Exploitation of Phase-Based Features for Whispered Speech Emotion Recognition

JUN DENG[1], XINZHOU XU[2,3], ZIXING ZHANG[1], (Member, IEEE), SASCHA FRÜHHOLZ[4,5,6], AND BJÖRN SCHULLER[1,7], (Senior Member, IEEE)

[1]Chair of Complex and Intelligent Systems, University of Passau, Passau 94032, Germany
[2]Machine Intelligence and Signal Processing Group, Mensch–Maschine–Kommunikation, Technische Universität München, Munich 80333, Germany
[3]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210018, China
[4]Institute of Psychology, University of Zurich, Zürich 8006, Switzerland
[5]Neuroscience Center Zurich, University of Zurich and ETH Zürich, Zürich 8092, Switzerland
[6]Zurich Center for Integrative Human Physiology, University of Zurich, Zürich 8006, Switzerland
[7]Department of Computing, Imperial College London, London SW7 2AZ, U.K.

Corresponding author: J. Deng (jun.deng@uni-passau.de)

This work was supported in part by the Bundesministerium für Bildung und Forschung within IKT 2020 through the Emotionssensitives Assistenzsystem for Mensches met Behinderungen Project under Grant 16SV7213, in part by the European Communitys Seventh Framework Programme within the European Research Council Starting Grant through the iHEARu Project under Grant 338164, and in part by the European Union's Horizon 2020 Programme within the Research and Innovation Action through the Multi-Modal Human-Robot Interaction for Teaching and Expanding Social Imagination in Autistic Children Project under Grant 688835. The work of X. Xu was supported by the Research Grants through the China Scholarship Council. The work of S. Frühholz was supported by the Swiss National Science Foundation under Grant PP00P1 157409/1.

ABSTRACT Features for speech emotion recognition are usually dominated by the spectral magnitude information while they ignore the use of the phase spectrum because of the difficulty of properly interpreting it. Motivated by recent successes of phase-based features for speech processing, this paper investigates the effectiveness of phase information for whispered speech emotion recognition. We select two types of phase-based features (i. e., modified group delay features and all-pole group delay features), both which have shown wide applicability to all sorts of different speech analysis and are now studied in whispered speech emotion recognition. When exploiting these features, we propose a new speech emotion recognition framework, employing outer product in combination with power and L2 normalization. The according technique encodes any variable length sequence of the phase-based features into a fixed dimension vector regardless of the length of the input sequence. The resulting representation is fed to train a classification model with a linear kernel classifier. Experimental results on the Geneva Whispered Emotion Corpus database, including normal and whispered phonation, demonstrate the effectiveness of the proposed method when compared with other modern systems. It is also shown that, combining phase information with magnitude information could significantly improve performance over the common systems solely adopting magnitude information.

INDEX TERMS Phase-based features, whispered speech emotion recognition, outer product.

I. INTRODUCTION

Speech emotion recognition is devoted to increase the user-friendliness and provide a more natural interaction experience for human-computer interaction and computer-mediated human communication [1]–[5]. A large number of efforts have been made to predict accurate emotional states from 'normal' speech [6]–[10]. Besides normal speech, in fact, whispered speech that is produced by speaking with high breathiness and no periodic excitation is another common speech mode. In our daily life, whispered speech appears when we want to intentionally confine the hearing of speech to listeners who are nearby. For example, people often whisper to the user interface of a smartphone to send privacy information in terms of credit card information and date of birth, etc. In addition, communication by whispered speech is of vital importance for patients with disabilities who are affected by disease of the vocal system such as functional aphonia [11] or laryngeal disorders [12]. In spite of the important role of whispered speech, in the community, there have been only a handful of efforts at whispered speech emotion recognition by now [13], [14]. Hence, this present work sheds light on whispered speech emotion recognition particularly in terms of using phase-based features.

To exploit the short-term stationary properties of natural sounds, speech processing fundamentally depends on the Short-Time Fourier Transform (STFT), which is

used to determine the magnitude and phase content of a speech signal. Normally, the magnitude content is retained to derive acoustic features needed for speech emotion recognition [3], [15]–[17]. For example, the most frequently used acoustic features Mel-Frequency Cepstral Coefficients (MFCCs), which have been constantly proven useful in speech emotion recognition and speech recognition, are derived by applying a nonlinear mel-scale filter bank on the power spectrum. Further, [18] and [19] suggested that the lower order MFCCs are more relevant to affect and paralinguistic speech tasks. Accordingly, [19] presented a minimalistic set of voice parameters for affective computing by leveraging the first four MFCCs, resulting in a promising performance on a wide range of affective computing tasks. Depending on three types of MFCC features, an 'EmoNet' integrating with deep learning approaches was the winning submission in the emotion recognition in the wild challenge 2013 [7], [20].

In contrast to the widely accepted magnitude spectrum, the role of the phase spectrum of the signal has been largely ignored because of the difficulties in phase wrapping [21], [22]. On the one hand, early studies suggested the unwarranted use of phase processing for speech processing [23]–[25]. On the other hand, a variety of recent studies have reported the importance of the phase information for different audio processing applications, including speech recognition [26], [27], speech enhancement [28], [29], and speaker recognition [30], [31].

Recently, phase-based representation processing is a growing trend in the speech enhancement community [28], [29]. In human listening tests, [32] found that the phase spectrum can contribute to the speech intelligibility as much as the magnitude spectrum. Reference [33] systematically investigated whether speech enhancement approaches can benefit from modifying the phase spectrum in terms of speech quality. The results showed that significant improvements of speech quality are possible especially when the clean phase spectrum is known. And some improvements were obtained when only the noise phase spectrum was available for processing. Besides, several studies demonstrated that using an independently estimated phase results in improvement in the perceived quality [28], [34], [35]. Instead of estimating the magnitude and phase spectrum separately, [36] proposed a Wiener filtering method for speech enhancement to jointly estimate the phase and magnitude spectrum, yielding an improvement over the traditional Wiener filtering method.

Phase spectrum also is attracting increasing interest in the speaker recognition and verification domain [30], [31], [37]–[40]. Reference [41] analytically demonstrated that group delay based features, which are derived from the phase spectrum [42], [43], are robust to additive noise. The experimental result showed that, the group delay based features obtain lower error rates when compared to the traditional MFCC featuers for a speaker verification task on the NIST 2003 database. Further, phase spectrum can be considered as complementary information to augment magnitude spectrum (e. g., MFCCs) and then result in obvious improvement against the common systems solely adopting the magnitude spectrum [37], [39]. In the first automatic speaker verification spoofing and countermeasures challenge 2015, the phase-based feature representations were a highly popular choice in the feature extraction [38], [39], [44].

These aforementioned examples suggest that, incorporating the phase information can extend the horizon for audio and speech signal processing beyond the current limit of phase-independent solutions employed for long time by speech scientists. However, in the domain of speech emotion recognition, there exists very little research with respect to phase-based features. Recently, [45] investigated the phase distortion for emotional valence recognition, defined as the derivative of the relative phase shift. More recently, we have preliminarily demonstrated the usefulness of phase-based features for whispered speech emotion recognition in [46]. Such work made use of the *modified group delay feature* in conjunction with a Fisher kernel and a linear kernel Support Vector Machine (SVM) for improved speech emotion recognition.

In this work, we extend our previous work in [46] and continue to look into the phase-based features for whispered speech emotion recognition. In addition to the modified group delay feature, another phase-based feature derived by the *group delay function of all-pole models* is investigated, which was proposed for speaker recognition in [40]. To the best of our knowledge, the group delay function of all-pole models has not been applied for speech emotion recognition by now. Besides, we propose a novel SVM-based speech emotion recognition framework that enjoys the benefit of SVMs and manipulates variable length sequences of segmental features in a very efficient way. To this end, we use the simple, yet efficient *outer product* technique to map the varied length series of the segmental phase-based feature vectors into a fixed length feature vector, which meets the requirements of the input form of SVMs. The great advantage of the outer product over other so-called length normalisation or bag-of-words techniques, such as the Fisher kernel encoding is that it completes the mapping by just the tensor product, instead of by the involvement of any generative models (e. g., Gaussian Mixture Models (GMMs) in the Fisher kernel encoding). Lastly, we adopt a linear kernel SVM to train the emotion recognition model with the resulting outer product vectors. Analogous to the work in [46], the proposed work is evaluated on the Geneva Whispered Emotion Corpus (GeWEC) under various mismatched speech mode conditions.

The organization of this paper is as follows. Section II first introduces the proposed methods, including two selected phase-based features, the outer product of the trajectory matrix, and normalisation. Next, Section III presents the experimental results on the GeWEC database. In Section IV, finally, we conclude this paper and point out future work.

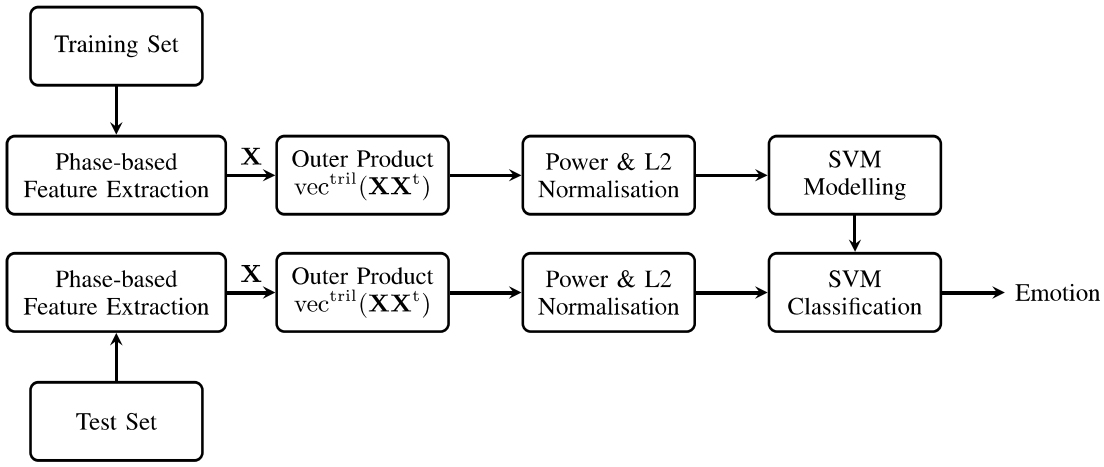


FIGURE 1. Block scheme of the proposed speech emotion recognition system, using phase-based feature extraction, outer product, the power and L2 normalisation, and SVMs.

II. PROPOSED METHODS

A. SYSTEM DESCRIPTION

The proposed system, illustrated in Figure 1, generally consists of four major modules, including a *phase-based feature extraction module*, an *outer product module*, a *normalisation module*, and a *linear kernel SVM module*. Specifically, this work employs two types of phase-based features: the modified group delay feature [43] and the group delay function of all-pole models [40]. To ease the way to build an efficient emotion recogniser, SVMs, which have been found very powerful in a wide range of applications, are chosen as the back-end model. In order to deploy SVMs for classifying the emotion label, the extracted feature sequences of the input utterances, which are usually of variable length, need to be mapped to vectors of fixed-length. This is accomplished by using the outer product in this work. Besides, the *power and L2 normalisation* are considered to correct the values of the resulting outer product vector to an appropriate range for the SVM modelling.

B. PHASE-BASED FEATURES

The Fourier transform of a discrete time digital signal $x(n)$ can be computed in the polar form as

$$X(\omega) = |X(\omega)|e^{j\phi(\omega)}, \tag{1}$$

where $|X(\omega)|$ and $\phi(\omega)$ are the magnitude and phase spectrum.

The common spectral features for speech processing only make use of the magnitude spectrum whereas often rejecting the use of the phase spectrum. Extracting useful features from the phase spectrum is a challenging task because of wrapping of the phase spectrum and its dependency on the window position [47], [48]. In the community, however, a large body of previous studies have shown that extracting the phase information of a signal is applicable and systems using the extracted phase information deliver promising performance. Among them, the Modified Group Delay feature (referred to as MGD) and the All-Pole Group Delay feature (referred to as APGD) are widely used for speech recognition [27], [49], speaker recognition [31], [40], [43], [50], speaker verification [39], and environmental sound events detection [48]. Inspired by the big success of them, we explore these two phase-based features to build up a speech emotion recognition system in this work.

1) MODIFIED GROUP DELAY FEATURE

The group delay function is derived as the negative derivative of the Fourier phase spectrum and can be explicitly written as follows [42], [43], [46]

$$\begin{aligned}\tau_g(\omega) &= -\frac{d(\phi(\omega))}{d\omega} \\ &= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2},\end{aligned} \tag{2}$$

where the angular frequency ω is limited in $[0, 2\pi]$, $|X(\omega)|$ is the magnitude of the Fourier transform of $x(n)$, $Y(\omega)$ is the Fourier transform of the signal $y = nx(n)$, and the subscripts R and I indicate real and imaginary parts, respectively.

It has been observed that, the features derived by the group delay function are discriminative and additive for recognition [51]. But the function often leads to an erroneous representation of a given speech signal. The term $|X(\omega)|$ in the denominator of the group delay function in Section II-B.1 goes towards zeros, especially, when the zeros of the system transfer function are very close to the unit circle in the z-plane [26], [27]. For this reason, the group delay function at frequency bins near these zeros inevitably results in spurious spikes and becomes ill-behaved although it is able to produce a meaningful representation of a signal to a certain extent.

To overcome the spiky nature of the group delay feature, a modification of the group delay function is proposed in [26], which is computed as

$$\tau_m(\omega) = \frac{\tau_p(\omega)}{|\tau_p(\omega)|}|\tau_p(\omega)|^{\alpha}, \tag{3}$$

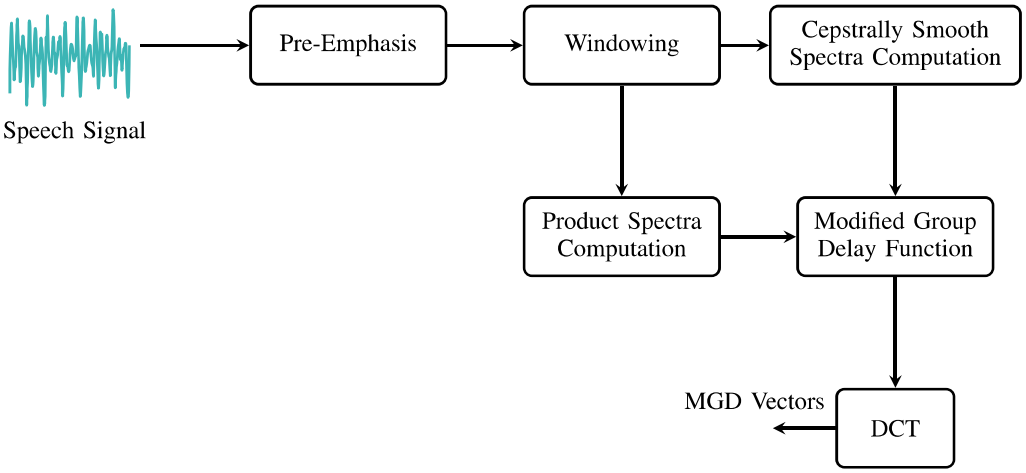


FIGURE 2. Modified Group Delay feature (MGD) Computation Process.

where

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{4}$$

and $S(\omega)$ is a cepstrally smoothed form of $|X(\omega)|$. The two tuning parameters γ and α control the range dynamics of the MGD spectrum. Note that, $P(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)$, called the product spectra, includes information from both the magnitude and phase spectrum [52]. In practice, the cepstrally smoothed form $|S(\omega)|$ is commonly derived as follows [53]:

1) Take the log-magnitude Fourier spectra of $X(\omega)$ and smooth the log spectra by applying the fifth order of the median filter.
2) Take the Discrete Cosine Transform (DCT) of the log spectra and retain the first 30 coefficients.
3) Take the inverse DCT of the cepstral coefficients to result in the smoothed spectra $|S(\omega)|$.

Figure 2 shows the complete computation process of the MGD-based feature extraction. In a manner similar to the computation of MFCCs, the speech signal is first being pre-emphasised and then framed by a Hamming window. Afterwards, the MGD features are computed by Section II-B.1. Finally, the DCT is applied on the MGD features so as to perform a decorrelation. In general, the first coefficient obtained by the DCT is excluded to avoid the effects of the average value.

2) GROUP DELAY FUNCTION OF ALL-POLE MODELS

Recently, [40] presented a new phase representation for speaker recognition in a form of the group delay function of all-pole models. The APGD is capable of interpreting properly the phase information and it avoids extensive parameter adjustment in comparison with the MGD feature. This feature has been successfully used in formant extraction [54], speaker recognition [40], musical instrument recognition [55], and environmental sound event recognition [55].

Unlike the MGD feature computing the group delay function directly from the signal, the APGD feature is rooted in group delay functions of parametric all-pole models of speech signals. Linear prediction analysis approximates the short-term power spectrum by means of all-pole models [56]. From that point, linear prediction is formulated as

$$H(\omega) = \frac{G}{1 - \sum_{k=1}^{p} a(k)e^{-j\omega k}}, \tag{5}$$

where $a(k)$ are the prediction coefficients, G is the signal dependent gain, and p is called the prediction order. We calculate coefficients $a(k)$ by minimising the least square errors between the power spectrum of $H(\omega)$ and the speech power spectrum $|X(\omega)|^2$. Here, the gain G is simply set to 1 for normalisation purpose.

Based on the delay function Section II-B.1, then, the all-pole group delay function is defined as the phase response of this filter formed by $H(\omega)$. The overall computation process of APGD is presented in Figure 3, which also ends up performing the DCT for a decorrelation.

C. OUTER PRODUCT OF TRAJECTORY MATRIX

A robust speech emotion recognition should be capable of recognising emotions from different utterances with varying lengths. In feature extraction, short-time analysis of an utterance leads to a sequence of segmental feature vectors. The length of the sequence may vary due to the different length of given utterances. Hence, care must be taken to process sequences of variable length in recognition modelling. A popular method is to directly train dynamical models, such as Hidden Markov Models (HMMs) [57], [58], on segmental features (e. g., MFCCs). Another widely used method is to project segmental feature vectors onto a fixed-length vector by applying various functionals (e. g., max and min functionals) over time before modelling. The resulting features are known as supra-segmental features [16]. Then, powerful static models, such as SVMs and k-Nearest Neighbours (k-NN), are used to analyse emotions in speech. In contrast to segmental features based methods, these methods simplify the process of building the speech

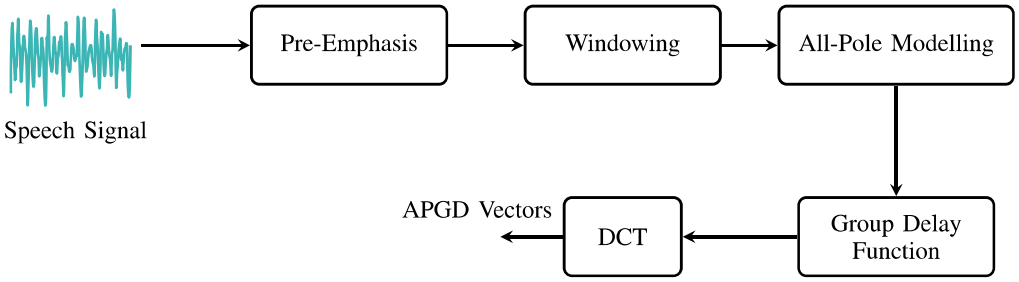


FIGURE 3. All-Pole Group Delay feature (APGD) computation process.

emotion recognition system and constantly save computation cost and test time, especially when the input speech is long.

Following the line of supra-segmental based features approaches, this present work investigates SVM-based emotion recognition models by leveraging the outer product of the trajectory matrix, which is a simple, but efficient sequence length normalisation technique. Such a sequence length normalisation method obtains a fixed-length vector, which is independent of the length of the input utterance, and produces also salient feature representation necessary for speech emotion recognition. To the best of our knowledge, it is the first time to adopt the outer product based length normalisation method for speech emotion recognition.

The outer product of the trajectory matrix was proposed for acoustic modelling using SVMs by [59], where a sequence is viewed as a trajectory in segmental-feature space. Since then, the outer product method was also used for acoustic event classification [60] and emergency state detection [61]. More recently, [62] extended the outer product idea to improve acoustic separability in feature extraction for deep neural networks based speech recognition. Instead of using the standard filter-bank feature, such work projected the filter-bank outputs onto a tensor product space using decorrelation followed by a bilinear map.

The trajectory matrix of an utterance is defined as the sequence of the T feature vectors with dimension d, which is mathematically written as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathbf{R}^{d \times T}. \tag{6}$$

The outer product feature vector of the trajectory matrix is then defined as

$$\hat{\mathbf{x}} = \text{vec}^{\text{tril}} \left(\frac{1}{T} \mathbf{X}\mathbf{X}^{\text{t}} \right) \in \mathbf{R}^{D}, \tag{7}$$

where $D = \frac{d \times (d+1)}{2}$, and the vectorised function $\text{vec}^{\text{tril}}(\cdot)$ concatenates all the elements in the lower triangular part of the symmetric matrix $\mathbf{X}\mathbf{X}^{\text{t}}$, resulting in a vector with dimension D. To alleviate the effect of the length T, this work further normalises the outer product value by the length. It is obvious that, the outer product generates a fixed dimension vector regardless of the length of the sequence in the input. Using such a vector as features enables us to exploit an SVM for classification.

The outer product feature vector represents the second-order correlation information of the spectral features. This is, because Section II-C can also be written as

$$\hat{\mathbf{x}} = \text{vec}^{\text{tril}} \left(\frac{1}{T} \mathbf{X}\mathbf{X}^{\text{t}} \right) = \text{vec}^{\text{tril}} \left(\frac{1}{T} \sum_{i=1}^{T} \mathbf{x}_i \mathbf{x}_i^{\text{t}} \right), \tag{8}$$

which represents that, the outer product vector $\hat{\mathbf{x}}$ can be considered as the average of the product of all possible pairs of column vectors of a given matrix $\mathbf{X}$.

Figures 4 and 5 visualise examples of the MGD and APGD features and their outer product vectors for the whispered pseudoword belam from the GeWEC corpus expressed in anger, fear, happiness, and neutral voice. Based on the two figures, it can be observed that, the emotion patterns are clearly visible on the MGD and APGD features. Besides, it seems that the resulting outer product vectors are easier to distinguish than the original MGD and APGD features. To further verify this findings, we conduct a similarity analysis to investigate the effectiveness of the outer product vector of phase-based spectral features for the four chosen utterances. Cosine distance is used as a similarity measure to indicate emotion separability of the standard MGD features and their outer product vectors. A large cosine distance reflects high dissimilarity between two vectors. Figure 6 shows the cosine distances among the four chosen utterances. The figure indicates that outer product vectors can improve the emotion separability of MGD features.

D. NORMALISATION

In preliminary experiments, we observed that outer product vectors of spectral features are sparse, which delivers a poor performance. Inspired by the fisher kernel framework for large-scale image classification [63] in which normalisation methods are used to process fisher vectors, this work applies both power normalisation and L2 normalisation to outer product vectors in hope that the processed outer product vectors contain more expressive information needed for emotion recognition.

Power normalisation is an element-wise function, defined as

$$f(x) = \text{sign}(x)|x|^{\beta}, \tag{9}$$

where $0 \leq \beta \leq 1$ is a predefined parameter of the normalisation. Based on previous work (e. g., [63]), here, β is set to 0.5, and this transformation is known as 'signed square-rooting'.

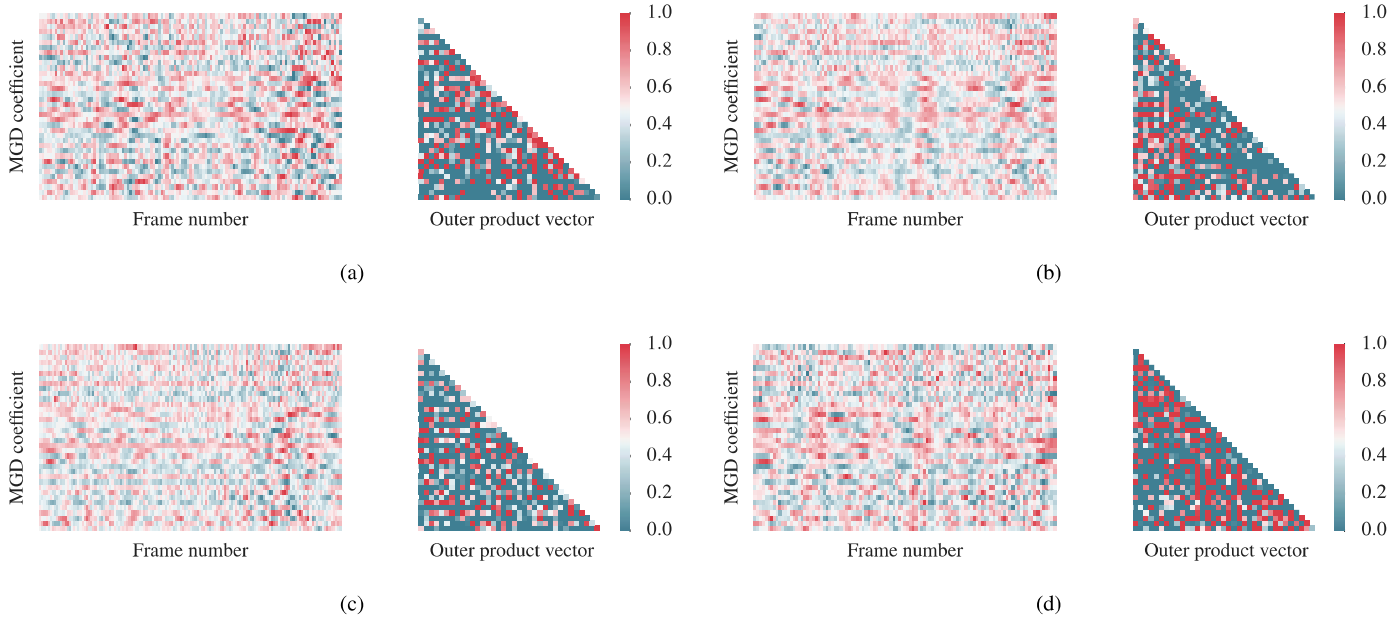


FIGURE 4. Examples of the *MGD* features and their outer product vectors for the whispered pseudoword "belam" expressed in four emotions: anger, fear, happiness, and neutral, taken from the GeWEC corpus. Best viewed in colour. (a) Anger. (b) Fear. (c) Happiness. (d) Neutral.

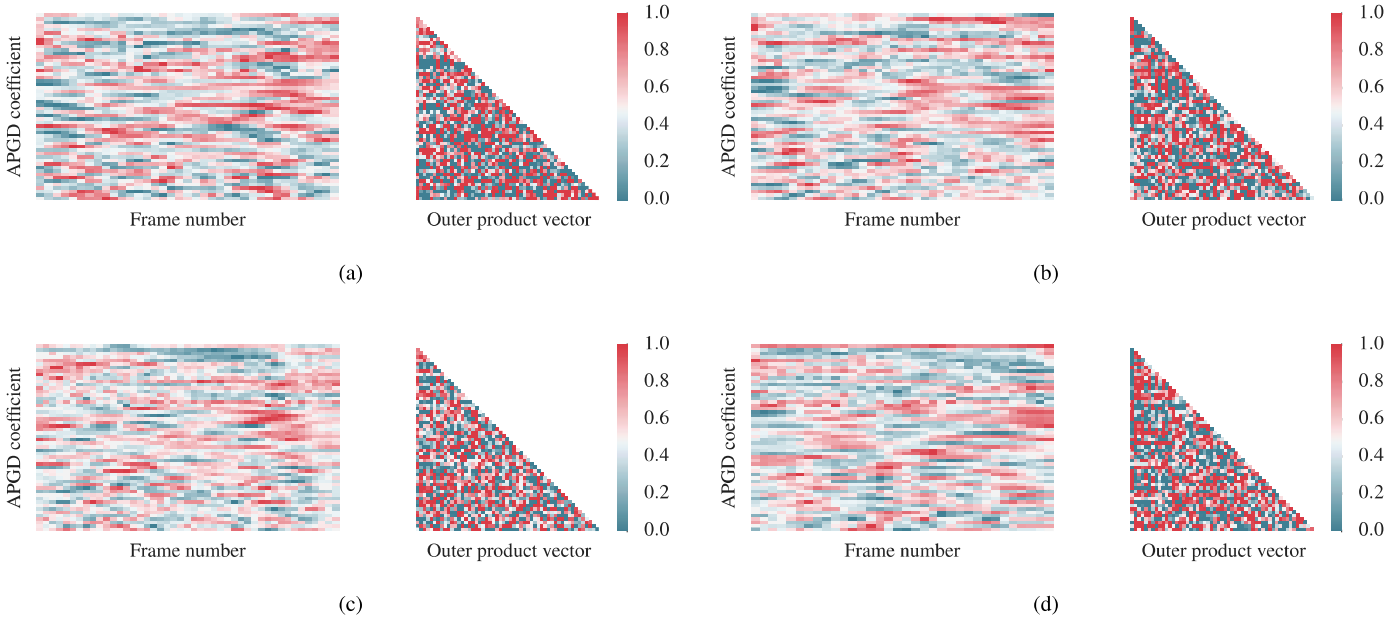


FIGURE 5. Examples of the *APGD* features and their outer product vectors for the whispered pseudoword "belam" expressed in four emotions: anger, fear, happiness, and neutral, taken from the GeWEC corpus. Best viewed in colour. (a) Anger. (b) Fear. (c) Happiness. (d) Neutral.

Besides, L2 normalisation is also investigated to reduce the dependence on the amounts of individual speaker information and improve the performance as long as classifiers (e. g., SVMs) involve dot-products. Given a feature vector $\mathbf{x} \in \mathbf{R}^d$, the L2 normalisation is written as

$$l^2(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\sum_{i=1}^{d} x_i^2}}. \tag{10}$$

When using both the power and L2 normalisation, we apply the power normalisation ahead of the L2 normalisation. Figure 7 shows the effect of the combination of power normalisation and L2 normalisation on the distribution of the first dimension of the outer product vector of APGD. As can be seen from Figure 7, the normalisation combination can unsparsify outer product features.

III. EXPERIMENTS

A. GENEVA WHISPERED EMOTION CORPUS

We employ the Geneva Whispered Emotion Corpus (GeWEC) to evaluate the effectiveness of the

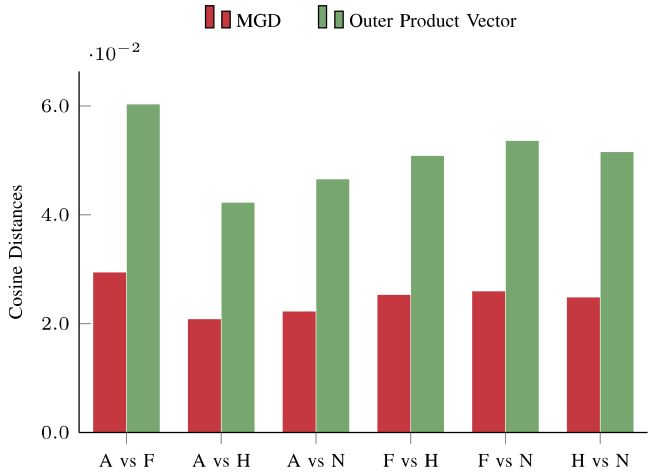


FIGURE 6. Cosine distances among examples (used in Figures 4 and 5) for anger (A), fear (F), happiness (H), and neutral (N) with the standard modified group delay features (MGD) and their outer product features of the trajectory matrix. The larger the cosine distance is, the more dissimilar the two feature vectors are.

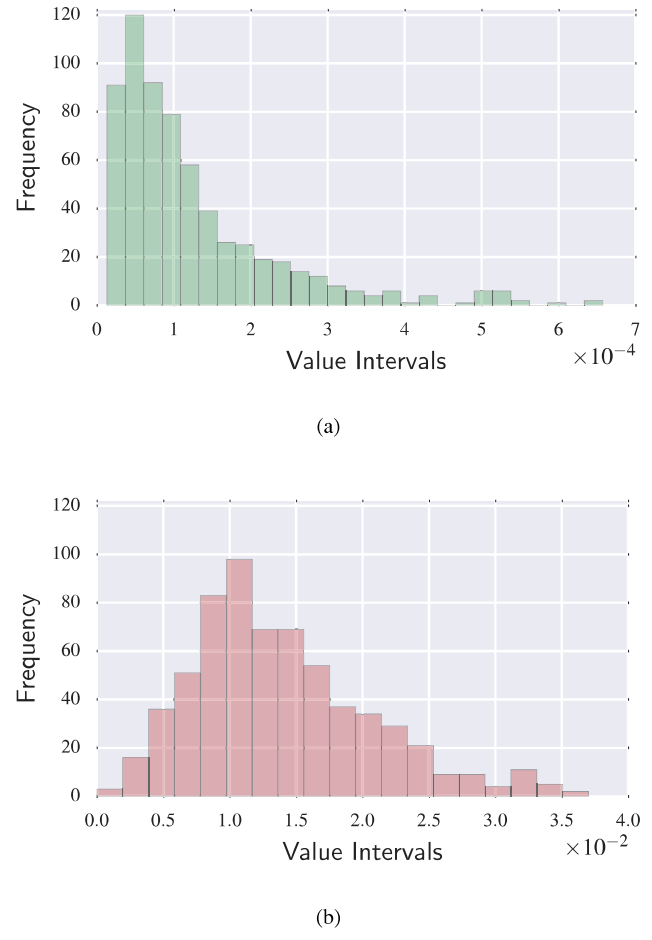


FIGURE 7. Effect of L2 and power normalisation on the distribution of the first dimension of the outer product vector of APGD. The histograms are estimated from the whispered speech mode of the GeWEC. (a) Outer Product Vector of APGD without L2 and Power Normalisation. (b) Outer Product Vector of APGD with L2 and Power Normalisation.

proposed system. The corpus provides normal and whispered paired utterances. Two male and two female professional French-speaking actors in Geneva Switzerland were recruited to speak eight predefined French pseudo-words (e. g., *"belam"* and *"molen"*) with a given emotional state in both normal and whispered speech modes as in the GEMEP-corpus that was used in the Interspeech 2013 Computational Paralinguistics Challenge [64]. Speech was expressed in four emotional states: *angry*, *fear*, *happiness*, and *neutral*. The actors were requested to express each word in all four emotional states five times. The utterances were labelled based on the state they should be expressed in, i. e., one emotion label was assigned to each utterance. As a result, GeWEC consists of 1 280 instances in total. To keep in line with the previous work [46], a cross-speech-mode evaluation is considered in the experiments. That is, one speech mode of the GeWEC data is used for training while the other speech mode data is used for testing.

B. EXPERIMENTAL SETUP

As for the feature extraction, pre-emphasis is first conducted. Afterwards, frame windowing is performed using a Hamming window with a frame-length of 25 ms and a frame-shift of 10 ms. When computing the MGD features based on Section II-B.1, we set the two tuning parameters α and γ to 0.1 and 0.2 based on the preliminary experiments, which ends up in 36-dimensional MGD features including delta and acceleration coefficients. As for the APGD features, the order of the all-pole mode p is 30, and 18 APGD coefficients are kept. Delta and acceleration coefficients are appended to the APGD coefficients to form 54-dimensional APGD feature vectors. In addition to the two phase-based features, the most frequently used magnitude-based features, 36-dimensional MFCCs, are also extracted. As for the basic supervised learner in the classification step, we use linear SVMs implemented in LIBLINEAR [65].

Since the proposed method is a general SVM-based speech emotion recognition system, we choose various related computational paralinguistics recognition systems as the baselines depending on the open source openSMILE toolkit [66], [67]. These chosen have served as the baselines for Interspeech Challenges, which use supra-segmental features and choose SVMs as the classifier. In details, the supra-segemental feature sets used include Interspeech Challenges on Emotion in 2009 [68] (referred to as IS09), Level of Interest in 2010 [69] (referred to as IS10), Speaker States in 2011 [70] (referred to as IS11), Speaker Traits in 2012 [71] (referred to as IS12), Emotion in 2013 [64] (referred to as IS13), and the most recently proposed Geneva minimalistic acoustic parameter set (referred to as GeMAPS) and the extended Geneva minimalistic acoustic parameter set (referred to as eGeMAPS) [19].

Unweighted Average Recall (UAR) is used as a performance metric as in these challenges. Besides, statistical significance tests are conducted by computing a one-sided z-test.

C. RESULTS OF PHASE-BASED FEATURES

Following the experimental setting, we first investigate the performance of the proposed systems using the

TABLE 1. UAR for four-way cross-mode emotion recognition on GeWEC: When one speech mode of GeWEC (normal speech (norm.) or whispered speech (whisp.)) is used for training, the other one is used for testing. The proposed system compares with different state-of-the-art SVM-based emotion recognition systems using supra-segmental features. Significant results (p-value < 0.05, one-sided z-test) are given in parentheses. Maximal UAR is highlighted in bold.

UAR [%]	Norm. (train), Whisp. (test)	Whisp. (train), Norm. (test)
IS09 + SVM	35.5	53.4
IS10 + SVM	39.5	52.3
IS11 + SVM	40.3	52.8
IS12 + SVM	33.3	46.4
IS13 + SVM	36.4	48.4
GeMAPS + SVM	34.1	32.0
eGeMAPS + SVM	41.9	38.9
MGD + Fisher Vectors [46]	50.3	54.8
MGD + Outer Product	49.1 ($p < 0.005$)	53.4
APGD + Outer Product	**53.9** ($p < 0.001$)	55.2
MFCC + Outer Product	50.3 ($p < 0.002$)	**59.7** ($p < 0.05$)

phase-based features. Table 1 presents the experimental results on the GeWEC data. In Table 1, p-values in parentheses refer to the significance level of rejecting the null hypothesis that the best baseline systems (e. g., eGEMAPS + SVM) are superior to the proposed system.

As can be seen from Table 1, the proposed systems achieve an appealing performance for the two cross-mode speech emotion tasks. For the first one, where the model is trained on 'normal' speech while tested on whispered speech, the MGD and APGD systems reach 49.1 % UAR and 53.9 % UAR, which both outperform the other supra-segmental features based methods statistically significantly by a large margin. As for the second setting, the proposed methods with the MGD and APGD features achieve 53.4 % UAR and 55.2 % UAR, which are as competitive as the other approaches. Besides, the proposed SVM-based system using the outer product is comparable in performance to the Fisher vector system previously shown in [46], which has yielded the best performance on GeWEC by now.

We further run experiments to show that the proposed system is a general framework for speech emotion recognition, which is comparable to the magnitude-based features, MFCCs, as well. Table 1 reports the results. The proposed system with MFCCs gives the best performance (i. e., 59.7 % UAR) for the normal speech emotion test and produces reliable performance (50.4 % UAR) for the whispered speech emotion test. Furthermore, when comparing the MGD and APGD features with the MFCCs features, we observed that, these two phase-based features are as effective as the MFCCs features in terms of performance in speech emotion recognition.

D. RESULTS OF COMBINING PHASE FEATURES WITH MAGNITUDE FEATURES

Next, we investigate whether the phase-based features and the magnitude-based features are complementary, leading to performance improvement. To this end, the MGD features

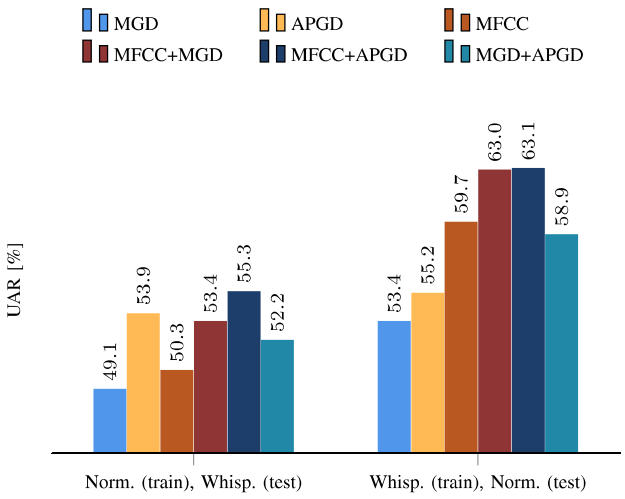


FIGURE 8. UAR for different combinations of the two phase-based features (i. e., MGD and APGD) and the magnitude-based features (i. e., MFCCs) in the proposed method.

and APGD features are respectively concatenated with the MFCCs features to form a long feature vectors. Figure 8 presents the experimental results of different feature combinations in the evaluation of the two tasks. The corresponding confusion matrices for the feature combination of APGD and MFCCs on the GeWEC database are given in Figure 9. According to Figure 8, it is easily observed that, when the MGD or APGD features are combined with the MFCCs features, each combination yields a notable increase in performance over each other in isolation for the two evaluation tasks. In addition, the combination of the MFCCs and APGD features even performs statistically significantly better than the MFCCs features for the whispered emotion recognition test. The combination of the MGD and APGD features achieves modest performance improvement when compared to the original MGD or APGD features.

These results echo the observations found in [37], [39], and [47] that the phase-based features can have a complementary role in speech processing, and combining

	A	F	H	N
A	97	17	20	26
F	18	56	51	35
H	6	16	116	22
N	52	3	20	85

(a)

	A	F	H	N
A	77	5	56	22
F	27	100	13	20
H	5	14	113	28
N	14	8	24	114

(b)

FIGURE 9. Confusion matrices obtained by the proposed method using the feature combination of APGD and MFCCs in the 4-way (i. e., anger (A), fear (F), happiness (H), and neutral (N)) cross-mode emotion recognition. (a) Norm. (train), Whisp. (test). (b) Whisp. (train), Norm. (test).

the magnitude features with the phase-based features together can improve the performance.

IV. CONCLUSIONS

In this paper, we focused on improving whispered speech emotion recognition in a challenging whispered vs non-whispered speech and vice-versa cross-mode setting. When exploiting the effectiveness of two types of phased-based features (i. e., modified group delay features and all-pole group delay features), we presented a novel speech emotion recognition framework with outer product and power normalisation plus L2 normalisation. Cross-speech-mode experiments on the GeWEC data were conducted, demonstrating that the present framework is competitive with or superior to other modern emotion recognition models. We also found empirical evidence that the two phase-based representations could be used to achieve comparable performance to the common magnitude-based Mel-Frequency cepstral coefficients in speech emotion recognition. Furthermore, it was demonstrated that, the combination of the magnitude and the phase information could significantly improve performance over the conventional systems purely adopting magnitude information.

To further improve the performance, one potential direction is to use transfer learning [72] to address the difference among the cross-mode speech settings. Besides, future work could extend the proposed work to various computational paralinguistics tasks such as native language recognition task between whispered and non-whispered speech [73].

REFERENCES

[1] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] J. B. Walther and K. P. D'Addario, "The impacts of emoticons on message interpretation in computer-mediated communication," *Soc. Sci. Comput. Rev.*, vol. 19, no. 3, pp. 324–347, 2001.

[3] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2012.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[5] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. New York, NY, USA: Wiley, Nov. 2013.

[6] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy, "Amplitude modulation features for emotion recognition from speech," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 2420–2424.

[7] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[8] X. Xu, J. Deng, W. Zheng, L. Zhao, and B. Schuller, "Dimensionality reduction for speech emotion features by multiscale kernels," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1532–1536.

[9] J. Deng, W. Han, and B. Schuller, "Confidence measures for speech emotion recognition: A start," in *Proc. ITG Symp.*, Braunschweig, Germany, 2012, pp. 1–4.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 223–227.

[11] K. Tsunoda, S. Sekimoto, and T. Baer, "An fMRI study of whispering: The role of human evolution in psychological dysphonia," *Med. Hypotheses*, vol. 77, no. 1, pp. 112–115, 2011.

[12] K. Simonyan and C. L. Ludlow, "Abnormal activation of the primary somatosensory cortex in spasmodic dysphonia: An fMRI study," *Cerebral Cortex*, vol. 20, no. 11, pp. 2749–2759, 2010.

[13] C. Gong, H. Zhao, W. Zou, Y. Wang, and M. Wang, "A preliminary study on emotions of Chinese whispered speech," in *Proc. IFCSTA*, vol. 2. Chongqing, China, 2009, pp. 429–433.

[14] J. Zhou, R. Liang, L. Zhao, L. Tao, and C. Zou, "Unsupervised learning of phonemes of whispered speech in a noisy environment based on convolutive non-negative matrix factorization," *Inf. Sci.*, vol. 257, pp. 115–126, Feb. 2014.

[15] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 344–347.

[16] B. W. Schuller, *Intelligent Audio Analysis* (Signals and Communication Technology). New York, NY, USA: Springer, 2013.

[17] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," in *Proc. ICASSP*, Vancouver, BC, Canada, 2013, pp. 7527–7531.

[18] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions—A new perspective on feature selection," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 2044–2048.

[19] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr./Jun. 2016.

[20] S. E. Kahou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. ICMI*, Sydney, NSW, Australia, 2013, pp. 543–550.

[21] P. Mowlaee, R. Saeidi, and Y. Stylianou, "INTERSPEECH 2014 special session on phase importance in speech processing applications," in *Proc. INTERSPEECH*, Singapore, 2014, p. 5.

[22] B. Yegnanarayana, J. Sreekanth, and A. Rangarajan, "Waveform estimation using group delay processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 832–836, Aug. 1985.

[23] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

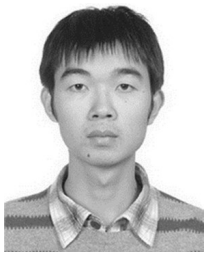
[24] P. Vary and M. Eurasip, "Noise suppression by spectral magnitude estimation: Mechanism and theoretical limits," *Signal Process.*, vol. 8, no. 4, pp. 387–400, 1985.

[25] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

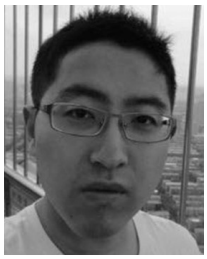
[26] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. ICASSP*, Hong Kong, 2003, pp. 68–71.

[27] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 190–202, Jan. 2007.

[28] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. ICSLP*, Hong Kong, 2012, pp. 1–4.
[29] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.
[30] N. Wang, P. C. Ching, and T. Lee, "Exploitation of phase information for speaker recognition," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2126–2129.
[31] I. Hernáez, I. Saratxaga, J. Sanchez, E. Navas, and I. Luengo, "Use of the harmonic phase in speaker recognition," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 2757–2760.
[32] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Commun.*, vol. 45, no. 2, pp. 153–170, 2005.
[33] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.
[34] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.
[35] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, Feb. 2013.
[36] J. Le Roux and E. Vincent, "Consistent wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.
[37] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012.
[38] J. A. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2067–2071.
[39] M. J. Alam, P. Kenny, and T. Stafylakis, "Combining amplitude and phase-based features for speaker verification with short duration utterances," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 249–253.
[40] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 2489–2493.
[41] R. Padmanabhan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase-based features for speaker recognition," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 2355–2358.
[42] H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *Proc. ICASSP*, Seattle, WA, USA, 1998, pp. 861–864.
[43] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Process.*, vol. 22, no. 3, pp. 259–267, 1991.
[44] Z. Wu *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.
[45] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. ICSP*, Beijing, China, 2012, pp. 693–696.
[46] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Proc. IWSDS*, Saariselkä, Finland, 2016, p. 6.
[47] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2026–2038, Sep. 2011.
[48] A. Diment, E. Cakir, T. Heittola, and T. Virtanen, "Automatic recognition of environmental sound events using all-pole group delay features," in *Proc. EUSIPCO*, Nice, France, 2015, pp. 729–733.
[49] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," in *Proc. EUSIPCO*, Antalya, Turkey, 2005, pp. 1–4.
[50] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, p. 4.
[51] K. M. Murthy and B. Yegnanarayana, "Effectiveness of representation of signals through group delay functions," *Signal Process.*, vol. 17, no. 2, pp. 141–150, 1989.
[52] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. ICASSP*, Montreal, QC, Canada, 2004, pp. 125–128.
[53] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. ICASSP*, Vancouver, BC, Canada, 2013, pp. 7234–7238.
[54] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *J. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1638–1640, 1978.
[55] A. Diment, P. Rajan, T. Heittola, and T. Virtanen, "Modified group delay feature for musical instrument recognition," in *Proc. CMMR*, Marseille, France, 2013, pp. 431–438.
[56] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
[57] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[58] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. ICASSP*, Hong Kong, 2003, pp. 1–4.
[59] R. Anitha, D. Satish, and C. Sekhar, "Outerproduct of trajectory matrix for acoustic modeling using support vector machines," in *Proc. MLSP*, Maranhão, Brazil, 2004, pp. 355–363.
[60] A. Temko, E. Monte, and C. Nadeu, "Comparison of sequence discriminant support vector machines for acoustic event classification," in *Proc. ICASSP*, 2006, pp. 721–724.
[61] E. Principi, S. Squartini, E. Cambria, and F. Piazza, "Acoustic template-matching for automatic emergency state detection: An ELM based algorithm," *Neurocomputing*, vol. 149, pp. 426–434, Feb. 2015.
[62] T. Ogawa, K. Ueda, K. Katsurada, T. Kobayashi, and T. Nitta, "Bilinear map of filter-bank outputs for DNN-based speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 16–20.
[63] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. ECCV*, Crete, Greece, 2010, pp. 143–156.
[64] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 148–152.
[65] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.
[66] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. MM*, Florence, Italy, 2010, pp. 1459–1462.
[67] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. MM*, Barcelona, Spain, 2013, pp. 835–838.
[68] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.
[69] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2794–2797.
[70] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 3201–3204.
[71] B. Schuller *et al.*, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, pp. 1–4.
[72] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.
[73] B. Schuller *et al.*, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 1–5.



JUN DENG received the bachelor's degree in electronic and information engineering from Harbin Engineering University in 2009, and the master's degree in information and communication engineering from the Harbin Institute of Technology, China, in 2011, the Ph.D. degree in electrical engineering and information technology from TUM, Munich, Germany, in 2016, with a focus on feature transfer learning for speech emotion recognition. He is currently a Post-Doctoral Researcher at the Chair of Complex and Intelligent Systems with the University of Passau, Passau, Germany. His interests are machine learning methods such as transfer learning and deep learning with an application preference to affective computing.



XINZHOU XU received the bachelor's degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, and the master's degree from Southeast University, Nanjing, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with Southeast University. He is also with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany, and the Chair of Complex and Intelligent Systems with the University of Passau, Germany. His research interests include spoken signal processing, pattern recognition, machine learning, and affective computing.

ZIXING ZHANG (M'15) received the Ph.D. degree in engineering from the Institute for Human-Machine Communication with Technische Universität München, Germany, in 2015, and the master's degree in physical electronics from the Beijing University of Posts and Telecommunications, China, in 2010. He is also a Post-Doctoral Researcher with the University of Passau, Germany. He has authored over 30 publications in peer-reviewed journals and conference proceedings. His research interests mainly lie in deep learning, semi-supervised learning, active learning, and multi-task learning, in the application of computational paralinguistics, and robust automatic speech recognition.

SASCHA FRÜHHOLZ received the degree in science of education in 2001 and the degree in psychology in 2006, and the Ph.D. degree with Bremen University in 2008, with a focus on the neural mechanisms of facial expressions. He is currently SNSF Professor with the Department of Psychology, University of Zurich, Zurich, Switzerland. He is also with the Neuroscience Center Zurich, ZNZ, University of Zurich and ETH Zurich, Switzerland, and the Zurich Center for Integrative Human Physiology, ZIHP, University of Zurich, Switzerland. His current projects deal with dynamic connectivity patterns of local and remote brain regions during affective voice processing using high-resolution brain scans and specific connectivity modeling approaches for functional imaging data.

BJÖRN SCHULLER (M'05–SM'15) received the Diploma degree in 1999, the Ph.D. degree in automatic speech and emotion recognition in 2006, and the Habilitation degree from TUM, all in electrical engineering and information technology. He is currently an Adjunct Teaching Professor in signal processing and machine intelligence in 2012. He is also a Tenured Full Professor heading the Chair of Complex & Intelligent Systems with the University of Passau, Germany, and a Reader in machine learning with the Department of Computing, Imperial College London, London, U.K. He has co-authored five books and over 550 publications in peer-reviewed books, journals, and conference proceedings leading to over 11 000 citations (h-index=51). He is also the President-Emeritus of the Association for the Advancement of Affective Computing (AAAC), and an elected member of the IEEE Speech and Language Processing Technical Committee, and a member of the ACM and ISCA.