2018 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2018, AALBORG, DENMARK

SPEECH EMOTION RECOGNITION USING CYCLOSTATIONARY SPECTRAL ANALYSIS

Amin Jalili[†]*, Sadid Sahami*[‡]*, Chong-Yung Chi*[†] *and Rassoul Amirfattahi*[‡]

[†]Institute of Communications, National Tsing Hua University, Hsinchu, Taiwan
[‡]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
amin.jalili@ieee.org, s.sahami@ieee.org, cychi@ee.nthu.edu.tw, fattahi@cc.iut.ac.ir

ABSTRACT

Inspired by the modulated and non-stationary nature of speech signals, this paper proposes a new feature extraction scheme for speech emotion recognition (SER) using cyclostationary spectral analysis (CSA). This spectral analysis discloses the underlying *first-order* and *second-order (hidden) periodicities* in emotional speech signals using the estimated spectral correlation function (SCF) via FAM algorithm. Experiments on the Berlin database of emotional speech (EmoDB) show that the proposed scheme using cyclostationary spectral features (CSFs) significantly outperforms state-of-the-art methods in terms of recognition accuracy.

Index Terms— speech emotion recognition, artificial intelligence, cyclostationarity, FAM algorithm, cyclic spectral analysis, human machine interaction.

1. INTRODUCTION

Have we ever imagined an intelligent humanoid robot can feel our emotion and react to us appropriately? This may be in the realms of fantasy decades ago, but now it seems quite achievable. Roughly speaking, the current technology is mainly enriched by cognitive intelligence, however, it is predictable that the future generation of artificial intelligence will be equipped with emotional intelligence. The primary approach for detecting emotions has long been the facial recognition. However, speech emotion recognition (SER) has become a trend in behavioral/speech signal processing and artificial intelligence. SER attempts to recognize the underlying emotional state of a speaker from the speech signal and has demonstrated its effectiveness in human-machine interactions [1, 2]. In a recent review of tracing of SER in the past years, Schulller [3] noted that despite significant advances of SER, until now only few commercial products of automatic emotion recognition had the chance to get a place in the market for widely-spreading daily life usages.

The two major categories for feature extraction in SER are *prosodic* and *spectral* features. The former is broadly studied as the most commonly used type of SER features [4, 5]. On the other hand, the latter conveys frequency information of the speech signal and plays an important role in SER such as linear predictor cepstral coefficients (LPCC) [6] and Mel-frequency cepstral coefficients (MFCC) [7]. Moreover, [8] proposed a model using statistical features of Fourier representation over frames of the emotional speech signals for SER. One drawback behind [8] is to consider a subset of Berlin database of emotional speech (EmoDB) [9] and hence it lacks the completeness of analyzing the whole database. Recently, [10] presented a biologically inspired method for SER which operates directly on the speech signal without the block of feature extraction, however, its recognition accuracy for four out of seven emotions is just around 75% over EmoDB. It is also noteworthy that the *stationarity* of speech signals in short time segments is a cornerstone for many research works focusing on spectral features. Despite many insightful research works for SER, there are still serious challenges on this complicated task [11–14].

This paper proposes a *new stochastic and statistical signal processing approach toward SER*. Our philosophy is to pay particular attention to the non-stationary nature of emotional speech signals for *automatic affective information* recognition rather than focusing on the assumption of stationarity of this type of signals in short time intervals. Accordingly, a new feature extraction scheme for affective information recognition using cyclostationary spectral analysis (CSA) is proposed. To this end, the spectral correlation function (SCF) will be estimated by the fast Fourier transform accumulation method (FAM) [15] which is a computationally efficient algorithm and this attempts to uncover the *first-order* (corresponding to degenerate cyclic frequency) and *second-order periodicities* (nonzero cyclic frequencies) [16] buried in the signal by analyzing the quadratic form of the signal. To the best of our knowledge, this is the first research work that interprets an emotional speech signal as a cyclostationary signal.

The rest of this paper is organized as follows. Section 2 presents the track of cyclostationarity in speech signals. Section 3 details the proposed scheme based on the CSA. Section 4 provides the experimental results using the EmoDB. Finally, concluding remarks are given in Section 5.

This work was supported by the Ministry of Science and Technology, R.O.C., under Grant MOST 105-2221-E-007-020-MY2.

978-1-5386-5477-4/18/$31.00

2. CYCLOSTATIONARITY IN SPEECH SIGNALS

2.1. Background of cyclostationary analysis

There is a special type of non-stationary stochastic processes when their statistical properties vary periodically with time called *cyclostationary* processes [17]. An important subclass of cyclostationary processes is the one that exhibits cyclostationarity in its mean and autocorrelation function which is called *wide-sense cyclostationary* stochastic process. To begin with, let us consider a wide-sense cyclostationary process $\{x(n;w) \in \mathbb{R} \mid n \in \mathbb{Z}, w \in \Omega\}$ with period $N_0 \in \mathbb{Z}$ where Ω is the sample space, $\mathbb{R}$ and $\mathbb{Z}$ are the sets of real and integer numbers, respectively. For this process we have $\mathbb{E}\{x(n+N_0;w)\} = \mathbb{E}\{x(n;w)\}$ and $R_x(n+N_0,\ell) = R_x(n,\ell)$ where $R_x(n,\ell) \stackrel{\text{def}}{=\!=} \mathbb{E}\{x(n;w)x(n-\ell;w)\}$ is the autocorrelation function, $\mathbb{E}\{\cdot\}$ accounts for statistical expectation and $\ell \in \mathbb{Z}$.

As the autocorrelation function is periodic in the time domain, it can be expressed using the Fourier series representation as [17]

$$R_x(n,\ell) = \sum_{\alpha \in \mathcal{A}} r_x^{\alpha}(\ell) e^{\mathrm{i}2\pi\alpha n}, \tag{1}$$

where $\mathrm{i} = \sqrt{-1}$ and $\mathcal{A}$ is a finite set of cyclic frequencies and

$$r_x^{\alpha}(\ell) \stackrel{\text{def}}{=\!=} \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n+\ell)x(n)e^{-\mathrm{i}2\pi\alpha n}, \tag{2}$$

is called the discrete-time *cyclic autocorrelation function* [17–19] at cyclic frequency α provided that $x(n;w)$ is a cycloergodic process and $2N+1$ is the truncation length.

Furthermore, the cyclostationary process can be characterized in the frequency domain using SCF which can be obtained from *cyclic Wiener relation* [20] as

$$S_x^{\alpha}(f) \stackrel{\text{def}}{=\!=} \sum_{\ell=-\infty}^{\infty} r_x^{\alpha}(\ell) e^{-\mathrm{i}2\pi f\ell}, \tag{3}$$

where f is the cross spectrum frequency and SCF is the Fourier transform of the cyclic autocorrelation function on a bifrequency plane (α, f).

In practice, a single realization of the finite length of this stochastic process is available and it is characterized as the *cyclostationary signal*. Let $x[n], 0 \le n \le N-1$ be a cyclostationary signal with sampling interval of $T_s = 1/f_s$. In this paper we use the FAM algorithm [15] which is based on the time-smoothed cyclic periodogram. First, the signal is windowed and for that the Hamming window $T = N'T_s$, denoted as $w(p)$, $-N'/2 \le p \le N'/2-1$, is hopped over the signal in blocks of L samples which leads to $P = N/L$ segments. Then the complex demodulate (short-time Fourier transform with N'-point FFT) of these windowed segments can be obtained as [15]

$$X_T(n,f) = \sum_{p} w(p)x(n-p)e^{-\mathrm{i}2\pi f(n-p)/N'}, \tag{4}$$

where n is the center of a time segment. By inherent properties of this algorithm, it estimates SCF over diamond shape regions of support, which are called channel-pairs and the pattern of tiled bifrequency plane is depicted in Fig. 1 for $N' = 8$. The center of each channel-pair region is characterized by an ordered pair as

$$(j,i) \stackrel{\text{def}}{=\!=} (\frac{l+l'}{2}, l-l'), \quad l,l' = -N'/2, ..., N'/2-1. \tag{5}$$

Then, the SCF is estimated over each channel-pair region for P-point estimation (indexed by q) and separated by $\Delta\alpha = f_s/N$ (*cyclic frequency resolution*) as

$$S_{x_T}^{\alpha_{i,q}}(f_j) = \sum_{r=0}^{P-1} X_T(rL, f_l)X_T^*(rL, f_{l'})g(r)e^{-\frac{\mathrm{i}2\pi rq}{P}}, \tag{6}$$

where $g(n)$ is the data tapering Hamming window of length P, $\alpha_{i,q} \stackrel{\text{def}}{=\!=} \hat{\alpha}_i + q\Delta\alpha$ and f_l (and $f_{l'}$) can be obtained as

$$f_l \stackrel{\text{def}}{=\!=} l(f_s/N'), \quad l = -N'/2, ..., N'/2-1. \tag{7}$$

is the discretized frequencies of each complex demodulate and the bifrequency coordinate associated with the center of each channel-pair region is determined as

$$(f_j, \hat{\alpha}_i) \stackrel{\text{def}}{=\!=} (f_{\frac{l+l'}{2}}, \alpha_{l-l'}), \quad \text{(cf. (5))}, \tag{8}$$

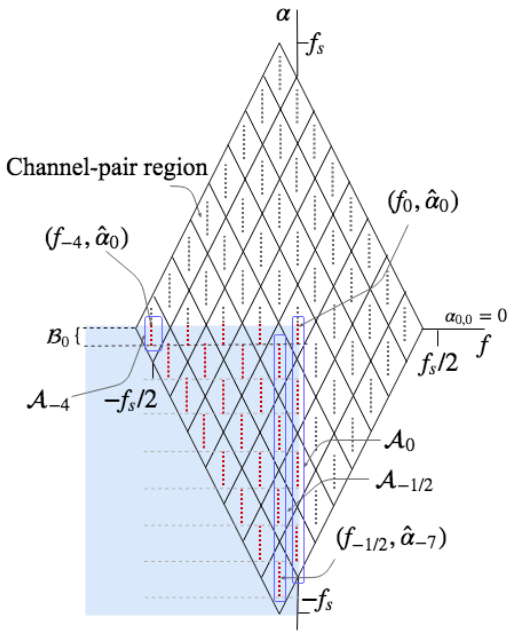

Fig. 1: Tiling of the bifrequency plane coressponding to the location of estimated SCF channel pair regions

where the coordinate of channel-pair region is denoted by indicies of frequencies. In fact, estimated SCF is the Fourier transformation of products of the complex demodulates, $X_T(rL, f_l)X_T(rL, f_{l'})$, with P-point FFT. Besides, it is noted in [15] that due to minimizing the variability of SCF estimation near the top and bottom of each channel-pair region, the point estimates within the range of $-Q/2 \leq q \leq Q/2-1$ where $Q = PL/N'$ must be retained (cf. Fig. 1).

Moreover, the physical interpretation behind $S^{\alpha}_{x_T}(f)$ is the correlation of spectral components of the signal $x(n)$ over time span of NT_s with *frequency resolution* of $1/T$. Besides, it is noted in [15], $L = N'/4$ is a good choice for practical use. For more details about the FAM algorithm, one can refer to [15]. Moreover, without loss of generality it can be assumed that $f_s = 1$ and then (f, α) will contain normalized frequencies.

2.2. Tracking Cyclostationarity in speech signals

The modulation theory presented in [21] assumes speech signal to be the modulated output of different components such as affective control, speech gesture, and vocal carrier signal. Due to such model, it can be assumed that affective information modulates the vocal carriers in speech signal. Consistent with the *Napolitano's* definition for cyclostationary process, which is *the modulation of a periodic process by a random one* [22], affective information can be extracted using CSA. Consequently, the underlying second-order (hidden) periodicity, can be uncovered by analyzing the quadratic form of the signal. The estimated SCF of a frame of speech signal from EmoDB database is shown in Fig. 2. The estimated SCF shows noticeable spectral lines ($|S^{\alpha}_x(f)| \neq 0$) for some $\alpha \neq 0$. This implies a cyclostationarity (of second order) behavior of the emotional speech signal to some extent. However, there are very few papers in which the cyclostationary analysis for speech signals are discussed. Recently [23] proposes a new approach to track the fundamental frequency using the cyclostationary analysis on the intrinsic mode functions of the target vowel for automatic speech recognition. It is also interestingly specified in [23] that the fundamental frequency in speech signal is equivalent to the cyclic frequency in CSA.

Now, the critical question that arises here is: *can this trace of cyclostationarity in speech signals lead to an effective feature extraction scheme for emotion recognition?* In the following sections, we will answer this question.

3. PROPOSED METHOD

As the preprocessing step, to preserve the information from variations in an utterance due to the dynamic nature of the speech signal, it is necessary to partition the speech signal into K overlapped frames of the length τ ms with $\Delta\tau$ ms step size and let $x_k[n]$ be the k-th frame of the speech signal. Then, it is beneficial to multiply the speech segment by a Hamming window to tone down the abrupt changes at the edges.

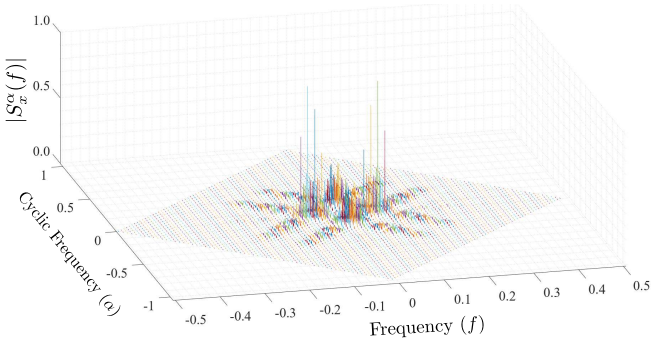

Fig. 2: Estimated SCF of a frame of a sample speech signal from EmoDB

3.1. Feature extraction scheme

In this section, our objective is to take advantage of the *statistical properties* of the channels in the bifrequency plane of the estimated SCF (as seen in Fig. 2) for feature extraction. To do so, for each windowed frame of an utterance, the SCF is estimated using FAM algorithm which computes the SCF in diamond-shape channel pair regions on the bifrequency plane. Moreover, it is notable that for an N'-point channelizer in FAM algorithm, due to the symmetry in SCF, it is enough to estimate only ${N'}^2/4$ channels (the highlighted quadrant in Fig. 1).

3.1.1. Features from low cyclic frequency band

Here, we aim to derive features from spectral lines corresponding to the *low cyclic frequency band* $\mathcal{B}_0$ (cf. Fig. 1 and it is the set of cyclic frequencies close to the degenerate cyclic frequency $\alpha_{0,0}$). Let us define

$$\mathcal{T}_{k,h} \stackrel{\text{def}}{=} \{\frac{\sum_j |S^{\alpha_{0,h}}_{x_k}(f_j)|}{N'+1} : -N'/2 \leq j \leq 0\}, \qquad (9)$$

where $-Q/2 \leq h \leq 0$. It is needed first to normalize this set and we have $\overline{\mathcal{T}}_{k,h} = \mathcal{T}_{k,h}/\eta_k$ where $\eta_k \stackrel{\text{def}}{=} \max_{-Q/2 \leq h \leq 0} \mathcal{T}_{k,h}$. Then the block of features are defined as

$$\boldsymbol{\psi}_u \stackrel{\text{def}}{=} [\psi_u(\overline{\mathcal{T}}_h)]^0_{h=-Q/2},\ u = 1, \ldots, 4, \qquad (10)$$

where the features $\psi_1(\overline{\mathcal{T}}_h)$, $\psi_2(\overline{\mathcal{T}}_h)$, $\psi_3(\overline{\mathcal{T}}_h)$ and $\psi_4(\overline{\mathcal{T}}_h)$ are the mean, minimum, standard deviation and median of $\overline{\mathcal{T}}_h$, respectively.

3.1.2. Features from dynamics of cyclic frequencies

First, for the $x_k[n]$, it needs to define the set corresponding to the magnitude of spectral lines for the cyclic frequency

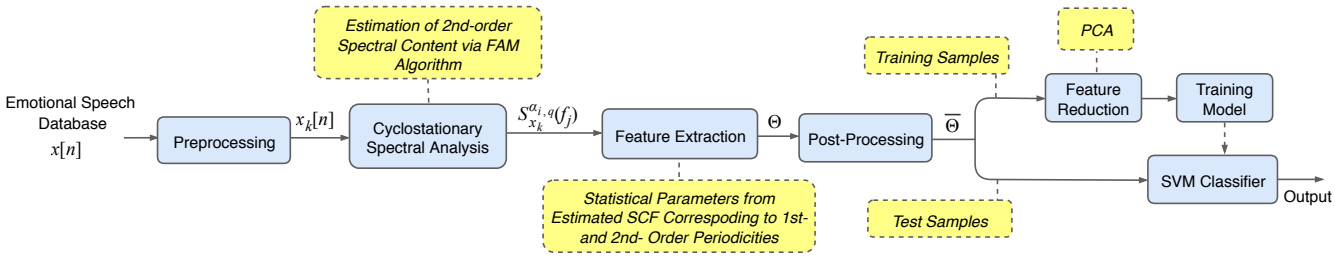

Fig. 3: SER system using the proposed CSA based feature extraction scheme

bands of the highlighted quadrant (cf. Fig. 1) and we have $\mathcal{S}_k \stackrel{\text{def}}{=} \cup_{j=-N'/2}^{0} \mathcal{S}_{k,j}$ where

$$\mathcal{S}_{k,j} \stackrel{\text{def}}{=} \{|S_{x_k}^{\alpha_{i,q}}(f_j)| : \alpha_{i,q} \in \mathcal{A}_j\}, \tag{11}$$

and f_j is defined in (8),

$$\mathcal{A}_j \subset \mathcal{A} = \{m/N : m = 0, 1, \ldots, \frac{(2N'-1)Q}{2}\},$$

which is the unified cyclic frequency set shown in Fig. 1 containing cyclic frequencies corresponding to the frequency f_j. For instance in Fig. 1, $\mathcal{A}_0$ is the set contains all the point estimates depicted by red colors. Due to the high dynamic range of SCF, first, it is needed to normalize it with respect to its maximum as $\overline{\mathcal{S}}_{k,j} = \mathcal{S}_{k,j}/\rho_{k,j}$ where j is defined in (5) and

$$\rho_{k,j} \stackrel{\text{def}}{=} \max_{\alpha_{i,q} \in \mathcal{A}_j} |S_{x_k}^{\alpha_{i,q}}(f_j)|.$$

The corresponding feature space includes six blocks of measures denoted by $\boldsymbol{\varphi}_v \stackrel{\text{def}}{=} [\varphi_v(\overline{\mathcal{S}}_{k,j})]_{j=-N'/2}^{0}$, $v = 1, \ldots, 6$ where each component is denoted as

$$\varphi_v(\overline{\mathcal{S}}_{k,j}) = \frac{1}{K} \sum_{k=1}^{K} \varphi_{v,k}(\overline{\mathcal{S}}_{k,j}), \tag{12}$$

and the features $\varphi_{v,k}(\overline{\mathcal{S}}_{k,j})$ are defined as follows. Here, $\varphi_{1,k}(\overline{\mathcal{S}}_{k,j})$ is the standard deviation of $\overline{\mathcal{S}}_{k,j}$. The second measure $\varphi_{2,k}(\overline{\mathcal{S}}_{k,j})$ is the skewness of $\overline{\mathcal{S}}_{k,j}$ which is the amount and direction of asymmetry over $\mathcal{A}_j$. The next measure $\varphi_{3,k}(\overline{\mathcal{S}}_{k,j})$ is the *kurtosis* of $\overline{\mathcal{S}}_{k,j}$, as the fourth standardized moment, representing sharpness of the central peak over $\mathcal{A}_j$. The fourth measure is the *spectral flatness* of $\overline{\mathcal{S}}_{k,j}$ and is defined as

$$\varphi_{4,k}(\overline{\mathcal{S}}_{k,j}) \stackrel{\text{def}}{=} \frac{\left(\prod_{\alpha_{i,q} \in \mathcal{A}_j} |S_{x_k}^{\alpha_{i,q}}(f_j)|/\rho_{k,j}\right)^{1/n(\mathcal{A}_j)}}{\sum_{\alpha_{i,q} \in \mathcal{A}_j} |S_{x_k}^{\alpha_{i,q}}(f_j)|/\left(\rho_{k,j} n(\mathcal{A}_j)\right)}, \tag{13}$$

where $n(\cdot)$ is the cardinality of corresponding set. Another block of feature is the *modified spectral centroid* defined as

$$\varphi_{5,k}(\overline{\mathcal{S}}_{k,j}) \stackrel{\text{def}}{=} \frac{\sum_{\alpha_{i,q} \in \mathcal{A}_j} (|f_j| + |\alpha_{i,q}|)|S_{x_k}^{\alpha_{i,q}}(f_j)|}{\sum_{\alpha_{i,q} \in \mathcal{A}_j} |S_{x_k}^{\alpha_{i,q}}(f_j)|}. \tag{14}$$

Algorithm 1: Our proposed scheme

input : $x[n] \in$ *EmoDB*, $n = 1, \ldots, N$
output: $\overline{\boldsymbol{\Theta}}$, the standardized feature vector
begin initialization
  set N', τ, and $\Delta\tau$
  $L \leftarrow N'/4$, $K \leftarrow \left\lceil \frac{N - \lfloor \tau f_s \rfloor}{\lfloor \Delta\tau f_s \rfloor} \right\rceil$, $P \leftarrow N/L$
end
for *frame* $k \in \{1 \rightarrow K\}$ **do**
  Compute k-th frame, x_k, using $\tau, \Delta\tau$
  Compute $|S_{x_k}^{\alpha_{i,q}}(f_j)|$ (Eq. (6))
  Compute $\frac{\sum_{j=-N'/2}^{0} |S_{x_k}^{\alpha_{0,h}}(f_j)|}{(N'+1)}$ (Eq. (9))
  Compute $\varphi_{v,k}(\overline{\mathcal{S}}_{k,j})|_{v=1}^{6}$ (Eq. (12))
end
$\boldsymbol{\psi}_u = [\psi_u(\overline{\mathcal{T}}_h)]_{h=-Q/2}^{0}$
$\boldsymbol{\varphi}_v = [\varphi_v(\overline{\mathcal{S}}_{k,j})]_{j=-N'/2}^{0}$
$\boldsymbol{\Theta} = [\boldsymbol{\psi}_u, \boldsymbol{\varphi}_v]$, $1 \le u \le 4$, $1 \le v \le 6$
$\overline{\boldsymbol{\Theta}} = [\overline{\vartheta}_i]_{i=1}^{M}$ (Eq. (15))

Table 1: Confusion table for the proposed CSFs using PCA

Emotion	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	CCR (%)
Anger	**124**	0	1	1	1	0	0	97.6
Boredom	0	**76**	0	0	0	0	5	93.8
Disgust	2	0	**40**	1	0	0	3	87.0
Fear	5	1	0	**61**	0	0	2	88.4
Happiness	22	0	0	1	**47**	1	0	66.2
Sadness	0	2	0	0	0	**59**	1	95.2
Neutral	0	9	1	0	0	1	**68**	86.1
Percision (%)	81.0	86.4	95.2	95.3	97.9	96.7	86.1	88.8

The next measure $\varphi_{6,k}(\overline{\mathcal{S}}_{k,j})$ is the *trimmed mean* of the $\overline{\mathcal{S}}_{k,j}$ which is defined as the difference of the mean of upper-half of $\overline{\mathcal{S}}_{k,j}$ and the mean of lower-half of this set for each unified-channel[1]. Now, the complete raw feature space can be constructed by concatenation of the extracted

[1]Here, the lower-half/upper-half of a data set is defined as the set of all values of the ascending ordered data set that are less/greater than the mean value.

Table 2: Recognition results in terms of weighted accuracy for the CSF and various state-of-the-art methods

Feature Scheme	CCR (%)							Weighted Accuracy (%)	
	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	Achievable	Average[†]
Wu: PROS [11]	87.4	82.7	78.3	79.7	49.3	83.9	82.3	78.7	*NA*
Wu: MSF [11]	91.3	86.4	78.3	71.0	60.6	88.7	83.5	81.3	*NA*
Zhang [14]	90.8	78.8	84.8	81.1	65.8	88.7	75.9	79.6	*NA*
Bhargava [13]	83.5	84.8	84.1	80.9	60.6	88.7	73.9	80.3	*NA*
Tawari [12]	93.7	92.6	80.4	76.8	57.7	91.9	**88.6**	84.5	*NA*
CSF[*]	93.7	90.1	84.8	87.0	**67.6**	93.6	87.3	87.1 ▲	85.0 ± 0.6
CSF[**]	**97.6**	**93.8**	**87.0**	**88.4**	66.2	**95.2**	86.1	**88.8** ▲	86.1 ± 0.4

[†]Over 400 Monte Carlo runs; *NA*[‡]: Not available; [*]Without feature reduction; [**]Using PCA

features from spectral information corresponding to the first and second-order periodicities consequently and we have $\boldsymbol{\Theta} \stackrel{\text{def}}{=} [\boldsymbol{\psi}_u, \boldsymbol{\varphi}_v], 1 \leq u \leq 4,\ 1 \leq v \leq 6$ (the overall number of features is $M = 4(Q/2+1) + 6(N'+1)$). Finally, it is essential to normalize the feature matrix and we use the z-score technique which makes the values of each feature have zero-mean ($\mu = 0$) and unit-variance ($\sigma^2 = 1$); the standardized feature is denoted by $\overline{\boldsymbol{\Theta}} = [\overline{\boldsymbol{\vartheta}}_i]_{i=1}^{M}$ where

$$\overline{\boldsymbol{\vartheta}}_i = \frac{\boldsymbol{\vartheta}_i - \mu \mathbf{1}}{\sigma}, \tag{15}$$

where $\boldsymbol{\vartheta}_i$ is the i-th column of $\boldsymbol{\Theta}$ and $\mathbf{1}$ is an all-one vector.

Feature reduction is an important step in machine learning applications in order to avoid challenges due to the curse of dimensionality and further decreases the computational cost and complexity for the ensuing classification task [24]. To this end, we use the cross validated *principal component analysis* (PCA) [24]. Here, our objective is to investigate the effectiveness of our CSFs rather than using advanced techniques for feature reduction which often lead to higher classification rate. Fig. 3 shows our proposed SER system using CSA. Moreover, the proposed feature extraction scheme is summarized in Algorithm 1 which provides all the features $\overline{\boldsymbol{\Theta}}$ for training and testing.

4. EXPERIMENTAL RESULTS

The EmoDB [9] is a publicly available database which consists of utterances expressed by ten German actors for ten sentences and six basic emotions as well as neutral speech including anger (127), boredom (81), disgust (46), fear (69), happiness (71), sadness (62) and neutral (79). Our experiments are based on all the 535 available utterances in the EmoDB.

In our scheme, we set the window length for framing of the speech signal to $\tau = 150$ ms with $\Delta\tau = 15$ ms step size. For SCF estimation, we set $N' = 64$, $L = 16$, and finally 522 features are extracted. For classification, we use the support vector machine (SVM) [24] with radial basis kernel of MATLAB 2016b for training and testing the data. Moreover, all of our results in this paper are based on the stratified 10-fold cross validation[2]. The confusion table for the proposed CSFs resulting from the PCA method for 30 features is shown in Table 1, where the correct classification rate (CCR) column lists the correct recognition rate per class. Clearly, the CCR for emotions *Anger*, *Boredom* and *Sadness* are quite satisfactory.

Table 2 presents the CCR of the proposed CSFs per emotion and various state-of-the-art methods. All those methods have been tested on the EmoDB using SVM with radial basis kernel and stratified 10-fold cross validation. Wu [11] reaches up to 78.7% and 81.3% weighted accuracy[3] for prosodic (PROS) and modulation spectral features (MSF), respectively. Besides, Zhang [14] achieved 79.6% weighted accuracy using an enhanced kernel isomap. Bhargava [13] achieved 80.3% of the weighted accuracy using rhythm and temporal feature. Tawari [12] achieved 84.5% weighted accuracy by proposing a new set of features based on cepstrum analysis of pitch and intensity contours. However, it is mentioned in [12] that the feature selection is prior to the training and testing phase in the system model. This violates the cross validation procedure and can mislead to subsequent increase in the recognition accuracy. Whereas in our system depicted in Fig. 3, the test set is kept *completely unseen* from the training including the feature reduction stage. The proposed CSFs reaches up to 87.1% weighted accuracy even without any feature reduction. Finally, our scheme reaches up to 88.8% weighted accuracy for seven emotions using cross-validated PCA with only 30 selected features. Due to that training and testing sets for randomized cross validation can be different in each run, the accuracy would have some tolerances and hence doing Monte Carlo runs is necessary. The obtained average of weighted recognition accuracy of our scheme over 400 Monte Carlo simulation runs is 86.1 ± 0.4 which is superior to the achievable accuracy of all the other methods (such Monte Carlo simulation averages are not available for any of the other methods). Clearly, our method significantly outperforms the other methods in terms of recognition accuracy.

[2]The MATLAB code of this research work is publicly available at: http://www1.ee.nthu.edu.tw/cychi/links.php.

[3]The weighted accuracy is the ratio of the total number of correctly recognized samples to the total number of samples.

5. CONCLUSION

We have presented a new feature extraction scheme for SER using CSA. This includes features from spectral content related to the low cyclic frequency band and features from spectral information corresponding to the dynamics of cyclic frequencies in emotional speech signals. The SCF of segmented speech signal are computed using FAM algorithm and then blocks of features are extracted using statistical parameters of magnitudes of the spectral lines over the bifrequency plane. Experimental results have demonstrated that this scheme significantly boosts the emotion recognition accuracy and meanwhile open up new avenues for further research in SER. Our future work will be focused on further analysis and experiments in broader datasets and noisy environments.

6. REFERENCES

[1] 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.

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

[3] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[4] S. Gharsellaoui, S. A. Selouani, and A. O. Dahmane, "Automatic emotion recognition using auditory and prosodic indicative features," in *Proc. 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*, May 2015, pp. 1265–1270.

[5] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[6] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[8] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affective Computing*, vol. 6, no. 1, pp. 69–75, Jan 2015.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *INTERSPEECH*, 2005, pp. 1517–1520.

[10] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5135–5139.

[11] 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.

[12] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 502–509, Oct. 2010.

[13] M. Bhargava and T. Polzehl, "Improving automatic emotion recognition from speech using rhythm and temporal feature," *CoRR*, vol. abs/1303.1761, 2013.

[14] S. Zhang, X. Zhao, and B. Lei, "Speech emotion recognition using an enhanced kernel isomap for human-robot interaction," *International Journal of Advanced Robotic Systems*, vol. 10, no. 2, pp. 1–7, 2013.

[15] R. S. Roberts, W. A. Brown, and H. H. Loomis, "Computationally efficient algorithms for cyclic spectral analysis," *IEEE Signal Processing Magazine*, vol. 8, no. 2, pp. 38–49, 1991.

[16] W. A. Gardner, "The spectral correlation theory of cyclostationary time-series," *Signal Processing*, vol. 11, no. 1, pp. 13 – 36, 1986.

[17] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: Half a century of research," *Signal Processing*, vol. 86, no. 4, pp. 639 – 697, 2006.

[18] W. A. Gardner, "Exploitation of spectral redundancy in cyclostationary signals," *IEEE Signal Processing Magazine*, vol. 8, no. 2, pp. 14–36, April 1991.

[19] A. Napolitano, *Generalizations of Cyclostationary Signal Processing: Spectral Analysis and Applications*, Piscataway, NJ: Wiley–IEEE Press, 2012.

[20] W. A. Gardner, *Statistical Spectral Analysis: A Nonprobabilistic Theory*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1986.

[21] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, no. 1-3, pp. 170–183, 1994.

[22] A. Napolitano, "Cyclostationarity: New trends and applications," *Signal Processing*, vol. 120, pp. 385–408, 2016.

[23] B. Dong, "Characterizing resonant component in speech: A different view of tracking fundamental frequency," *Mechanical Systems and Signal Processing*, vol. 88, pp. 318–333, 2017.

[24] M. B. Christopher, *Pattern Recognition and Machine Learning*, Springer-Verlag, New York, 2016.