Maximal Information Coefficient and Predominant Correlation-Based Feature Selection Toward A Three-Layer Model for Speech Emotion Recognition

Xingfeng Li * and Masato Akagi †
Japan Advanced Institute of Science and Technology, Nomi, Japan
E-mail: lixingfeng@jaist.ac.jp, akagi@jaist.ac.jp

***Abstract*—This paper describes an efficient attempt to build a three-layer emotion perception model consisting of acoustic features, semantic primitives, and emotion dimensions with a focus on acoustic feature subset selection. Previous studies using this model focused on the most relevant acoustic features using a Pearson correlation coefficient-based filter approach, which could only capture the relation limited to linear function well. However, perception of human emotion is vague; linear correlation measures could not capture the relations that are not linear in nature. In this study, we introduce a novel feature selection algorithm based on the maximal information coefficient and predominant correlation, which can identify relevant features between paired variables in spite of linear or nonlinear relations and remove redundancies among the relevant features. Experimental results on the Berlin Emo-DB and Chinese Emotional Speech Corpus demonstrated that the proposed algorithm achieves an improvement to estimation of emotion dimensions, resulting in a smaller mean absolute error and higher correlation coefficient between estimations and human evaluations, compared with the referred Pearson correlation coefficient-based method, and the commonly used wrapper-based method of sequential floating forward selection.**

I. Introduction

Speech emotion recognition, which aims to identify an emotional state from human voices, has gained more attention in the area of affective computing. The purpose is to enrich user-friendly computer-based interaction in a system to understand human behaviors not only in terms of what a person says, but also how it is expressed [1]. In call-center applications, for instance, speech emotion recognition could be adopted to enable the application to respond better upon identifying frustration or annoyance in a customer's voice [2]. Likewise, information about the mental state of a driver can be also provided to an in-car board system to initiate strategies for his or her safety [3]. Moreover, an automatic speech-to-speech translation system may be greatly enriched by understanding the emotions of a speaker in affective communication between parties [4]. In this paper, we focus on automatic emotion recognition from speech, and specifically, presenting an algorithm of acoustic feature subset selection with the goal toward implementing a three-layer emotion perception model.

In recent years, there are two major approaches that define vocal emotions: categorical and dimensional-based. The categorical-based approach divides human emotions into a small set of discrete categories, such as happiness, anger, sadness, and so on [5] [6]. However, the emotional expression of natural speech in general is not binary, i.e., anger or sadness, but may change the intensities of a certain emotion over time, such as a little anger or much anger [7] [8]. To describe the rich variation of the intensity of emotional states, the approach for emotion definition has shifted from categorical to dimensional-based, where emotions can be defined as points in a dimensional space spanned by valence (pleasant and unpleasant) and arousal (relaxed and aroused). Our study uses this two-dimensional emotion space to characterize emotions embedded in speech.

The dimensional-based approach contributes to resolving a significant challenge in speech emotion recognition, namely describing emotion transitions gradually and continuously. However, studies related to speech emotion recognition still face several challenges, one of which is the accurate estimation of dimensions of valence and arousal. Many studies focused on acoustic features and their correlates to emotion dimensions by incorporating different estimators such as a fuzzy inference system (FIS) and support vector regression [9] [10]. The limitation of these works, however, is that performance has been poor in terms of valence. This is possibly due to the difficulty of determining a set of distinguished acoustic features for estimating valence.

To overcome this challenging task on improving estimation accuracies on emotion dimensions directly from acoustic features, reference [11] presented a three-layer model, inspired by an adapted version of Brunswik's lens model [13], to describe human emotion perception as a multi-layer process. This model consists of acoustic features, semantic primitives, and emotion dimensions, where it assumed that human perception of emotion embedded in speech did not originate directly from a change in acoustic cues but from an indirect route of a more subtle perception of semantic primitives. This reference later ensured the fact that the selected acoustic features that were highly correlated

with semantic primitives were promising for improving estimation accuracies of emotion dimensions, especially for valence.

As this three-layer model benefits from determining relevant acoustic features with emotion dimensions through semantic primitives, a new challenge arose: how to select relevant acoustic features for semantic primitives from a rough set of data. One way to select a set of compact features for pattern recognition that has been heavily used is to use a wrapper-based method [14]. This approach generally incorporates a specific learning algorithm into a close-loop for searching an optimal subset of features, and can potentially achieve a better learning performance. However, it also has a higher risk of over-fitting and is very computationally intensive [15] [16]. As an alternative to the wrappers, a filter-based approach is comparatively appealing for feature selection on the grounds that it could easily be scaled up to high-dimensional speech features, it is computationally simple and fast, and it is not dependent on the classifiers or estimators [17].

Reference [11] adopted a feature selection algorithm by incorporating the Pearson correlation coefficients to select relevant features by a linear measure (hereafter P-CCFS). They first calculated the Pearson correlation coefficient (PCC) between acoustic features and semantic primitives individually, and the degree of relevance was quantified within the range -1–1, where -1 is total negative linear correlation, 1 is total positive linear correlation, and 0 is no linear correlation. These correlation coefficients were then ranked on the basis of their absolute scale; those greater than 0.45 were finally selected as relevant acoustic features for semantic primitives.

Despite the substantial advances in the three-layer model, the limitation of P-CCFS, however, are considerable due to following reasons. First, the PCC can naively capture linear relationship between features and target, but can not capture correlations that are not linear in nature. Human emotion perception, for instance, is vague, complex, and has multi-processes; it does not suffice to always use linear correlation to capture the association between acoustic feature and semantic primitives. Second, although P-CCFS could define the relevant features by a decided threshold 0.45, it cannot determine any redundancies among them. Several studies on feature selection have further claimed that in addition to irrelevant features, redundant features affect the performance of learning algorithms as well, and thus, should also be removed [18] [19].

The motivation of the present study is to introduce a novel feature selection algorithm with the capability to effectively determine the relevant features and remove any redundancies among them. To this end, we first present a correlation measure to quantify the degree of relevance between features and target on the basis of the maximal information coefficient (MIC) [20], which could capture a wide range of associations between paired variables in spite of linear or non-linear relations, that is well suited to measure relations between acoustic features and semantic primitives in human vague emotion perception. Additionally, we adopt a concept of predominant correlation (PC) after [21] by incorporating the MIC to remove any redundancies among the relevant acoustic features. The proposed feature selection algorithm is hereafter called MIC-PCFS.

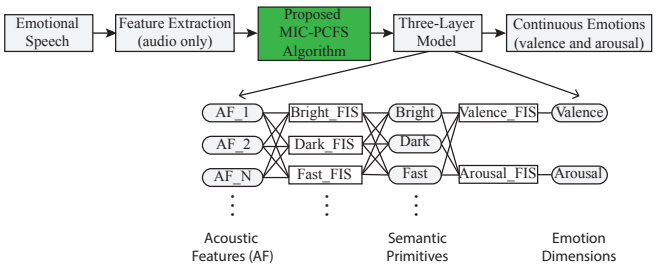

Fig. 1. *Schematic diagram of the three-layer emotion perception model*

The remainder of the paper is organized as follows. Section 2 gives a description of the three-layer emotion perception model and introduces the emotional corpora used for the experiments. Section 3 presents MIC-PCFS in detail. Section 4 reports the results of the experiments on emotion estimation by incorporating MIC-PCFS, and comparisons with other representative feature selection methods. Final remarks are given in Section 5.

II. Emotional Corpus And Recognizer

Fig. 1 illustrates a schematic diagram of the present study's system. The colored block highlights the scope of this research to introduce a novel feature selection algorithm, with the aim to implement a three-layer model for speech emotion recognition. After receiving emotional speech as input, features are then extracted. A set of compact relevant acoustic features is then selected by MIC-PCFS. The three-layer model incorporating fuzzy inference systems finally takes the relevant features as input and maps them into valence and arousal dimensions through semantic primitives.

A. Emotional Corpus

The Berlin emotional speech database (Berlin Emo-DB), which was previously used in the three-layer model in [11], is first selected to analyze the performance of MIC-PCFS for speech emotion recognition. Additionally, we further evaluate the proposed approach using the Chinese emotional speech corpus (CASIA) for discussion.

Berlin Emo-DB The German corpus was released by the Institute of Speech and Communication, Technical University of Berlin. 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 as follows: 127 anger, 81 boredom, 46 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness. Finally, 200 utterances previously used in [11] were selected from this corpus with 50 utterances in each of the four emotion categories: neutral, joy, anger, and sadness.

CASIA The Chinese corpus was released by the Institute of Automation, Chinese Academy of Sciences. 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 the actors covering four similar emotions as those in Berlin Emo-DB (neutral, happiness, sadness, and anger) were selected, i.e. 50 utterances for each emotion.

B. Emotional Recognizer

The elements in terms of acoustic features, semantic primitives, and emotion dimensions towards constructing the three-layer model are detailed in this subsection.

Acoustic Features Feature extraction was performed per utterance by the widely used OpenSMILE feature extraction toolkit incorporating $emo\_large.conf$ [22], producing 6552 audio-based features. This feature set includes 56 low-level descriptors (LLDs), such as signal log-energy, fundamental frequency, 13 Mel-frequency cepstrum coefficients etc., and the first and second derivatives of each LLD. 39 statistical functional parameters such as mean, standard deviation, skewness, kurtosis, etc., are applied to the LLDs compute each of the emotional utterances. For more details on the $emo\_large.conf$ set, the reader is referred to [23].

Semantic Primitives and Emotion Dimensions We reported a human-perceptual-based framework to estimate emotions from 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 a more subtle perception of semantic primitives. Low arousal and negative valence speech often implies that the speech was uttered 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 [12] 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 two 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, as this study characterizes emotions using a dimensional space spanned by valence and arousal, the 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).

Ten native Chinese speakers (five males and five females) were asked to evaluate CASIA. However, it was impossible for us to recruit enough native German speakers for the listening test. Fortunately, several studies on psychology have confirmed that the human ability to perceive emotion is cross-lingual, even without understanding the linguistic information expressed [4], [24], [25]. So, we asked nine Japanese speakers (eight males and one female) to evaluate 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 was designed 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.

The averaged results of inter-evaluator correlation for the semantic primitives in terms of Berlin Emo-DB and CASIA were almost identical with values ranging from 0.84–0.93 and 0.82–0.92, respectively. In addition, the average correlation between evaluators over valence, and arousal was 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 in terms of valence compared to that of arousal.

III. Feature Selection Algorithm

This section describes the proposed acoustic feature selection algorithm to define the relevant features for the semantic primitives, removing any possible redundancies among them, enabling the implementation of a three-layer model for speech emotion recognition.

A. Maximal Information Coefficient

The MIC was originally noted by Reshef [20] to identify a novel measure of dependence for two-variable relationships, which enables the capturing of a wide range of both functional and nonfunctional associations. As the processing of human vague emotion perception is complex due to the possibility of both linear and nonlinear associations existing, this measurement is well suited to quantify the degree of relevance between acoustic features and semantic primitives.

Formally, the MIC can be given as follows: let D be a finite set of two-variable data whose value is into the range 0–1, as (X,Y). Supposing X is partitioned into x bins, and Y is partitioned into y bins. For a grid G, let $D|_G$ represent the probability distribution induced by the D in the cells of G, and let $I(\cdot)$ denote mutual information.

$$I^*(D,x,y) = max_G I(D|_G) \tag{1}$$

where the maximum is taken from all x-*by*-y grids G. Furthermore, the MIC is obtained as:

$$MIC(D) = \underset{xy<N^{0.6}}{max} \frac{I^*(D,x,y)}{\log_2 min\{x,y\}} \tag{2}$$

where N is the sample size.

The value of the MIC falls into the range 0–1; a higher value of MIC suggests a higher relevance between a feature and target, and a value of zero denotes an independent relationship. MIC is additionally symmetric in that MIC(X,Y)=MIC(Y,X). The calculation function $MINE$ for MIC has been implemented and is available at $http://minepy.readthedocs.io/en/latest/$.

In this stage, a set of relevant acoustic features for each of the semantic primitives can be defined on the basis of the MIC value as follows: let $F = \{f_i | i = 1, 2, ..., N_f\}$ be a rough set of acoustic features, where N_f is the total number of acoustic features, and let $S = \{s_j | j = 1, 2, ..., N_j\}$ be a set of semantic primitives, where N_s is the total number of semantic primitives. $MIC_{i,j}$ denotes a MIC value between an acoustic feature f_i and a semantic primitive s_j. A set of relevant acoustic features for a semantic primitive s_j can be obtained as F_{rel}, where $F_{rel} = \{f_i | f_i \in F, MIC_{i,j} \geq \delta_{s_j}\}$, and δ_{s_j} is a pre-defined threshold MIC value.

B. MIC-based Predominant Correlation

A set of relevant features is traditionally good as long as the weight of relevance for each relevant feature is greater than a threshold value, even if some of these relevant features are highly related with each other [26]. However, studies on feature selection have shown that combining relevant features does not necessarily result in a good performance, moreover, the redundant features should also be considered [18] [19] [27] [28].

To determine a compact set of relevant acoustic features, this stage adopts a concept of predominant correlation after [21] by incorporating the MIC to eliminate redundancies as follows.

◎ The correlation between an acoustic feature f_i and a semantic primitive s_j is predominant iff $MIC_{i,j} \geq \delta_{s_j}$, and $\forall f_m \in F_{rel}$ $(m \neq i)$, there exists no f_m such that $s.t.$ $MIC_{m,i} \geq MIC_{i,j}$. In particular, if there exists such f_m to a feature f_i, then f_m is a redundant peer to f_i.

◎ An acoustic feature is predominant to a semantic primitive, iff its correlation to the semantic primitive is predominant or can be predominant after eliminating its redundant peers.

An acoustic feature f_i is good if its relevance to a semantic primitive s_j, i.e. $MIC_{i,j}$, is at the maximum when compared with the rest of the acoustic features in the relevant set F_{rel}, and is not redundant with those that have already been decided. The idea of the acoustic feature selection algorithm is to identify and keep the predominant acoustic features.

The pseudo code of MIC-PCFS is as follows.

As demonstrated in Algorithm 1, MIC-PCFS defines a set of predominant acoustic features S^j_{Rel} for each semantic primitive per round, corresponding to bright (1), dark (2), high (3), low (4), strong (5), weak (6), calm (7), unstable (8), well-modulated (9), monotonous (10), heavy (11), clear (12), noisy (13), quiet (14), sharp (15), fast (16), and slow (17). It is a two-stage procedure: the first part starts from line 3–8, which calculates the MIC value for each acoustic feature stores the relevant acoustic features into F_{rel} on the basis of the threshold value of a semantic primitive δ_{s_j}, and sorts these features in descending order on the basis of their MIC values. The second part from line 9–22 focuses on removing the redundancies among the relevant acoustic features.

Algorithm 1 MIC-PCFS Algorithm

Input: S($f_1, f_2, ..., f_{N_f}, s_j$), and δ_{s_j}

Output: S^j_{Rel} // a set of defined relevant features for a semantic primitive j

1: for j = 1 to N_s do begin
2: **begin**
3: for i = 1 to N_f do begin
4: calculate $MIC_{i,j}$ for f_i ;
5: if $MIC_{i,j} \geq \delta_{s_j}$
6: stored f_i to F_{rel} ;
7: end ;
8: sorting F_{rel} by the $MIC_{i,j}$ value in descending order ;
9: f_m = $getFirstFeature(F_{rel})$;
10: do begin
11: f_n = $getNextFeature(F_{rel}, f_m)$;
12: if (f_n <> null)
13: do begin
14: f_n^{temp} = f_n ;
15: if ($MIC_{m,n} \geq MIC_{n,j}$)
16: remove f_n from F_{rel} ;
17: f_n = $getNextFeature(F_{rel}, f_n^{temp})$;
18: else
19: f_n = $getNextFeature(F_{rel}, f_n)$;
20: end until f_n == null ;
21: f_m = $getNextFeature(F_{rel}, f_m)$;
22: end until f_m == null ;
23: S^j_{Rel} = F_{rel} ;
24: **end**
25: end

IV. Experiment

This section aims to investigate whether MIC-PCFS provides a promising framework for implementing a three-layer model and improves the accuracy in tracking continuous emotions in valence-arousal space. To this end, we compare MIC-PCFS with two representative algorithms. P-CCFS was first selected on the grounds that it was the first attempt adopted for the task of implementing a three-layer model for emotion recognition. P-CCFS explores the $N - best$ relevant acoustic features with their weights of relevance greater than a threshold value. The sequential floating forward selection algorithm (SFFS) was also selected as it is one of the most promising feature selection methods for wrappers. It is an iterative algorithm that evaluates a selected subset and takes the combined effect of features into account.

Adaptive neuro-fuzzy inference systems (ANFISs) were used in the three-layer model to estimate valence and arousal. The ANFIS was selected on the grounds that it could efficiently model nonlinear input and output relations by incorporating human knowledge with a lower root mean square error [29]. Correspondingly, the nature of perception of speech emotion was fuzzy and vague [9]. Furthermore, our three-layer model incorporated human knowledge from evaluations of semantic primitives and emotion dimensions, which involved nonlinear processing in accordance with human emotion perception. More specifically, to estimate continuous emotion, each of the semantic primitives in the middle layer was predicted separately from the relevant acoustic features using 17 FISs. Beyond that, the estimation of emotion dimensions was done from 17 estimated adjectives in the previous part by another two FISs. Experimental results were reported by leave-one-speaker-out cross-validation on the emotional corpora of Berlin Emo-DB and CASIA.

The correlation coefficient (CC) and mean absolute error (MAE) were calculated to assess the performance of the estimation of semantic primitives and emotion dimensions. For the values of a semantic primitive or emotion dimension estimated by the system, i.e. X_n, and the corresponding averaged values of a semantic primitive or emotion dimension given by human estimators, i.e. Y_n, the CC and MAE are individually calculated as:

$$CC = \frac{\sum_1^N \left(X_n - \overline{X}\right)\left(Y_n - \overline{Y}\right)}{\sqrt{\sum_1^N \left(X_n - \overline{X}\right)^2 \sum_1^N \left(Y_n - \overline{Y}\right)^2}} \tag{3}$$

$$MAE = \frac{\sum_1^N |X_n - Y_n|}{N} \tag{4}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X_n and Y_n, respectively. In addition, N is the number of utterances of each emotional corpus. Notably, the values of the CC trend to 1 for a system's estimation closer to human evaluations, and the values of the MAE trend to 0 for a better performance of system's estimation.

Mean values of the CC and MAE for semantic primitives and emotion dimensions of Berlin Emo-DB, averaged over all speakers, are individually shown in Tables I and II, respectively. As can be seen from Table I, MIC-PCFS achieved a better estimation performance for most of the semantic primitives with high CC values ranging between 0.759–0.964, and low MAE values ranged between 0.052–0.096. On the other hand, as shown in Table II, MIC-PCFS also achieved a promising performance for estimating valence and arousal as indicated by the highest CC values of 0.849 and 0.970 and the lowest MAE values of 0.113 and 0.046 when compared with those obtained by P-CCFS and SFFS. This closer estimation reveals that the gain of MIC-PCFS over P-CCFS and SFFS is mainly due to the former's effectiveness on identifying both irrelevant and redundant acoustic features.

TABLE I
COMPARISON OF THE AVERAGING CC AND MAE FOR SEMANTIC PRIMITIVES OF BERLIN EMO-DB CORRESPONDING TO MIC-FS, P-CCFS, AND SFFS USING A THREE-LAYER MODEL.

Semantic	CC			MAE		
Primitives	**MIC-PCFS**	P-CCFS	SFFS	**MIC-PCFS**	P-CCFS	SFFS
1 bright	0.840	0.801	**0.844**	**0.092**	0.105	**0.092**
2 dark	**0.944**	0.922	**0.944**	0.076	0.088	**0.073**
3 high	**0.904**	0.866	0.897	**0.079**	0.093	0.083
4 low	**0.945**	**0.945**	0.943	**0.069**	0.071	**0.069**
5 strong	**0.932**	0.895	0.930	0.075	0.093	**0.071**
6 weak	**0.963**	0.925	0.961	**0.061**	0.088	0.062
7 calm	**0.948**	0.910	0.927	**0.064**	0.084	0.075
8 unstable	**0.932**	0.860	0.912	**0.067**	0.093	0.074
9 well modulate	**0.930**	0.879	0.921	**0.072**	0.095	0.076
10 monotonous	**0.922**	0.891	0.909	**0.068**	0.073	0.070
11 heavy	0.759	**0.775**	0.754	0.096	**0.092**	0.093
12 clear	**0.873**	0.802	0.838	**0.073**	0.086	0.080
13 noisy	**0.930**	0.871	0.916	**0.062**	0.087	0.069
14 quiet	**0.964**	0.938	0.947	**0.052**	0.066	0.060
15 sharp	0.921	0.895	**0.933**	0.068	0.079	**0.064**
16 fast	**0.856**	0.776	0.812	**0.074**	0.091	0.086
17 slow	**0.922**	0.860	0.890	**0.068**	0.085	0.078

TABLE II
COMPARISON OF THE AVERAGING CC AND MAE FOR EMOTION DIMENSIONS OF BERLIN EMO-DB CORRESPONDING TO MIC-FS, P-CCFS, AND SFFS USING A THREE-LAYER MODEL.

Emotion	CC			MAE		
Dimensions	**MIC-PCFS**	P-CCFS	SFFS	**MIC-PCFS**	P-CCFS	SFFS
Valence	**0.849**	0.821	0.814	**0.113**	0.121	0.116
Arousal	**0.970**	0.956	0.962	**0.046**	0.058	0.053

TABLE III
COMPARISON OF THE AVERAGING CC AND MAE FOR SEMANTIC PRIMITIVES OF CASIA CORRESPONDING TO MIC-FS, P-CCFS, AND SFFS USING A THREE-LAYER MODEL.

Semantic	CC			MAE		
Primitives	**MIC-PCFS**	P-CCFS	SFFS	**MIC-PCFS**	P-CCFS	SFFS
1 bright	**0.798**	0.663	0.537	**0.120**	0.150	0.180
2 dark	**0.794**	0.683	0.516	**0.124**	0.160	0.186
3 high	**0.783**	0.685	0.550	**0.143**	0.172	0.185
4 low	**0.801**	0.705	0.688	**0.122**	0.157	0.161
5 strong	**0.748**	0.693	0.498	**0.146**	0.158	0.181
6 weak	**0.769**	0.647	0.561	**0.150**	0.165	0.190
7 calm	**0.817**	0.623	0.346	**0.123**	0.185	0.217
8 unstable	**0.867**	0.716	0.691	**0.107**	0.145	0.152
9 well modulate	**0.698**	0.605	0.421	**0.175**	0.203	0.242
10 monotonous	**0.525**	0.505	0.315	**0.206**	0.199	0.233
11 heavy	0.649	**0.706**	0.549	0.171	**0.141**	0.190
12 clear	**0.566**	0.513	0.431	**0.185**	0.186	0.210
13 noisy	**0.749**	0.718	0.564	**0.146**	0.147	0.178
14 quiet	**0.863**	0.664	0.434	**0.103**	0.157	0.193
15 sharp	**0.779**	0.685	0.529	**0.128**	0.157	0.187
16 fast	**0.807**	0.778	0.609	**0.117**	0.130	0.158
17 slow	**0.835**	0.761	0.658	**0.118**	0.137	0.159

TABLE IV
COMPARISON OF THE AVERAGING CC AND MAE FOR EMOTION DIMENSIONS OF CASIA CORRESPONDING TO MIC-FS, P-CCFS, AND SFFS USING A THREE-LAYER MODEL.

Emotion	CC			MAE		
Dimensions	**MIC-PCFS**	P-CCFS	SFFS	**MIC-PCFS**	P-CCFS	SFFS
Valence	**0.510**	0.488	0.470	**0.160**	0.169	0.175
Arousal	**0.782**	0.781	0.687	**0.126**	0.130	0.148

In addition, Tables III and IV further depict the results of averaged CC and MAE values for semantic primitives and emotion dimensions of CASIA. It is clear from these tables that MIC-PCFS consistently outperforms P-CCFS and SFFS on estimation irrespective of semantic primitives or emotion dimensions. Such findings again resonates with the aforementioned fact that redundancies among features need to be removed after defining the relevant acoustic features. Moreover, compared with the algorithm's performance with Berlin Emo-DB, the results with CASIA were comparatively lower. This is because Berlin Emo-DB contains emotional utterances that are more typical and clear from actors or actresses, while CASIA contains emotional utterances that are more spontaneous, and do not sufficiently simulate emotions in a natural or clear manner. Another reason might that the Berlin Emo-DB has a large number of speakers (ten actors), which might improve the generalization to different speakers, however, CASIA has a small number of speakers (four actors), and therefore, limiting the generalization of the achieved performance.

Overall, the three-layer model implemented on the basis of MIC-PCFS improved the estimation performance on valence and arousal, resulting in both higher CC and smaller MAE values, compared with the performances obtained by other feature selection approaches and presenting comparable results to human evaluators.

V. Conclusion

This study presented a novel acoustic feature selection algorithm based on the maximal information coefficient and an adapted concept of predominant correlation and demonstrated that it is efficient to implement a three-layer model for speech emotion recognition. Promising results on estimation of emotion dimensions are reported over the emotional corpora of Berlin Emo-DB and CASIA, and outperformed those achieved by related feature selection algorithms. This performance can advance the SER accuracies. The main advantage of the proposed algorithm is the fact that the maximal information coefficient could capture a wide range of associations between a feature and target and is well suited to human vague emotion perception which may exist both functional and non-functional percepts. Besides, the concept of an adopted predominant correlation does a great contribution to eliminate the redundancy among relevant features to improve the accuracy of emotion estimation.

Acknowledgment

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

References

[1] M. Grimm, K. Kroschel, and S. Narayanan, "Modeling emotion expression and perception behavior in auditive emotion evaluation," *Proc. ISCA 3rd Internat. Conf. on Speech Prosody, Dresden, Germany,* pp. 9-12, 2006.

[2] J. Ma, H. Jin, L. Yang, and J. Tsai, "Ubiquitous Intelligence and Computing: Third International Conference, UIC 2006, Wuhan, China, September 3-6, 2006," *Proc. (Lecture Notes in Computer Science), Springer-Verlag New York, Inc., Secaucus, NJ,* USA, 2006.

[3] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *Proc. ICASSP 2004,* vol. 1, pp. 577-580, 2004.

[4] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific,* pp. 1-10, 2006.

[5] S. Yildirim et al., "An acoustic study of emotions expressed in speech," *Proc. ICSLP, Jeju Island, Korea, October, 2004* pp. 2193-2196, 2004.

[6] D. Ververidis, and C. Kotropoulos, " Automatic speech classification to five emotional states based on gender information," *Proc. EUSIPCO, 2004* pp. 341-344, 2004.

[7] J.A. Russell, " A circumplex model of affect," *J. of personality and social psychology* vol. 39, pp. 1161-1178, 1980.

[8] H. Gunes et al., "Emotion representation, analysis and synthesis in continuous space: A survey," *Proc. Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011* pp. 827-834, 2011.

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

[10] D. Wu, T.D. Parsons, and S.S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," *Proc. Eleventh Annual Conference of the International Speech Communication Association,2010* pp. 785-788, 2010.

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

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

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

[14] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," *Proc. Eighteenth International Conference on Machine Learning, 2001* pp. 74-81, 2001.

[15] P. Langley, "Selection of relevant features in machine learning," *Proc. the AAAI Fall symposium on relevance, 1994* vol. 184, pp. 245-271, 1994.

[16] Y. Saeys, I. Inza, and P. Larraaga "A review of feature selection techniques in bioinformatics," *J. of Bioinformatics, 2007* vol. 23, no. 19, pp. 2507-2517, 2007.

[17] J. Rong, G. Li, and Y.P.P. Chen " Acoustic feature selection for automatic emotion recognition from speech," *J. of Information processing and management, 2009* vol. 45, no. 3, pp. 315-328, 2009.

[18] R. Kohavi, and G. John "Wrappers for feature subset selection," *J. Artificial intelligence, 1997* vol. 97, pp. 273-324, 1997.

[19] M. Hall, " Correlation-based feature selection for discrete and numeric class machine learning," *Proc. the Seventeenth International Conference on Machine Learning, 2000* pp. 359-366, 2000.

[20] D. Reshef et al., " Detecting novel associations in large data sets," *J. Science, 2011* vol. 340, no. 6062, pp. 1518-1524, 2011.

[21] L. Yu, and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," *Proc. of the 20th international conference on machine learning, 2003* pp. 856-863, 2003.

[22] F. Eyben, M. Wollmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," *Proc. of ACM Multimedia (MM), Florence, Italy, 2010* pp. 14591462, 2010.

[23] A.V. Ivanov, G. Riccardi, A.J. Sporka, and J. Franc "Recognition of personality traits from human spoken conversations," *Proc. of Twelfth Annual Conference of the International Speech Communication Association, 2011* pp. 15491552, 2011.

[24] A. Tickle, "English and Japanese speakers emotion vocalizations and recognition: A comparison highlighting vowel quality," *Proc. of ISCA Workshop on Speech and Emotion. Citeseer, 2000* 2000.

[25] R. Huang, and C. Ma "Toward a speaker-independent real-time affect detection system," *Proc. of Pattern Recognition, 2006, ICPR 2006* vol. 1, pp. 12041207, 2006.

[26] K. Kira, and L. Rendell "The feature selection problem: Traditional methods and a new algorithm," *Proc. of the Tenth National Conference on Artificial Intelligence, 1992* pp. 129-134, 1992.

[27] H. Peng, F. Long, and C. Ding "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *J. IEEE Transactions on pattern analysis and machine intelligence,2005* vol. 27, no. 8, pp. 1226-1238, 2005.

[28] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C.S. Cruz "Quadratic programming feature selection," *J. Machine Learning Research, 2010* pp. 1491-1516, 2010.

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