Speech Emotion Classification Using Multiple Kernel Gaussian Process

Sih-Huei Chen[1], Jia-Ching Wang[1], Wen-Chi Hsieh[1], Yu-Hao Chin[1], Chin-Wen Ho[1], and Chung-Hsien Wu[2]
[1]Department of Computer Science and Information Engineering
National Central University, Taiwan, R.O.C.
[2]Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.

Abstract—**Given the increasing attention paid to speech emotion classification in recent years, this work presents a novel speech emotion classification approach based on the multiple kernel Gaussian process. Two major aspects of a classification problem that play an important role in classification accuracy are addressed, i.e. feature extraction and classification. Prosodic features and other features widely used in sound effect classification are selected. A semi-nonnegative matrix factorization algorithm is then applied to the proposed features in order to obtain more information about the features. Following feature extraction, a multiple kernel Gaussian process (GP) is used for classification, in which two similarity notions from our data in the learning algorithm are presented by combining the linear kernel and radial basis function (RBF) kernel. According to our results, the proposed speech emotion classification apporach achieve an accuracy of 77.74%. Moreover, comparing different apporaches reveals that the proposed system performs best than other apporaches.**

Index Terms—**Speech emotion classification, multiple kernel Gaussian process, semi-nonnegative matrix factorization.**

I. Introduction

As technology advances significantly contribute to the perception of human emotions with a computer or a machine in daily life, the pervasiveness of emotion measurement technology will allow individuals to develop innovative ways of communication. Among the numerous applications of this technology include home health care systems, social networking, and e-learning systems. In human-computer interaction (HCI), emotion classification by a computer still remains a challenging task, especially speech emotion.

The ability of a classification system to demonstrate good results involves selecting proper features and using a reliable classifier. In speech emotion classification, many low-level or high-level acoustic features have been devised to isolate information of emotions in speech signals, including pitch, energy, timing, and voice quality. Low-level features contain prosodic features extracted from each frame or total speech signal duration. However, high-level features derive from low-level features including mean, variance, maximum, minimum, range, skew, and kurtosis [21]. Many works [1, 2] confer that high-level features are superior to low-level features in terms of classification accuracy and dimensionality. Besides using of prosodic features, some studies have been adopted spectral and cepstral features, including log-filter power coefficients (LFPCs), Mel frequency cepstral coefficients (MFCCs), and linear predictive cepstral coefficients (LPCCs). Improved results can be obtained when different features are combined to increase information of signals. For example, Chuang *et al.* combined acoustic features and textual content together for emotion recognition [3], and Schuller *et al.* fused acoustic and linguistic features to perform speaker emotion recognition [4]. In addition to using acoustic features, this work applies a novel matrix factorization method to the original features to obtain more information about the features.

Primitive data sets are often organized as data matrices, and factorized into two matrix for linear combination representation of bases or dimensionality reduction. Conventional matrix factorization methods, such as principal components analysis (PCA), vector quantization (VQ), linear discriminant analysis (LDA), independent component analysis (ICA), are well-known exemplars of matrix factorization. Lee and Seung [5] proposed a standard paradigm of matrix factorization called nonnegative matrix factorization (NMF) as a highly effective means of factorizing a matrix as the product of two matrices, in which all elements of these three matrices are nonnegative. The NMF algorithm has been successfully applied, such as in image processing, source separation, and musical genre classification [6-8].

Besides the traditional NMF, many variations of NMF algorithm are created by adding or changing constraints to the decomposition. For instance, the sparse NMF proposed in [9] imposes the sparseness constraint, and discriminant NMF [10] integrates label information into objective function, i.e. the within-class scatter and the between-class scatter. In terms of changing constraints, semi-NMF removes the non-negative constraints on the data and basis matrix, making it applicable to both the matrix of mixed signs and many other fields [19].

The choice of classifiers also holds a major role in classification systems. Previous works have used various classifiers for emotion classification, including k-nearest neighbors (k-NNs), hidden Markov models (HMMs), Gaussian mixture models (GMMs), support vector machines (SVMs), and artificial neural networks (ANNs). Each classifier has its own advantages and limitations. For example, k-NNs are easily implemented, yet computationally intensive for large training sets. Of these classifiers, Gaussian mixture models [13] have been studied the most. While efficient in modeling multi-modal

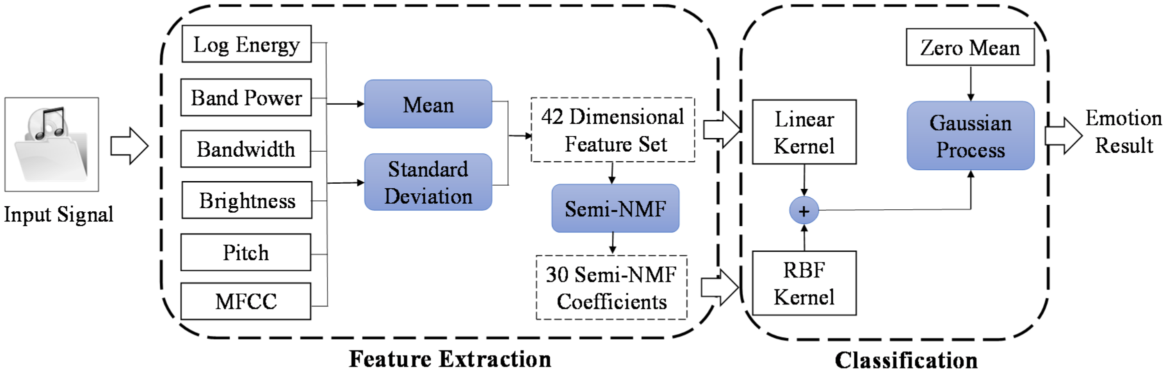

Fig. 1. A block diagram of our proposed emotion classification system.

distributions [22], GMMs are appropriate for global features; however, GMMs suffer from the dimensionality problem [11]. Commonly used in emotion classification, HMMs are stochastic processes that comprise a Markov chain with the hidden states are generated by the observation sequence. Although capable of capturing the temporal structure of the data, HMMs are trained by using only positive data [12]. Among the advantages that SVMs have over GMM and HMM include effectiveness with limited training data and global optimality of algorithm. Finally, in addition to increasing the effectiveness of modeling nonlinear mappings, ANNs have a better classification performance than HMM and GMM when the number of training data is relatively small [14].

This work explores the feasibility of using Gaussian process (GP) classifiers [17] in speech emotion classification. While playing a major role in solving problems such as nonlinear regression and classification, GP is a stochastic process that models distributions over function spaces. The Gaussian process is characterized by its ability to be specified by a covariance function which observes the relation between data points by use of kernel functions. Many works have used the Gaussian process for classification purposes [15, 17]. Moreover, kernel tricks are applicable such that the covariance function can use a combination of more than one kernel function. Using multiple kernels instead of a single kernel can offer different notions of similarity and yield improved results [16, 18]. In this paper, we present a novel emotion classification system based on a multiple kernel Gaussian process.

II. System Overview

Figure 1 schematically depicts the proposed speech emotion classification system, in which the multiple kernel Gaussian process is used. The proposed system can be divided into two parts: feature extraction and classification.

In the feature extraction stage, two acoustic features are selected: prosodic features and some general features used in audio classification. Six features are extracted from input signals, i.e. log energy (1), band power (4), bandwidth (1), brightness (1), pitch (1), and MFCCs (13), where the notation inside the parentheses represents the number of dimensions of features. The signal is decomposed into frames and then these 6 features of each frame are extracted. Next, statistical values such as mean and standard deviation of frame-based features are calculated to create a 42 dimensional feature vector. Following extraction of the 42 dimensional feature from the input signal, the semi-nonnegative matrix factorization (Semi-NMF) method [19] is applied to derive the Semi-NMF coefficients to improve classification accuracy since these coefficients preserve more information of the input signal.

The proposed system also incorporates the use of the multiple kernel Gaussian process based classifier [18], which combines different kernels to present different characteristics of various features. Based on the linear kernel, our 42 dimensional feature is modelled and then combined with the RBF kernel. By using this kernel, the similarity of Semi-NMF coefficients is described using a weighted linear combination approach. Finally, the speech emotion classification results are obtained.

For obtaining objective results, the proposed speech emotion classification system is analyzed using a professional database of seven emotions, i.e. anger, boredom, disgust, fear, joy, neutral, and sadness. Experimental results demonstrate how to decide the number of Semi-NMF coefficients and what kernels are used in the multiple kernel Gaussian process.

III. Semi-NMF Coefficients

Non-negative matrix factorization (NMF) is an unsupervised learning method for locating a meaningful and physically interpretable part-based representation [5]. A NMF problem can be stated as follows.

Given a non-negative data matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ and a constant $r \in \mathbb{N}^+$, NMF decomposes $\mathbf{X}$ into two matrices $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$, such as $\mathbf{X}_+ \approx \mathbf{W}_+\mathbf{H}_+$, where $\mathbf{W}$ is regarded as part-based bases of the data matrix, and $\mathbf{H}$ denotes the combinational coefficients of bases.

Ding *et al*. proposed new variations of NMF algorithm called Semi-NMF [19]. For Semi-NMF, the data matrix $\mathbf{X}$ may have a mixed sign, and only matrix $\mathbf{H}$ is constrained to be non-negative while placing no restriction on the signs of $\mathbf{W}$. The matrix sign definition of Semi-NMF is written as

$$\mathbf{X}_\pm \approx \mathbf{W}_\pm \mathbf{H}_+ \tag{1}$$

By applying results of the algorithm of Ding *et al*. [19], an alternately iterative procedure for computing the nonnegative solution is created.

$$\mathbf{W} = \mathbf{X}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{H}^{\mathrm{T}})^{-1} \tag{2}$$

$$\mathbf{H} \leftarrow \mathbf{H}\sqrt{\frac{(\mathbf{X}^{\mathrm{T}}\mathbf{W})^{+} + [\mathbf{H}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-}]}{(\mathbf{X}^{\mathrm{T}}\mathbf{W})^{-} + [\mathbf{H}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{+}]}} \tag{3}$$

where the positive and negative parts of a matrix $\mathbf{A}$ are defined as $\mathbf{A}^{+} = (|\mathbf{A}| + \mathbf{A})/2$ and $\mathbf{A}^{-} = (|\mathbf{A}| - \mathbf{A})/2$ respectively.

Since Semi-NMF is more flexible than NMF, this work applies Semi-NMF to our 42 dimensional feature set in order to extract coefficient matrix $\mathbf{H}$ and use it for classification. Experimental results indicate that the fusion of the original features and its Semi-NMF coefficients using the proposed system increase the classification accuracy.

IV. Multiple Kernel Gaussian Process for Classification

As a stochastic process, a Gaussian process (GP) models distributions over function spaces [17]. Distribution over the random variables has a joint Gaussian probability. A Gaussian process can be specified by a mean function $\mu(\mathbf{x})$ and a covariance function $k(\mathbf{x},\mathbf{x}')$. The mean function and covariance function of a GP are defined as $\mu(\mathbf{x}) = \mathrm{E}[f(\mathbf{x})]$ and $k(\mathbf{x},\mathbf{x}') = \mathrm{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))]$ respectively. The GP is written as $f(\mathbf{x}) \sim GP(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}'))$.

Given a data set $D = \{(\mathbf{x}_i, y_i) \mid i = 1,\ldots,N\}$, the Gaussian process for classification identifies an appropriate mapping $f(\cdot)$, which satisfies a situation in which $y_i = f(\mathbf{x}_i)$ can match the correct results for data set D where the label of $\mathbf{x}_i$ is y_i. For a two-class classification problem, the probability of a data instance $\mathbf{x}_i$ belonging to a label y_i is defined as

$$p(y_i = +1 \mid f_i) = \sigma(y_i f_i) \tag{4}$$

where $f_i \equiv f(\mathbf{x}_i)$ and $\sigma(\cdot)$ denotes a sigmoid function such as logistic or probit function.

To perform prediction, the probability of a new sample $\mathbf{x}_*$ is obtained by integrating over the latent function f_*.

$$p(y_* = +1 \mid \mathbf{x}_*, \mathbf{y}, \mathbf{X}) = \int p(y_* \mid f_*) p(f_* \mid \mathbf{x}_*, \mathbf{y}, \mathbf{X}) df_* \tag{5}$$

where $\mathbf{y} = [y_1,\ldots,y_N]^{\mathrm{T}}$ and $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_N]^{\mathrm{T}}$.

As is generally assumed, using multiple kernels rather than only a single kernel provides more flexibility, possibly preserving information from different features. Moreover, the different kernels may offer different notions of similarity used together without choosing the only one that works the best. Here, different features are treated using different kernels.

Among the several ways to combine different kernels [18] include linear or multiplication combination. In this work, some general kernels are fused together by using a linear combination method. Hence, our kernel becomes

$$k(\mathbf{x},\mathbf{x}') = a \cdot k_1(\mathbf{x}_1,\mathbf{x}_1') + b \cdot k_2(\mathbf{x}_2,\mathbf{x}_2') \tag{6}$$

where $\mathbf{x} = [\mathbf{x}_1,\mathbf{x}_2]^{\mathrm{T}}$, a and b are the weighting of each kernel (set them to be 1 here).

V. Experimental Results

The proposed speech emotion classification system was tested on the German emotional speech database (GES) [20]. The GES database consists of 800 utterances with 7 emotions, including anger, boredom, disgust, fear, joy, neutral, and sadness. Notably, the number of valid utterances is 535 since this work uses only those utterances which are voted by over 80% voters. Those utterances are 127 angry, 81 boredom, 46 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness. Finally, a half of utterances of each emotion are used for training, and the remaining 50% are used for testing.

A. Evaluation of the Number of Bases in Semi-NMF

The optimal number of bases that are appropriate for the proposed system in Semi-NMF is estimated by evaluating several dimensions of coefficients with three single kernel Gaussian processes. Figure 2 summarizes those results. The horizontal axis denotes the number of bases, and the vertical axis represents classification accuracy of the feature. The proposed features obtain the best result when the number of bases is 30. Therefore, in the following experiments, the dimension of relevant Semi-NMF coefficients is set to the number which reaches the maximum accuracy in this experiment.

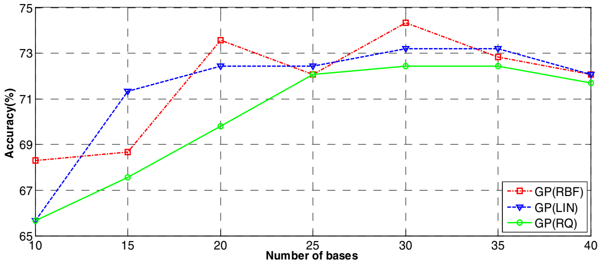

Fig. 2. Classification accuracy for different number of Semi-NMF bases. RBF: radial basis function, LIN: linear, RQ: Rational Quadratic.

B. Selection of Appropriate Kernels for Features

The proposed speech emotion classification system models different features by different kernels, and then, combines them by using a linear combination method. Therefore, the second experiment attempts to identify appropriate kernels for the features.

Table I summarizes the experimental results. The features are divided into three parts, i.e. the original feature vector, the Semi-NMF coefficients extracted from the original feature vector using Semi-NMF method (as mentioned in Section III), and the fusion feature of these two features. The features are applied to three single kernel GPs. The third part 'fusion feature' denotes the traditional directly concatenated feature when it is applied to a single kernel Gaussian process. Relatively, the third part 'fusion feature' used in the proposed speech emotion classification system represents the use of multiple kernel GP. The original feature vector is modelled using the linear kernel and the relevant Semi-NMF coefficients are modelled using the RBF kernel.

Based on Table I, we can infer the following:

- Combining those kernels which are appropriate for those features using the proposed system may allow us to obtain a better result than when only using the original feature vector with the single kernel GP.

- The direct concatenation of two feature sets may fail to achieve improved results. Moreover, the proposed feature fusion method using multiple kernel GP achieves a higher or equal accuracy than the traditional directly concatenated features.

TABLE I
ACCURACY WITH DIFFERENT KERNEL GAUSSIAN PROCESS

Features	Kernel functions			
	LIN	RBF	RQ	Proposed System
Our Proposed Features	**77.36**	67.55	67.92	**77.74** (LIN+RBF)
30 Semi-NMF Coefficients	73.21	**74.34**	72.45	
Fusion Features	75.74	72.45	72.45	

C. *Results and Discussion*

In Table II, we summarize the results among three classification systems. First, a support vector machine (SVM) with RBF kernel is applied. The last column of Table II displays the results for the proposed multiple kernel Gaussian process. The proposed system slightly improves over that of single kernel GP. The proposed emotion classification system is advantageous in that it does not require an additional feature selection procedure. Once the original feature set is selected, the corresponding Semi-NMF coefficients are obtained for use in multiple kernel GPs. Furthermore, the proposed feature performs better in single kernel GPs, thereby increasing the accuracy of the proposed system. The proposed features with the proposed system achieve the best result.

TABLE II
COMPARISON OF ACCURACY BETWEEN DIFFERENT FEATURES AND CLASSIFIERS

Features	Classifiers (kernel)	Accuracy
Our Proposed Features	SVM (RBF)	68.68
Our Proposed Features	GP (LIN)	77.36
Our Proposed Features + 30 Semi-NMF Coefficients	Proposed System	**77.74**

VI. CONCLUSION

This work presents a novel speech emotion classification system using the multiple kernel Gaussian process. During feature extraction, the proper feature set is determined first. Six frame-based features are extracted and, then, two statistics such as mean and standard deviation are calculated for each feature. Once the features are selected, the Semi-NMF algorithm is applied to them in order to extract corresponding coefficients. According to our results, these coefficient are successful for classification. Moreover, accuracy of the original feature sets is increasing using the Semi-NMF coefficients.

For classification accuracy, a Gaussian process, which is a stochastic process that models distributions over function spaces, is applied. A multiple kernel GP rather than a single kernel is used, owing to its ability to provide more flexibility and possibly preserve information from different features. Experimental results indicate how appropriate kernels are selected to model the original feature set and its corresponding Semi-NMF coefficients, respectively. According to the experimental results, the Gaussian process performs better than the traditional SVM in terms of classification accuracy. Furthermore, the feature performs better in single kernel GPs, thus achieving a higher accuracy in the proposed system.

REFERENCES

[1] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 1175–1191, Oct. 2001.

[2] H. Hu, M. X. Xu, and W. Wu, "Fusion global statistical and segmental spectral features for speech emotion recognition," in *Proc. Interspeech*, 2007, pp.1013–1016.

[3] Z. Chuang, and C. Wu, "Emotion recognition using acoustic features and textual content," in *Proc. ICME*, 2004, pp. 53–56.

[4] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Proc. Interspeech*, 2005, pp. 805–808.

[5] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[6] I. Buciu and I. Pitas, "Application of non-negative and local non negative matrix factorization to facial expression recognition," in *Proc. ICPR*, 2004, pp. 288–291.

[7] A. Mehmood, T. Damarla, and J. Sabatier, "Separation of human and animal seismic signatures using non-negative matrix factorization," *Pattern Recognit. Lett.*, vol. 33, pp. 2085–2093, Dec. 2012.

[8] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 2, Feb. 2008.

[9] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[10] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Fisher non-negative matrix factorization for learning local feature," in *Proc. ACCV*, 2004, pp. 27–30.

[11] N. E. Gillian, "Gesture recognition for musician computer interaction," Ph.D. dissertation, Faculty of Arts, Humanities and Social Sciences, School of Music and Sonic Arts, Queen's University Belfast, Belfast, County Antrim, Northern Ireland, United Kingdom, Mar. 2011.

[12] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition using HMMs fusion system with relative features," in *Proc. ICINIS*, 2008, pp. 608–611.

[13] H. K. Mishra and C. C. Sekhar, "Variational Gaussian mixture models for speech emotion recognition," in *Proc. ICAPR*, 2009, pp. 183–186.

[14] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, Nov. 1996.

[15] K. Markov and T. Matsui, "Speech and music emotion recognition using Gaussian processes," in *Modern Methodology and Applications in Spatial-Temporal Modeling*, Japan: Springer, 2015, pp. 63–85.

[16] D. Tuia, G. Camps-Valls, G. Matasci and M. Kanevski, "Learning relevant image features with multiple-kernel classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3780–3791, Oct. 2010.

[17] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[18] M. Gonen and E. Alpaydm, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, 2011.

[19] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

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

[21] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Proc. Interspeech*, 2005, pp. 493–496.

[22] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, *et al.*, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Proc. ICACII*, 2007, pp. 488–500.