Journal of Southeast University (English Edition) Vol. 33, No. 4, pp. 437 - 443 Dec. 2017 ISSN 1003—7985

Emotion recognition of Uyghur speech using uncertain linear discriminant analysis

Tashpolat Nizamidin[1,2] Zhao Li[1] Zhang Mingyang[1] Xu Xinzhou[1] Askar Hamdulla[2]

([1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2]School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: To achieve efficient and compact low-dimensional features for speech emotion recognition, a novel feature reduction method using uncertain linear discriminant analysis is proposed. Using the same principles as for conventional linear discriminant analysis (LDA), uncertainties of the noisy or distorted input data are employed in order to estimate maximally discriminant directions. The effectiveness of the proposed uncertain LDA (ULDA) is demonstrated in the Uyghur speech emotion recognition task. The emotional features of Uyghur speech, especially, the fundamental frequency and formant, are analyzed in the collected emotional data. Then, ULDA is employed in dimensionality reduction of emotional features and better performance is achieved compared with other dimensionality reduction techniques. The speech emotion recognition of Uyghur is implemented by feeding the low-dimensional data to support vector machine (SVM) based on the proposed ULDA. The experimental results show that when employing an appropriate uncertainty estimation algorithm, uncertain LDA outperforms the conventional LDA counterpart on Uyghur speech emotion recognition.

Key words: Uyghur language; speech emotion corpus; pitch; formant; uncertain linear discriminant analysis (ULDA)
DOI: 10.3969/j.issn.1003 - 7985.2017.04.008

Speech is one of the most effective ways of human-computer interaction in the era of artificial intelligence. Therefore, in the human-computer interaction system[1], in order to make the machine understand human emotion, the identification of the emotional state in the speech signal becomes increasingly important. Speech emotion recognition (SER) involves several different fields, including speech signal processing, pattern recognition, machine learning, psychology, and so on.

For SER, it is generally regarded as the default method for capturing paralinguistic information to generate a single high-dimensional representation of an utterance from a set of underlying low-level acoustic descriptors. Previous investigations consistently demonstrate the usefulness of this technique when applied to a range of different SER problems.

Dimensionality reduction is frequently used in the preprocessing stage to make the input data more suitable for modeling. Linear discriminant analysis (LDA)[2] is one of the simplest and most popular transforms to enhance class separability for multi-dimensional observations. Conventional LDA assumes that each class follows a normal distribution and classes share the same covariance structure. Although these assumptions do not generally hold in practice, the conventional approach and its variants have been found useful in many applications including automatic speech and speaker recognition. When the dimensionality of the data becomes comparable with the number of samples per class, the sample covariance estimation becomes unstable. Regularization and Bayesian estimation of covariance models have been discussed in exiting literature to overcome this issue. It is also possible to obtain a nonlinear class separation using subclass discriminant analysis and the kernel trick in LDA. When each class is composed of several partitions, subclass discriminant analysis aims to maximize the distance between the class means and the subclass means in the same class at the same time.

Compared to the principal component analysis (PCA)[3], class-dependent dimensionality reduction is expected to be more effective in modeling classes. The extension of LDA includes heteroscedastic LDA, quadratic discriminant analysis, and mixture discriminant analysis. A distance preserving dimensionality reduction transform maps the D-dimensional data samples to a d-dimensional space ($d < D$), which means that nearby data samples are mapped to nearby low-dimensional representations. Considering K as the number of the classes in a dataset, the selection of less than $K - 1$ dimensions in LDA for data projection does not guarantee preserving the distance between classes from a classification perspective for $K > 2$.

In this paper, we address the task of finding linear discriminant directions, using a probabilistic description instead of a point-estimation for an observation. We achieve such a probabilistic description by using so-called observation uncertainties. In this approach, the feature

Received 2017-05-17, **Revised** 2017-08-30.

Biographies: Tashpolat Nizamidin (1988—), male, graduate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation item: The National Natural Science Foundation of China (No. 61673108, 61231002).

Citation: Tashpolat Nizamidin, Zhao Li, Zhang Mingyang, et al. Emotion recognition of Uyghur speech using uncertain linear discriminant analysis[J]. Journal of Southeast University (English Edition), 2017, 33(4): 437 - 443. DOI: 10.3969/j.issn.1003 - 7985.2017.04.008.

extraction process outputs the point-estimation of a feature vector along with an uncertainty. The point-estimation is assumed to form a Gaussian mean, while the corresponding variance is set as the estimated uncertainty. Throughout this paper, we note this process as an uncertain observation. Accordingly, we utilize uncertain LDA (ULDA)[4] to account for the observation uncertainties in estimating scatter matrices for LDA.

Based on the Uyghur speech emotion database, we set up a benchmark for Uyghur speech emotion recognition which involves a set of SER tasks in various training/test conditions. Additionally, the dimensionality reduction technique ULDA (uncertain linear discriminant analysis) is applied. We provide the complete data description, system architecture, experimental set-up and evaluation performance. These can be used as a full reference for Uyghur speech emotion recognition research.

1 Algorithm

Dimensionality reduction techniques are widely applied in speech emotion recognition research, such as PCA, LDA, locality preserving projections (LPP)[5], local discriminant embedding (LDE)[6], graph-based Fisher analysis (GbFA)[7] and so on. It is important to note that these methods do not solve recognition and hypothesis testing problems directly, and they are used as a pre-processing stage to reduce dimensionality. A conventional SER task requires a large number of features; hence, we should use an efficient dimensionality reduction technique to deal with this high-dimensional case. In this paper, we applied uncertain linear discriminant analysis (ULDA) to dimensionality reduction.

1.1 Conventional LDA

Let $\boldsymbol{X}=\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L$ be a set of L samples (features), each sample belonging to one of K classes and partitioning the data into clusters $C_1, C_2, \ldots, C_K$. The conventional LDA aims at finding a linear transformation of those features that can maximize the separability of the clusters. Each class is assumed to be Gaussian distributed and has the same Gaussian covariance structure. In order to find the discriminant directions, we first calculate the sample mean $\boldsymbol{\mu}$ and class mean $\boldsymbol{\mu}_k$ as

$$\boldsymbol{\mu}=\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{x}_l \tag{1}$$

$$\boldsymbol{\mu}_k=\frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{x}_l \tag{2}$$

where $|C_k|$ is the cardinality of class k. Next, the within-class $\boldsymbol{S}_{\mathrm{W}}$ and between-class scatters $\boldsymbol{S}_{\mathrm{B}}$ are given as

$$\boldsymbol{S}_{\mathrm{W}}=\sum_{k=1}^{K}\sum_{l\in c_k}|C_k|(\boldsymbol{x}_l-\boldsymbol{\mu}_k)(\boldsymbol{x}_l-\boldsymbol{\mu}_k)^{\mathrm{T}} \tag{3}$$

$$\boldsymbol{S}_{\mathrm{B}}=\sum_{k=1}^{K}(\boldsymbol{\mu}_k-\boldsymbol{\mu})(\boldsymbol{\mu}_k-\boldsymbol{\mu})^{\mathrm{T}} \tag{4}$$

The optimization problem is then solved by maximizing the Fisher-Rao criterion[1-3] as

$$\{\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2, \ldots, \hat{\boldsymbol{w}}_{K-1}\}=\underset{\boldsymbol{W}}{\operatorname{argmax}}\left\{\frac{|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{W}|}{|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{W}|}\right\} \tag{5}$$

where $\hat{\boldsymbol{w}}_i(i=1, 2, \ldots, K-1)$ is the i-th eigenvector corresponding to the i-th eigenvalue λ_i, which is obtained by solving $(\boldsymbol{S}_{\mathrm{B}}-\lambda_i\boldsymbol{S}_{\mathrm{W}})\hat{\boldsymbol{w}}_i=0$. The optimal projection matrix $\boldsymbol{W}^*$ is formed by putting d eigenvectors ($d\leqslant K-1$) corresponding to the largest eigenvalues together and the new representation of features is given by

$$\boldsymbol{Y}=\boldsymbol{W}^{*\mathrm{T}}\boldsymbol{X} \tag{6}$$

where $\boldsymbol{Y}=\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L$.

1.2 Uncertain LDA

ULDA is proposed for the case that the input data is available in the form of posterior distributions as $\underline{\boldsymbol{X}}=\underline{\boldsymbol{x}}_1, \underline{\boldsymbol{x}}_2, \ldots, \underline{\boldsymbol{x}}_L$, where each $\underline{\boldsymbol{x}}_l\in\mathbf{R}^{\mathrm{D}}$ is described by a respective probability density function $f_{\underline{\boldsymbol{x}}_l|I}(\underline{\boldsymbol{x}}_l)$ with I as the available information. The conventional LDA can deal with this type of data by only using the first-order statistics as $\boldsymbol{x}_l=\boldsymbol{\mu}_l=E(\underline{\boldsymbol{x}}_l)$ and continue to calculate between- and within-class scatter matrices $\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$. ULDA is developed in such a way that uses the second-order statistics of $\underline{\boldsymbol{x}}_l|I$, namely, $\Sigma_l=\mathrm{cov}(\underline{\boldsymbol{x}}_l)$ when calculating the expected scatter matrices $\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$. The eigenvectors calculated by employing expected scatter matrices are deemed to be more representative of variability in $\underline{\boldsymbol{X}}$ compared to those obtained by considering $\boldsymbol{X}=E(\underline{\boldsymbol{X}}|I)$. This claim is tested by the application of speech emotion recognition in this paper. In the following, we describe how to find expected scatter matrices.

For the sake of tractability, we assume that posterior distributions of observations are Gaussian as $\underline{\boldsymbol{x}}_l|I\sim\boldsymbol{N}(\boldsymbol{\mu}_l, \Sigma_l)$, which can be fully described by the first- and second-order statistics. In the following derivations, we rely on the fact that the sum of Gaussian variables is another Gaussian variable. By applying this principle, we obtain

$$\underline{\boldsymbol{\mu}}|I\sim N\left(\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mu}_l,\ \frac{1}{L^2}\sum_{l=1}^{L}\Sigma_l\right) \tag{7}$$

$$\underline{\boldsymbol{\mu}}_k\sim N\left(\frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\mu}_l,\ \frac{1}{|C_k|^2}\sum_{l\in C_k}\Sigma_l\right) \tag{8}$$

Next, we find the distribution for the sample mean deviation $\underline{\boldsymbol{\delta}}_{lk}$, and the class mean deviation $\underline{\boldsymbol{\delta}}_k$ as

$$\begin{aligned}&\underline{\boldsymbol{\delta}}_{lk}=\underline{\boldsymbol{x}}_l-\underline{\boldsymbol{\mu}}_k\Rightarrow\underline{\boldsymbol{\delta}}_{lk}|I\sim\boldsymbol{N}(E(\underline{\boldsymbol{\delta}}_{lk}),\ \mathrm{cov}(\underline{\boldsymbol{\delta}}_{lk},\ \underline{\boldsymbol{\delta}}_{lk}))\\&E(\underline{\boldsymbol{\delta}}_{lk})=\boldsymbol{\mu}_l-\frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\mu}_l\\&\mathrm{cov}(\underline{\boldsymbol{\delta}}_{lk},\ \underline{\boldsymbol{\delta}}_{lk})=\Sigma_l-\frac{2}{|C_k|}\Sigma_l+\frac{1}{|C_k|^2}\sum_{l\in C_k}\Sigma_l\\&\underline{\boldsymbol{\delta}}_k=\underline{\boldsymbol{\mu}}_k-\mu\Rightarrow\underline{\boldsymbol{\delta}}_{lk}|I\sim\boldsymbol{N}(E(\underline{\boldsymbol{\delta}}_k),\ \mathrm{cov}(\underline{\boldsymbol{\delta}}_k,\underline{\boldsymbol{\delta}}_k))\end{aligned} \tag{9}$$

$$E(\underline{\boldsymbol{\delta}}_k) = \frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\mu}_l - \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mu}_l \tag{10}$$

$$\operatorname{cov}(\underline{\boldsymbol{\delta}}_k, \underline{\boldsymbol{\delta}}_k) = \frac{1}{|C_k|^2}\sum_{l\in C_k}\Sigma_l - \frac{2}{L|C_k|}\sum_{l\in C_k}\Sigma_l + \frac{1}{L^2}\sum_{l=1}^{L}\Sigma_l$$

where we need to take into account the correlation between the mean and each sample. To calculate $\widehat{\boldsymbol{S}}_{\mathrm{W}}$ and $\widehat{\boldsymbol{S}}_{\mathrm{B}}$, we just need to apply the linearity of the expectation operator so that

$$\begin{aligned}\widehat{\boldsymbol{S}}_{\mathrm{W}} &= E\Big\{\sum_{k=1}^{K}\sum_{l\in C_k}(\underline{\boldsymbol{x}}_l - \underline{\boldsymbol{\mu}}_k)(\underline{\boldsymbol{x}}_l - \underline{\boldsymbol{\mu}}_k)^{\mathrm{T}}\Big\} = \\ &\sum_{k=1}^{K}\sum_{l\in C_k}E\{\underline{\boldsymbol{\delta}}_{lk}\underline{\boldsymbol{\delta}}_{lk}^{\mathrm{T}}\} = \\ &\sum_{k=1}^{K}\sum_{l\in C_k}\operatorname{cov}\{\underline{\boldsymbol{\delta}}_{lk}, \underline{\boldsymbol{\delta}}_{lk}\} + E\{\underline{\boldsymbol{\delta}}_{lk}\}E\{\underline{\boldsymbol{\delta}}_{lk}^{T}\}\end{aligned} \tag{11}$$

$$\begin{aligned}\widehat{\boldsymbol{S}}_{\mathrm{B}} &= E\Big\{\sum_{k=1}^{K}|C_k|(\underline{\boldsymbol{\mu}}_k - \underline{\boldsymbol{\mu}})(\underline{\boldsymbol{\mu}}_k - \underline{\boldsymbol{\mu}})^{\mathrm{T}}\Big\} = \\ &\sum_{k=1}^{K}|C_k|E(\underline{\boldsymbol{\delta}}_k, \underline{\boldsymbol{\delta}}_k^{\mathrm{T}}) = \\ &\sum_{k=1}^{K}|C_k|(\operatorname{cov}(\underline{\boldsymbol{\delta}}_k, \underline{\boldsymbol{\delta}}_k) + E\{\underline{\boldsymbol{\delta}}_k\}E\{\underline{\boldsymbol{\delta}}_k^{\mathrm{T}}\})\end{aligned} \tag{12}$$

leading to

$$\begin{aligned}\widehat{\boldsymbol{S}}_{\mathrm{W}} &= \boldsymbol{S}_{\mathrm{W}} + \sum_{k=1}^{K}\sum_{l\in C_k}\operatorname{cov}(\underline{\boldsymbol{\delta}}_{lk}, \underline{\boldsymbol{\delta}}_{lk}) = \\ &\boldsymbol{S}_{\mathrm{W}} + \sum_{k=1}^{K}\frac{|C_K| - 1}{|C_k|}\sum_{l\in C_k}\Sigma_l\end{aligned} \tag{13}$$

$$\begin{aligned}\widehat{\boldsymbol{S}}_{\mathrm{B}} &= \boldsymbol{S}_{\mathrm{B}} + \sum_{k=1}^{K}|C_k|\operatorname{cov}(\underline{\boldsymbol{\delta}}_k, \underline{\boldsymbol{\delta}}_k) = \\ &\boldsymbol{S}_{\mathrm{B}} + \sum_{k=1}^{K}\left(\frac{|C_K|}{L^2}\sum_{l=1}^{K}\Sigma_l + \frac{L - 2|C_k|}{L|C_k|}\sum_{l\in C_k}\Sigma_l\right)\end{aligned} \tag{14}$$

By removing uncertainties, i.e., setting $\Sigma_l = 0$, the posterior description of features becomes a Dirac delta function centered on x_l and hence $\widehat{\boldsymbol{S}}_{\mathrm{B}} = \boldsymbol{S}_{\mathrm{B}}$ and $\widehat{\boldsymbol{S}}_{\mathrm{W}} = \boldsymbol{S}_{\mathrm{W}}$.

By using $\widehat{\boldsymbol{S}}_{\mathrm{W}}$ and $\widehat{\boldsymbol{S}}_{\mathrm{B}}$ in Eq. (5), the ULDA transform is found to be $\widehat{\boldsymbol{W}}^*$ and the low-dimensional uncertain observations $\underline{\boldsymbol{Y}} = \underline{\boldsymbol{y}}_1, \underline{\boldsymbol{y}}_2, \ldots, \underline{\boldsymbol{y}}_L$ are obtained by Eq. (6) with $\widehat{\boldsymbol{W}}^*$ replacing $\boldsymbol{W}^*$. All $\underline{\boldsymbol{y}}_l \in \mathbf{R}^d (d < D)$ are then passed to the next steps of the classifier.

2 Uyghur Speech Emotion Database

At the INTERSPEECH conference, the Berlin Institute of Technology published a German language emotional speech database EMO-DB[8]. McGilloway et al.[9] recorded the Belfast emotional database. The Institute of Automation of Chinese Academy of Sciences published a Chinese speech emotion database CASIA. However, these related works have not focused on investigating emotion recognition in Uyghur speech. The Uyghur language is mainly used in Xinjiang, the Uyghur autonomous region of China. It is one of the most important minority languages of China. The Uyghur language belongs to the Altaic Turkic family of the west Hungarian branch, and its grammatical structure belongs to agglutinative type[10]. In view of this research gap, this paper established a Uyghur language speech emotion database (UYGSE-DB). The database has been collected and tagged with six basic emotional states, such as neutral, angry, happy, fear, sadness, and surprise.

2.1 Characteristics of Uyghur language

The Uyghur language is agglutinative and syllabic in nature, which means that each syllable contains a vowel and is surrounded by consonants. Most Uyghur syllables occur in the pattern [CC] V[CC], where consonants can vary from zero to two. A word can be formed using one or more syllable. The agglutinative nature of Uyghur results in many different patterns that create many lexical variations. Each word may consist of a stem or root, and one or more suffixes in different combinations, resulting in a very systematic but complicated morphology. Uyghur is a one letter one pronunciation language which is not common in English. For example, the alphabet "b" is not pronounced in word "debt" in English, but such a case would not appear in Uyghur. So, Uyghur has a high level of transcriptional ambiguity.

2.2 Uyghur speech emotion database

The emotional corpus is the basis of speech emotion analysis, and the quality of the database directly affects all aspects of the recognition system, such as feature extraction and emotion modeling. In this paper, we collected the Uyghur emotional speech and made an Uyghur speech emotion database. Ten native Uyghur speakers (5 female and 5 male) simulated six emotions, producing 10 Uyghur utterances (frequently used sentences in daily life) which are interpretable in all applied emotions. This database contains 600 sentences (six emotions × ten speakers × ten sentences). The collection step is shown in Fig. 1.

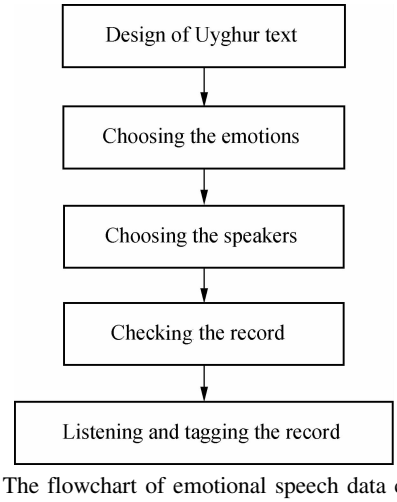

Fig. 1 The flowchart of emotional speech data collection

Our emotional corpus is collected by the acting of emotions. First, we design the non-emotional tendency text (10 sentences) as shown in Tab. 1. The UYGSE-DB was recorded by native Uyghur students with six target emotions. Considering the gap of the different gender's expression of emotion, we select 5 female and 5 male performers in the recording. The performers are twenty to thirty-five years old.

Tab. 1 Part of the Uyghur text

Text number	Uyghur sentences	English sentences
1	!چاقچاق قىلماڭلار تولا	No kidding!
2	نېمە ئىشىڭىز بارتى؟	What do you need?
3	بۈگۈن يەنە قار يېغىپتۇ.	It is snowing today.
4	بىز ياردەمگە مۇھتاج.	We need help
5	كېلىۋاتىدۇ. ئەركىن بۈيەرگە	Erkin is coming.

The collected target emotions include six basic emotions: happy, sadness, fear, surprise, anger, and neutral. These six basic emotions are widely used in other corpus, such as EMO-DB and CASIA[11]. We collected and produced the corpus with six kinds of emotion, which is conducive to the future of cross language comparison study.

We chose the recording environment in a quiet laboratory. Hardware devices for recording include a high performance computer, a SONY recording pen, a listening headset, etc. The recording set-up is single channel, 16 bit sampling accuracy, 48 kHz sampling rate and PCM encoding. In each recording, data validity was checked and recorded once again instead of the poor quality voice. In order to correct tagging of the emotion class, we chose another ten native Uyghur listeners to distinguish the emotional class of the utterances, and the average listening ear recognition rate reached 93.7% and the confusion matrix is shown in Fig. 2.

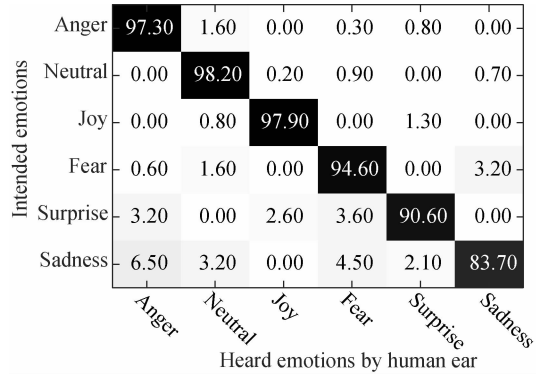

Fig. 2 Confusion matrix of Uyghur speech emotion recognition in listening experiment

3 Emotional Feature Analysis on Uyghur

In this paper, we use the Hamming window to segment the speech signal into frames. In order to maintain the smoothness between frames, we set half of the frame length for overlapping.

In the time domain, waveform can also be found in the characteristics of Uyghur speech emotions, which is shown in Fig. 3. Happy emotion speech has a short duration time and faster speech rate than neutral in Uyghur speech using the same text. The variation of sound intensity is also distinguished by the emotional state.

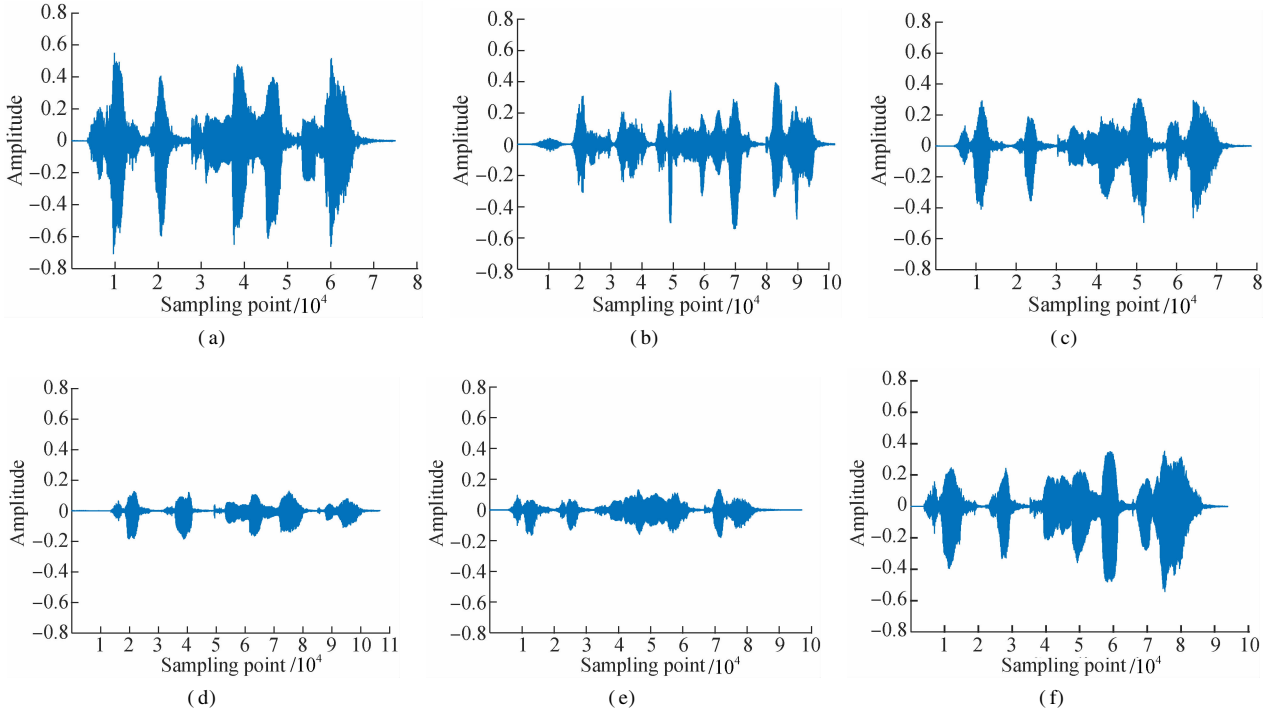

Fig. 3 Comparison of the time-domain waveform under six emotions on Uyghur. (a) Anger; (b) Fear; (c) Happy; (d) Neutral; (e) Sadness; (f) Surprise

Fundamental frequency (pitch) is one of the most important speech characteristics. F0 is related to the vocal cord and it reflects the emotional changes of human voices. We analyze the F0 feature on the same Uyghur text, as shown in Fig. 4. From the figure, we can find that the F0 feature is effective for acquiring emotional information in Uyghur speech. In addition, we fixed the contents of the Uyghur text, thus excluding the effects by phonemes. It is clearly observed from each F0 curve for different emotions that compared with neutral and sadness, the F0 curves of anger, fear, surprise and happy vary within a wide range.

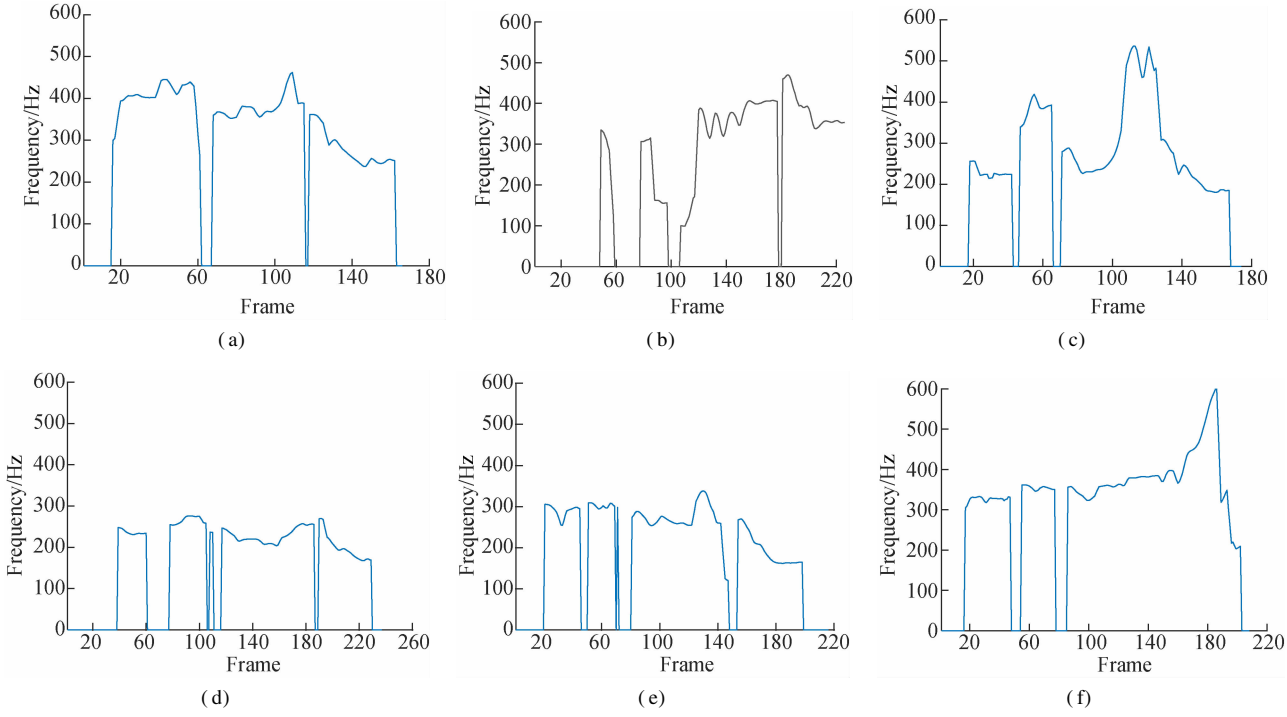

Fig. 4 The variation of the pitch contour under six emotions on Uyghur. (a) Anger; (b) Fear; (c) Happy; (d) Neutral; (e) Sadness; (f) Surprise

The formant frequency is an important acoustic feature. It is widely used in speech recognition and speaker verification. Therefore, we analyze the 1st, 2nd, 3rd and 4th formant of Uyghur emotional speech signals. Then we can perceive that each emotion has different formant distribution in Uyghur emotional speech.

In general, prosodic and formant related features can reflect the emotional variation on Uyghur language. Thus, it is a more effective feature to distinguish the emotions in speech.

However, a variety of languages and cultures will bring some differences. Referring to the domestic and foreign research on the speech emotion recognition, we use 998 features to represent the emotions in the Uyghur speech emotion corpus.

4 Experiments

Fig. 5 shows each step of the Uyghur speech emotion recognition task. First, the open-source openSMILE toolbox[12] is used for feature extraction. The features are extracted as 19 functionals and 26 acoustics low-level descriptor (LLD)[13] with the first-order difference and second-order difference, and the total number of features is 988 ($26 \times 19 \times 2$), as shown in Tab. 2 and Tab. 3.

To guarantee speaker independence, the whole data set is separated into 10 parts according to 10 speakers (five female and five male). At each time step, 4 speakers (two female and two male) are left for testing and the other 6 speakers are combined for training and form a 9-fold cross-validation[14].

The ULDA dimensionality reduction algorithm shows better performance compared with other techniques such as PCA and conventional LDA. The best average recognition accuracy on the Uyghur six emotion classification task was achieved using the ULDA + SVM algorithm, as shown in Tab. 4.

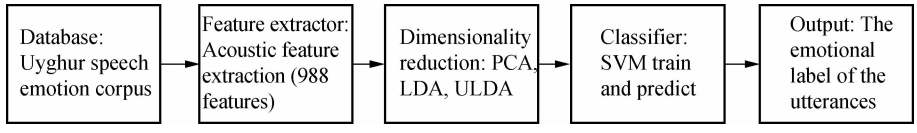

Fig. 5 Flowchart of the Uyghur speech emotion recognition system

Tab. 2 Acoustic descriptors

Descriptor	Number
Intense	1
Loudness	1
MFCC1-12	12
LSP 0-7	8
ZCR	1
Probability of voicing	1
F0	1
F0 envelope	1
Total	26

Tab. 3 Statistical functionals

Functionals	Number
Max, min, and range	3
Relative position of max and min	2
Arithmetic mean	1
Linear regression coefficients and corresponding appoximate error	4
Standard deviation, skewness, kurtosis	3
Quartiles and inter-quartile ranges	6
Total	19

Tab. 4 Average recognition accuracy on Uyghur speech emotion database

Algorithms	Accuracy/%
PCA + SVM	42.08
LDA + SVM	44.17
Proposed ULDA + SVM	48.75

5 Conclusions

In this paper, we present a Uyghur emotional database with basic emotions to analyze the emotional states in Uyghur speech. The emotional utterances are produced by Uyghur native speakers. Then, we propose a novel approach for coping with observation uncertainties in deriving an optimal linear discriminant feature transform. This so-termed uncertain LDA takes the probabilistic description of observations into account in finding the most discriminant directions. Compared to the existing algorithms, the proposed ULDA can effectively represent the emotional features in Uyghur speech. Our experiments indicate that by employing an appropriate uncertainty definition and a reliable uncertainty estimator, the Uyghur speech emotion recognition can be improved further when equipped with uncertain LDA.

References

[1] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases [J]. *Pattern Recognition*, 2011, **44**(3): 572 − 587. DOI: 10.1016/j.patcog.2010.09.020.

[2] Chu D, Liao L Z, Ng M K, et al. Incremental linear discriminant analysis: A fast algorithm and comparisons [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(11): 2716 − 2735. DOI: 10.1109/TNNLS.2015.2391201.

[3] Quan C, Wan D, Zhang B, et al. Reduce the dimensions of emotional features by principal component analysis for speech emotion recognition [C]//*Proceedings of the* 2013 *IEEE/SICE International Symposium on System Integration*. Kobe, Japan, 2013: 222 − 226. DOI: 10.1109/sii.2013.6776653.

[4] Saeidi R, Astudillo R F, Kolossa D. Uncertain LDA: Including observation uncertainties in discriminative transforms [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(7): 1479 − 1488. DOI: 10.1016/j.patcog.2010.09.020.

[5] Soldera J, Behaine C A R, Scharcanski J. Customized orthogonal locality preserving projections with soft-margin maximization for face recognition [J]. *IEEE Transactions on Instrumentation and Measurement*, 2015, **64**(9): 2417 − 2426. DOI: 10.1109/TIM.2015.2415012.

[6] Zhou Y, Peng J, Chen C L P. Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2015, **53**(2): 1082 − 1095.

[7] Li W, Du Q. Laplacian regularized collaborative graph for discriminant analysis of hyperspectral imagery [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, **54**(12): 7066 − 7076. DOI: 10.1109/tgrs.2016.2594848.

[8] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech [C]//*Proceedings of the* 2005 *INTERSPEECH*. Lisbon, Portugal, 2005: 1517 − 1520.

[9] McGilloway S, Cowie R, Douglas-Cowie E, et al. Approaching automatic recognition of emotion from voice: A rough benchmark [C]//*Proceedings of the* 2000 *ISCA Workshop on Speech and Emotion*: *A Conceptual Framework for Research*. Newcastle, Northern Ireland, UK, 2000: 207 − 212.

[10] Ablimit M, Eli M, Kawahara T. Partly supervised Uyghur morpheme segmentation [C]//*Oriental Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques Workshop*. Kyoto, Japan, 2008: 71 − 76.

[11] Pan S, Tao J, Li Y. The CASIA audio emotion recognition method for audio/visual emotion challenge 2011 [C]//*Affective Computing and Intelligent Interaction Fourth International Conference*. Memphis, USA, 2011: 388 − 395. DOI: 10.1007/978-3-642-24571-8_50.

[12] Eyben F, Wollmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor [C]//*ACM International Conference on Multimedia*. Firenze, Italy, 2010: 1459 − 1462.

[13] Xu X Z, Deng J, Zheng W M, et al. Dimensionality reduction for speech emotion features by multiscale kernels [C]//*Proceedings of Annual Conference of the International Speech Communication Association*. Dresden, Germany, 2015: 1532 − 1536.

[14] Wu S, Falk T H, Chan W Y. Automatic speech emotion recognition using modulation spectral features [J]. *Speech Communication*, 2011, **53**(5): 768 − 785. DOI: 10.1016/j.specom.2010.08.013.

基于不确定性线性判别分析的维吾尔语语音情感识别

塔什甫拉提·尼扎木丁[1,2] 赵 力[1] 张明阳[1] 徐新洲[1] 艾斯卡尔·艾木都拉[2]

([1] 东南大学水声信号处理教育部重点实验室, 南京 210096)
([2] 新疆大学信息科学与工程学院, 乌鲁木齐 830046)

摘要: 为了在语音情感识别中获得高效、紧凑的低维特征,提出了一种新的基于不确定线性判别分析的特征约简方法. 用与传统 LDA 相同的原则,在最大判别方向的估计中引入带噪声或失真输入数据的不确定性. 在维吾尔语语音情感识别任务上验证了不确定性判别分析的有效性. 在该情感数据上,分析了维吾尔语的语音情感特征,着重对维吾尔语语音的基音频率和共振峰频率进行了详细分析. 利用不确定性线性判别分析对特征维数进行了降维研究,获得了比其他的常用降维技术更好的结果. 通过不确定性线性判别分析获得的低维数据供给支持向量机,实现了维吾尔语的语音情感识别. 实验结果表明,采用适当的不确定性估计算法时,在维吾尔语音情感识别任务上,不确定性线性判别分析(ULDA)算法优于传统 LDA 降维算法.

关键词: 维吾尔语;语音情感数据库;基音频率;共振峰;不确定性线性判别分析

中图分类号: TP391