFs in total. The implementation we used was first published in [29].

Glottal waveform features: For the glottal waveform (GW) features, we used COVAREP (v1.4.2), a freely available open source Matlab and Octave toolbox for speech analysis [30]. Particularly, we extracted eight GW features including normalized amplitude quotient, quasi open quotient, the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum, parabolic spectral parameter, maxima dispersion quotient, spectral tilt/slope of wavelet responses, shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd), and the confidence value of Rd. The individual features were obtained by

calculating statistical values of each extracted parameter: mean, variance, standard deviation, skewness, and kurtosis. All together, we extracted 40 glottal waveform features per speech.

3.2. Primitives-based emotion evaluation

We defined a human-perceptual based framework to predict emotions from multilingual speech using a three-layer model, where it was assumed that the human perception of emotion embedded in speech did not originate directly from a change in acoustic cues, but from an indirect route of more subtle perception of semantic primitives. Low arousal and negative valence speech easily make an impression on listeners with dark and heavy feelings, but high arousal and positive valence speech is oftentimes uttered in a bright and well-modulated way. The set of semantic primitives derived from [4] that we examined and used in the three-layer model for describing emotional speech was: 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 through human listening tests. Emotional speech was evaluated 17 times by participants; 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. Additionally, in light of the fact that this study characterized emotions using a dimensional space spanned by valence and arousal, these corpora needed to be further annotated in terms of emotional dimensions. The same participants 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 dataset. Unfortunately, it was impossible for us to recruit enough native German speakers for the listening test; so we asked nine Japanese speakers (eight males and one female) 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 the listeners to understand the adjectives or dimensions. All stimuli were played randomly through binaural headphones at a comfortable sound pressure level in a soundproof room.

For each instance of speech n per corpus c, where $c \in \{Fujitsu, Berlin, Casia\}$, $1 \leq n \leq N$, the averaged ratings $\overline{x}_{n,c}^{(p)}$ of listeners' responses $\widehat{x}_{n,c}^{e,(p)}$ among all evaluators E were calculated for each semantic primitive (s) or emotion dimension (d).

$$\overline{x}_{n,c}^{(p)} = \frac{1}{E}\sum_{e=1}^{E}\widehat{x}_{n,c}^{e,(p)}, with\ p = \begin{cases} s,\ semantic\ primitive \\ d,\ emotion\ dimension \end{cases}. \quad (1)$$

The inter-evaluator agreement was evaluated using Eq. 2 following the related study reported in [10].

$$CC_c^{e,(p)} = \frac{\sum_{n=1}^{N}\left(\widehat{x}_{n,c}^{e,(p)} - \frac{1}{N}\sum_{n'=1}^{N}\widehat{x}_{n',c}^{e,(p)}\right)\left(\overline{x}_{n,c}^{(p)} - \frac{1}{N}\sum_{n'=1}^{N}\overline{x}_{n',c}^{(p)}\right)}{\sqrt{\sum_{n=1}^{N}\left(\widehat{x}_{n,c}^{e,(p)} - \frac{1}{N}\sum_{n'=1}^{N}\widehat{x}_{n',c}^{e,(p)}\right)^2}\sqrt{\sum_{n=1}^{N}\left(\overline{x}_{n,c}^{(p)} - \frac{1}{N}\sum_{n'=1}^{N}\overline{x}_{n',c}^{(p)}\right)^2}}. \quad (2)$$

Table 1: *Average correlation coefficient ($\overline{cc}$) for semantic primitives of Fujitsu, Berlin, and Casia corpora by human listeners; averaged of all speakers and whole utterances.*

	Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow
Fuji	.92	.90	.89	.91	.92	.91	.89	.90	.91	.89	.88	.88	.90	.93	.88	.85	.84
Berlin	.86	.90	.87	.89	.93	.93	.89	.85	.90	.87	.86	.89	.87	.91	.87	.84	.86
Casia	.82	.87	.86	.92	.91	.91	.88	.82	.82	.85	.85	.83	.89	.86	.89	.90	.91

Table 2: *Average correlation coefficient ($\overline{cc}$) for the emotion dimensions of Fujitsu, Berlin, and Casia corpora by human listeners; averaged of all speakers and whole utterances.*

	Valence	Arousal
Fuji	.96	.96
Berlin	.92	.94
Casia	.85	.91

The averaged results of inter-evaluator correlation for semantic primitives and emotion dimensions of the three emotional corpora were individually listed in Tables 1 and 2. The $\overline{cc}$ in the evaluations of the semantic primitives and emotion dimensions were identical between the three corpora with values ranging from 0.82-0.93 and 0.85-0.96, indicating good evaluation results. In particular, it can be found that the inter-rater agreement was generally lower for valence than for arousal, indicating human evaluations are more poorly correlated in terms of valence compared to that of arousal.

3.3. Feature selection

Large feature sets not only have exorbitant costs in terms of time for system training, but they also involve irrelevant features that reduce recognition accuracy. In this regard, we used the SFFS to select the best features from original sets of 260 acoustic features and 17 semantic primitives, separately. SFFS is an iterative algorithm to evaluate the selected subset and combined effects of features and k-nearest-neighbour classifier during the evaluation process. Nine acoustic features and four semantic primitives were finally used in this work.

4. Experiment

Three experiments were conducted to evaluate the efficiency of the proposed approaches. First, we showed the relevance of our developed set of features by comparing it to a previous study, which we named MultiBaseline (MB) that was mainly based on prosodic features [15]. Second, we carried out a leave-one-speaker-out (LOSO) validation to assess the speaker independence by comparing it to human evaluators as a reference. Finally, we defined three monolingual tests with the developed set of features and compared it to a system that was conducted with the best language-dependent features [15].

Adaptive neuro fuzzy inference systems (ANFIS) were used in the three-layer model to estimate continuous emotions. The ANFIS was chosen on the grounds that it could efficiently model nonlinear input and output relations by incorporating human knowledge with a lower root mean square error [31]. Correspondingly, the nature of perception of speech emotion was fuzzy and vague [10]. Furthermore, our three-layer model

incorporated human knowledge from evaluations of semantic primitives and emotion dimensions, which involved nonlinear processing according to human emotion perception. To estimate continuous emotion, each of the four semantic primitives in the middle layer was predicted separately from nine acoustic features using four FISs. Beyond that, the estimation of emotion dimensions was done from four estimated adjectives in the previous part by another two FISs.

4.1. Results on Multilingual evaluation

All results are given by a 10-fold cross-validation, where the training and test data were collected by merging the three corpora. We evaluated the systems using both the CC and mean absolute error (MAE). The CC measured the agreement between two variables, in this case, the averaged human evaluators and the estimations by systems. The CC was calculated by Eq. 2. The MAE was calculated as follows.

$$MAE^{(d)} = \frac{1}{N}\sum_{n=1}^{N} \mid \widehat{x}_n^{(d)} - \overline{x}_n^{(d)} \mid . \quad (3)$$

where $d \in \{valence, arousal\}$, $\widehat{x}_n^{(d)}$ is the output of the system, $\overline{x}_n^{(d)}$ is the averaged values from human evaluators, and N is the total number of utterances in the three corpora.

Table 3 lists the CC and MAE for each system, for valence and arousal separately. As seen, the proposed multilingual SER furnished a notable result in estimation of the valence and arousal, with a CC of 0.87 and 0.96, while the MAE was 0.38 and 0.22, respectively. This yielded a relative error reduction rate of 13% and 20% for CC, and 7.3% and 12% for MAE on valence and arousal respectively in comparison with that of the MB. It is notable that the combination of vocal features from prosodic, spectral, and glottal waveform domains improved the estimation of emotional dimensions in multiple languages significantly.

4.2. Results on LOSO

The LOSO cross-validation was performed by training on all but one speaker's data, and then was tested on the held data. The held-out speaker (S) was rotated until all speakers were tested. The LOSO results of the average correlation coefficient ($\overline{cc}$) and MAE ($\overline{mae}$) of valence and arousal for each speaker were reported in Table 4, and were compared with the $\overline{cc}$ and standard deviation of evaluations among human evaluators in the listening test as a reference.

As can be seen, for all speakers, the values of the $\overline{cc}$ were found to be positive; and for most of these speakers, we found a fairly high correlation in the range of 0.75-0.97. This result was comparable to that of the human evaluations with the values of the $\overline{cc}$ ranging from 0.82-0.96. Furthermore, the values of the $\overline{mae}$ obtained from the LOSO validation were between 0.14 and 0.51 for the different speakers, and were comparable to the $\overline{std}$ of the human evaluations. These errors are in the range of half the distance between two evaluated scales in the human listening test and thus notably small. It can be summarized that the proposed framework is well suited for valence and arousal-based detection for different speakers among languages.

4.3. Results on Monolingual evaluation

We defined a set of combined features that could efficiently work for multiple languages. To further assess the efficiency of this set, we performed three monolingual tests (MonoP) using these same features, and compared them to the systems conducted with the best language-dependent features (MonoB) [15]. All results were presented by using a 10-fold cross-validation.

The results of the CC and MAE for MonoP and MonoB are shown in Table 5. As can be seen, MonoP yielded both a higher CC and lower MAE for all emotion dimensions on all corpora. Again, this performance in turn proved that our developed features outperformed the language-dependent set of features in [15]. On average, the CC and MAE over valence and arousal was 0.99 and 0.14 for Fujitsu, 0.92 and 0.28 for Berlin, and 0.91 and 0.29 for Casia, which was consistent with human evaluators, cf. Table 2.

Table 3: *Results of CC and MAE obtained for Valence (V) and Arousal (A) by the proposed multilingual SER (MP) and MultiBaseline (MB)*

	CC		MAE		**Improvement rel. in %**	
	MP	MB	**MP**	MB	CC	MAE
V	**.87**	.85	**.38**	.41	**13**	**7.3**
A	**.96**	.95	**.22**	.25	**20**	**12**

Table 4: *Results of average CC and MAE of valence and arousal obtained for each speaker by LOSO validation compared to the mean CC and standard deviation of human evaluators (HEva)*

		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
LOSO	$\overline{cc}$	**.89**	**.93**	**.91**	**.97**	**.95**	**.97**	**.61**	**.97**	**.91**	**.96**	**.89**	**.88**	**.69**	**.77**	**.75**
	$\overline{mae}$	**.32**	**.33**	**.23**	**.19**	**.25**	**.14**	**.48**	**.25**	**.24**	**.21**	**.51**	**.41**	**.43**	**.43**	**.43**
HEva	$\overline{cc}$	.92	.96	.91	.92	.93	.91	.93	.94	.95	.94	.96	.92	.82	.88	.87
	$\overline{std}$	.43	.42	.49	.52	.49	.48	.44	.42	.37	.44	.38	.40	.49	.40	.47

Table 5: *Results of CC and MAE obtained on V and A by MonoP and MonoBaseline (MonoB) for Fujitsu, Berlin and Casia*

		Fujitsu		Berlin		Casia	
		V	A	V	A	V	A
CC	**MonoP**	**.99**	**.99**	**.86**	**.97**	**.88**	**.93**
	MonoB	.96	.99	.84	.96	.86	.93
MAE	**MonoP**	**.16**	**.12**	**.37**	**.18**	**.31**	**.27**
	MonoB	.26	.15	.41	.21	.37	.27

5. Conclusions

We presented a framework for multilingual SER. A set of generalizable features was addressed from the prosodic, spectral, and glottal waveform domains irrespective of languages. As demonstrated, this set significantly improved the estimation of emotion dimensions compared with results shared in related literature. It further confirms the validity of LOSO validation reports as a comparable performance to human evaluation; in particular, for estimating continuous emotion in a monolingual scenario. The developed features even outperformed a language-dependent feature set. The advantage of a multilingual continuous SER is that it could be used to build an affective speech-to-speech translation system that is capable of handling multiple input languages.

6. Acknowledgements

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026), and a China Scholarship Council (CSC) Scholarship.

7. References

[1] 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.

[2] A. Tickle, "English and japanese speakers emotion vocalizations and recognition: A comparison highlighting vowel quality," in *ISCA Workshop on Speech and Emotion*, 2000.

[3] R. Huang and C. Ma, "Toward a speaker-independent real-time affect detection system," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1204–1207.

[4] C. Huang, D. Erickson, and M. Akagi, "Comparison of japanese expressive speech perception by japanese and taiwanese listeners," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3323, 2008.

[5] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[6] K. Van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann, "Fully generated scripted dialogue for embodied agents," *Artificial Intelligence*, vol. 172, no. 10, pp. 1219–1244, 2008.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[8] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological Bulletin*, vol. 99, no. 2, p. 143, 1986.

[9] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[10] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[11] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation." in *INTERSPEECH*, 2010, pp. 785–788.

[12] 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.

[13] E. Brunswik, "Historical and thematic relations of psychology to other sciences," *Scientific Monthly*, vol. 83, no. 3, pp. 151–161, 1956.

[14] 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.

[15] X. LI and M. Akagi, "Multilingual speech emotion recognition using a three-layer model," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8–12, San Francisco, California, Proceedings*, pp. 3608–3612.

[16] 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.

[17] X. Li and M. Akagi, "Toward improving estimation accuracy of emotion dimensions in bilingual scenario based on three-layered model," in *O-COCOSDA/CASLRE*. IEEE, 2015, pp. 21–26.

[18] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," *Handbook of Emotions*, vol. 2, pp. 220–235, 2000.

[19] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1, pp. 5–32, 2003.

[20] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[21] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1322–1336, 2010.

[22] R. A. Calix, M. A. Khazaeli, L. Javadpour, and G. M. Knapp, "Dimensionality reduction and classification analysis on the audio section of the semaine database," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 323–331.

[23] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[24] R. H. Colton, "Discrimination of glottal wave form variations," in *Transcripts of the 11th Symposium Care of Professional Voice, Part 1: Scientific Papers*, 1982, pp. 61–68.

[25] K. E. Cummings and M. A. 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.

[26] E. Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–101.

[27] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[28] 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.

[29] 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.

[30] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 960–964.

[31] 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.