

Improved Emotion Recognition with Novel Global Utterance-level Features

Yongming Huang¹, Guobao Zhang¹, Xiong Li² and Feipeng Da¹

¹School of Automation, Southeast University, Nanjing Jiangsu 210096, China

²Institute of Image Processing & Pattern Recognition Shanghai Jiao Tong University, Shanghai 200240, China

Email Address: clixiong@sjtu.edu.cn

Received June 22, 2010; Revised March 2, 2011

Traditional features, which are extracted from each frame, can not reflect the dynamic characteristics of emotion speech signal accurately. To solve this problem, first, without dividing the emotion speech into frames, novel global utterance-level features are proposed with multi-scale optimal wavelet packet decomposition; then, considering the case of little training samples, a fusion strategy through metric learning, which is called weak metric learning in this work, is proposed for fusing the global and traditional features. The experimental results with LIBSVM show that fusing the novel global feature to traditional feature achieves significant improvements about 5.2% to 13.6% than merely using local utterance-level features.

Keywords: Speech emotion recognition, Multi-scale optimal wavelet packet decomposition, Utterance-level features fusion, Weak metric learning.

1 Introduction

Speech features that are commonly used in speech emotion recognition (SER) can be divided into two categories: utterance-level features[1,6] and frame-level features[2-5]. The former usually means calculating statistical information of raw frame-based features or frame-level features, such as max, min, mean, standard deviation. It should be noticed that frame-level features will decrease emotion recognition accuracy obviously if without being processed in advance, because they carry too much linguistic information. So, in this paper, utterance-level features are extracted for SER.

Yang et al.[3] proposed a new set of harmony features for SER. They showed that an improved recognition performance by using harmony parameters in addition to state of the art features. Kim et al.[4] focused on a thinking robot. A novel speaker-independent feature, the ratio of a spectral flatness measure to a spectral center, is proposed to solve the problem of diverse interactive users. Park et al.[5] proposed a feature vector classification to improve the performance in service robots, using local frame-level features (MFCCs and prosodic features). Chandaka et al. used five local utterance-level

features extracted from speech divided into frames (Mean, STD, MAD, K, S) [6]. Since the traditional prosodic features or complicated spectral features mentioned above are almost local features extracted from speech divided into frames, which, in turn, can not reflect the dynamic characteristics of emotion speech signal accurately, there would be not conducive to build a robust SER system.

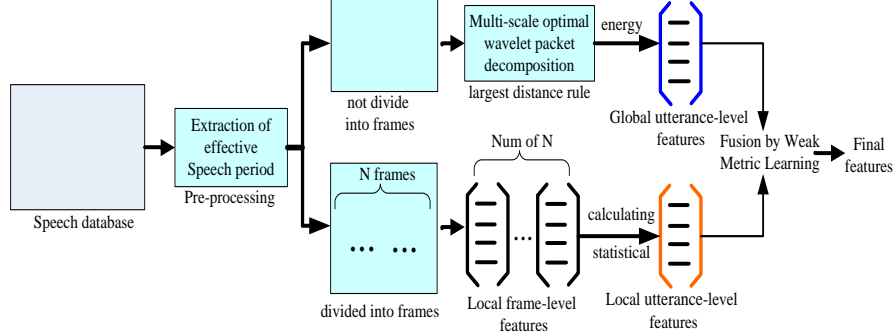


Figure 1.1: Illustration of our framework

In order to obtain a higher performance, we propose a new framework, as shown in Figure 1.1, in which global utterance-level features by multi-scale optimal wavelet packet decomposition (WPD) and local utterance-level traditional prosodic features are combined to realize emotion recognition from speech signals. However, in the process of features fusion, the problem of small training samples would appear when the number of speech training samples is less than the dimension of the extracted features. To address these problems, as illustrated in Figure 1.1, we introduce a kernel canonical correlation method to learn the metrics fusing global and local utterance-level features.

2 Global Utterance-level Features Based on Multi-scale Optimal WPD

If the binary discrete wavelet cluster $\{\psi_{j,k}(t) | j, k \in \mathbf{Z}\}$ constitute the Orthonormal basis of $L^2(\mathbf{R})$, then the Orthogonal decomposition of $x(t) \in L^2(\mathbf{R})$ can be formulated as:

$$x(t) = \sum_{j=1}^N \sum_{k \in \mathbf{Z}} d_k^j \psi_{j,k}(t) + \sum_{k \in \mathbf{Z}} c_k^N \phi_{N,k}(t), \quad (2.1)$$

where N is the decomposition level; d_k^j are the Wavelet coefficients; d_k^j are the Wavelet coefficients; c_k^N is the Scale factor of level N . Scaling function $\phi(t)$ and wavelet function $\psi(t)$ denote $u_0(t) = \phi(t)$, $u_1(t) = \psi(t)$, then the Two-scale equation can be formulated as:

$$u_{2n}(t) = \sqrt{2} \sum_{k \in \mathbf{Z}} h(k) u_n(2t - k), \quad (2.2)$$

$$u_{2n+1}(t) = \sqrt{2} \sum_{k \in \mathbb{Z}} g(k) u_n(2t - k), \quad (2.3)$$

where the defined set $\{u_n(t)\}_{n \in \mathbb{Z}}$ is the Wavelet packet which is determined by $u_0(t) = \phi(t)$; $g(k) = (-1)^k h(1 - k)$, which means the two coefficients satisfying the orthogonality relationship. Based on this, the wavelet decomposition in multiresolution analysis is extended to the wavelet packet decomposition, then the wavelet packet of emotional speech signal can be formulated as:

$$g_j^n(t) = \sum_k d_{j,n}^k u_n(2^j t - k), \quad (2.4)$$

where $d_{j,n}^k$ are coefficients from decomposition.

We use the energy got from the decomposition of the wavelet packet as the features of the speech signals and construct the eigenvectors as follows:

We make 5-level WPD of the signals and extract the 32 features of the signals S_{5j} ($j = 0, 1, \dots, 31$) from the low frequency to the high frequency at level 5 separately. Then the total signal is

$$S = S_{50} + S_{51} + \dots + S_{531}. \quad (2.5)$$

1) Getting the frequency band energy. Assuming that the energies of S_{5j} ($j = 0, 1, \dots, 31$) are E_{5j} ($j = 0, 1, \dots, 31$), then

$$E_{5j} = \int |S_{5j}(t)|^2 dt = \sum_{k=1}^n |x_{jk}|^2, \quad (2.6)$$

where x_{jk} represents the amplitude of S_{5j} .

2) Constructing the eigenvectors. The energies of different kinds of signals are different in the frequency band. Now we get the final utterance-level features using the energy

$$T = [E_{50}, E_{51}, \dots, E_{531}]. \quad (2.7)$$

3 Weak Metric Learning for Features Fusion

Metric learning is originally proposed to learn distance or similarity function by weighting each feature dimension. For a given speech, suppose similar feature elements correspond to the similar prototypes therefore they have similar metrics with similar weights. The continuous function $h \in H$ is used for assigning $w_i = h(x_i)/x_i$ to the global feature $\mathbf{x}(I) \in \mathbb{R}^m$. Then the weighed feature $\mathbf{x}'(I)$ could be formulated as:

$$\begin{aligned} \mathbf{x}'(I) &= \text{diag}(w_1, \dots, w_m) \mathbf{x}(I) \\ &= (h(x_1), \dots, h(x_m))^T \\ &= h \circ \mathbf{x}(I). \end{aligned} \quad (3.1)$$

It suggests that weighting feature with template derived weights equals to applying a nonlinear transformation on the feature. Because weights for a raw utterance-level

features are derived from the same template function h , the task of determining weight set $\{w_i\}_{i=1}^m$ is converted to determine the parameter set of the template function h . The weights are nonlinearly dependent because the number of free parameters of $\{w_i\}_{i=1}^m$ (equals to the parameter number of h) is much smaller than m . It leads to a weak learning scheme. However, with capacity increasing of the template function h , the weak metric learning scheme will approach the general metric learning.

The kernel version of canonical correlation[7] is used in the process of feature fusion, in our work, because it increases the flexibility of the feature selection through kernel trick. For the training emotion speech set \mathcal{I}_N , raw local utterance-level features $X_{N \times p} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ are calculated by means of Praat[9], and raw global utterance-level features $Y_{N \times q} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^T$ are computed by multi-scale optimal WPD. Given nonlinear transformations $g, h \in H$, the kernel canonical correlation of two weighted utterance-level features is defined as:

$$\phi(g, h, \alpha, \beta) = \text{corr}_{ker}(\alpha^T(g \circ X), \beta^T(h \circ Y)), \quad (3.2)$$

where $g \circ X$ represents applying transformation g on feature matrix X , as Eq.(3.3) formulated, and vectors $\alpha, \beta \in \mathbb{R}^N$ represent the combination coefficients of canonical correlation. We choose optimum nonlinear transformations by maximizing Eq. (3.2) stepwise

$$(g^*, h^*) = \arg \max_{g, h \in H} \widehat{\max}_{\alpha, \beta \in \mathbb{R}^N} \phi(g, h, \alpha, \beta), \quad (3.3)$$

where $\widehat{\max}_{\alpha, \beta}$ is a constrained maximizing process. We maximize Eq. (3.3) by enumerating g, h in function space H firstly. After g, h are given, we then further maximize $\phi(h, g, \alpha, \beta)$ in space \mathbb{R}^N . That is to maximize kernel canonical correlation:

$$\begin{aligned} & \max_{\alpha, \beta \in \mathbb{R}^N} \text{corr}_{ker}(\alpha^T(g \circ X), \beta^T(h \circ Y)) \\ \text{s.t. : } & \text{var}(\alpha^T(g \circ X)) = \text{var}(\beta^T(h \circ Y)) = 1 \end{aligned} \quad (3.4)$$

It can be solved using Lagrange method which leads to an eigenvalue decomposition problem. Then $\phi_{\max}(g, h)$ could be substituted into Eq. (3.3) to continue maximizing in function space. It is time consuming to enumerate function space H . A specific yet effective solving procedure is to solve the optimization problem in the parameter space of a certain function instead of in the function space. Specially, let H be a function family parameterized by $\theta \in \mathbb{R}^S$. Eq.(3.5) can be formulated as:

$$(\theta_g^*, \theta_h^*) = \arg \max_{\theta_g, \theta_h \in \mathbb{R}^S} \widehat{\max}_{\alpha, \beta \in \mathbb{R}^N} \phi(\theta_g, \theta_h, \alpha, \beta). \quad (3.5)$$

After parameter sets θ_g^* and θ_h^* are determined, two weighted global and local utterance-level features could be given by Eq. (3.1), leading to the final utterance-level features $(\mathbf{x}'(I)^T, \mathbf{y}'(I)^T)^T$. In the metric learning based fusion scheme, weighting on each feature element can be regarded as the adjusting process with feedback signals in acoustic cortex.

4 Experiments and Results

In this section, experiments were carried out on Berlin Emotional Database(EMODB)[8]. The first experiment is speaker-dependent experiment, and 10-fold cross validation is performed here. The second experiment is similar to the previous experiment but it's speaker-independent, meaning a "leaving-one-speaker-out" cross validation.

4.1 Recognition Performance of the Global Utterance-level Features Based on WPD

As the signal of emotional speech is a kind of non-stationary time and frequency signal, an algorithm based on the optimal wavelet packet basis (WPB) is proposed by using the arbitrary time-frequency decomposition of wavelet packet transform. The initial features are constituted by the filial generation energy of the WPD. Then, the best representative eigenvectors are obtained by the optimal WPB, which is chosen by the largest distance criteria.

Table 4.1: The distance between within-class sets and between-class sets for different WPB(%)

WPB	db1	dmey	db8	db10	sym3	coif1	bior1.1	bior1.3	rbio1.3	sym2
J_A	0.456	0.242	0.357	0.363	0.461	0.460	0.450	0.465	0.543	0.507

From Table4.1, where J_A is the distance between within-class sets and between-class sets, the experimental results show that J_A (0.543) of features based on WPB(rbio1.3) is the largest. So the rbio1.3 should be the optimal WPB, and the next experiments are all base on the WPB rbio1.3.

Table 4.2: Recognition performance of features base on multi-scale optimal WPD (%)

	happiness	anger	boredom	fear	sadness	neutral	disgust	Average
Speaker-dependent	45.07	73.23	64.20	50.72	58.06	43.04	60.87	56.46
Speaker-independent	40.08	66.14	56.79	43.48	48.39	36.71	52.17	49.12

As shown in Table 4.2, under speaker-dependent experiment, the utterance-level features base on multi-scale optimal WPD achieves low performance about 56.46% on the EMODB, and there is a significantly decrease in recognition rate in speaker-independent experiment. Overall, although the global utterance-level features have a certain effect to SER, it is unsuitable to use it merely. Next, we will introduce the kernel canonical correlation to learn the metrics to fuse the global utterance-level features and local utterance-level features.

4.2 Recognition Performance of the Feature Fusion Algorithm

To make a fusion, local utterance-level prosodic features were calculated by means of Praat, software for acoustic analysis[9]. First the pitch, F1, F2, F3, intensity curves were extracted from speech signals. For each curve, similarly, eight statistical features were computed, and this gives 40 prosodic features per utterance. As they are extracted from each frame, we call it local utterance-level features.

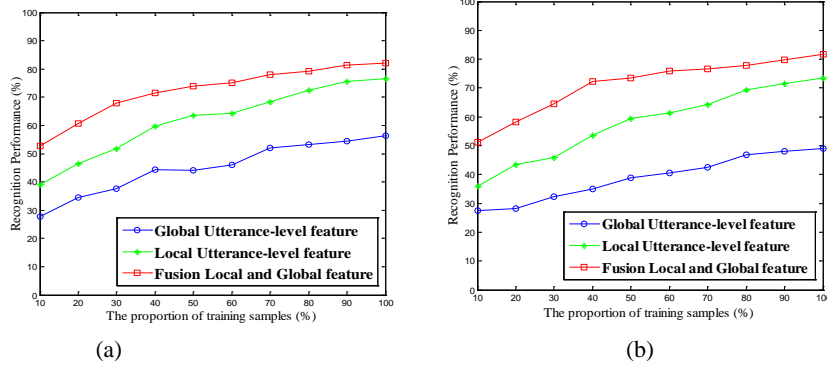


Figure 4.1: Recognition performance of the feature fusion algorithm. (a) Speaker-dependent experiment. (b) Speaker-independent experiment.

In Figure 4.1, the classification performance of local utterance-level Prosodic is better than the global utterance-level feature, and the recognition rate of speaker-dependent experiment is higher than speaker-independent experiment. Furthermore, Figure 4.1 indicates that introducing the kernel canonical correlation to learn the metrics to fuse the global features to local features achieves significant improvements than using local feature merely. As single utterance-level feature has low classification performance for training samples less than 30%, the advantage of feature fusion model seems more obvious when it comes to small sample. Overall, the Figure 4.1 reveal that our framework achieves significant improvements about 5.2% to 13.6% than using local utterance-level Prosodic merely, and classification performance is more robust to sample reduction. Consequently, in our framework, fusing the novel global utterance-level feature to local utterance-level feature can enhance the generalization ability of SER, especially for small training samples.

5 Conclusions and Discussions

In this work, we propose novel utterance-level features based on multi-scale optimal WPD, and the experimental results on EMODB show that the novel features are effective for SER. A weak metric learning algorithm is developed for high dimensional features and small samples towards constructing the feature fusion model. The metric learning based model is solved through maximal canonical correlation formulation, giving the final utterance-level features for SER towards small samples. Experiments using weak

metric learning for global and local utterance-level features fusion show an improved recognition performance. The fusion scheme, however, reaches the performance at the cost of much computing time, and it will be the further researched.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (no. 60805002), 863 program (no. 2009aa01z311).

References

- [1] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, *Speech Communication*. 52(2010), 613-625.
- [2] B.Elif, E.Engin, et al, Improving Automatic Emotion Recognition from Speech Signals, *10th Annual Conference of the International Speech Communication Association*(Brighton, United kingdom, September 6-10, 2009), 324-327.
- [3] B. Yang and M. Lugger, Emotion Recognition from Speech Signals Using New Harmony Features, *Signal Processing*. 90(2010), 1415-1423.
- [4] E. H. Kim, K. H. Hyun, S. H. Kim and Y. K. Kwak, J. Improved Emotion Recognition with a Novel Speaker-Independent Feature, *IEEE Trans. on Mechatronics*. 14(2009), 317-325.
- [5] J. S. Park, J. H. Kim and Y. H. Oh, J. Feature Vector Classification based Speech Emotion Recognition for Service Robots, *IEEE Trans. on Consumer Electronics*. 55(2009), 1590-1596.
- [6] C.Suryannarayana, C.Amitava, M.Sugata, Support vector machines employing cross-correlation for emotional speech recognition, *Measurement*. 42(2009), 611-618.
- [7] D.Hardoon, S.Szedmak, J.S.Taylor, Canonical correlation analysis: an overview with application to learning methods, *Neural Comput*. 16(2004), 2639-2664.
- [8] F.Burkhaedt, A.Paeschke, M.Rolfes, et al: A Database of German Emotional Speech, *9th European Conference on Speech Communication and Technology* (Lisbon, Portugal, September 4-8, 2005), 1517-1520.
- [9] P.Boersma, Praat - A system for doing phonetics by computer, *Glott International*. 5(2001), 341-345.



Yongming Huang received the MS degree in School of Automation from Southeast University in 2008, and then study for PhD degree in the University. His research interests are in the areas of Speech emotion recognition, Speech Processing, Speech recognition, and Facial expression recognition.