Journal of Southeast University (English Edition) Vol. 32, No. 4, pp. 402 - 407 Dec. 2016 ISSN 1003—7985

Auditory attention model based on Chirplet for cross-corpus speech emotion recognition

Zhang Xinran[1] Song Peng[2] Zha Cheng[1] Tao Huawei[1] Zhao Li[1]

([1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2]School of Computer and Control Engineering, Yantai University, Yantai 264005, China)

Abstract: To solve the problem of mismatching features in an experimental database, which is a key technique in the field of cross-corpus speech emotion recognition, an auditory attention model based on Chirplet is proposed for feature extraction. First, in order to extract the spectra features, the auditory attention model is employed for variational emotion features detection. Then, the selective attention mechanism model is proposed to extract the salient gist features which show their relation to the expected performance in cross-corpus testing. Furthermore, the Chirplet time-frequency atoms are introduced to the model. By forming a complete atom database, the Chirplet can improve the spectrum feature extraction including the amount of information. Samples from multiple databases have the characteristics of multiple components. Hereby, the Chirplet expands the scale of the feature vector in the time-frequency domain. Experimental results show that, compared to the traditional feature model, the proposed feature extraction approach with the prototypical classifier has significant improvement in cross-corpus speech recognition. In addition, the proposed method has better robustness to the inconsistent sources of the training set and the testing set.

Key words: speech emotion recognition; selective attention mechanism; spectrogram feature; cross-corpus

DOI: 10.3969/j.issn.1003-7985.2016.04.002

Speech emotion recognition (SER) can provide the most natural and fundamental interface for human-computer interaction (HCI). With the exponential growth in available computer power and the significant progress in speech technologies, the demand for expanding the generalization of emotion databases is growing in SER. The emotional feature extraction is crucial in each SER system and is focused on in this paper. The extracted features which have good robustness in various corpora, should carry sufficient information to represent the emotional states of the speakers.

Recently, several approaches have been presented for the task of SER on cross-corpus, such as hidden factor analysis[1], dimension mapping of arousal and valence[2], sparse autoencoder-based feature transfer learning[3] and multiple kernel learning scheme for the speaker-independent case[4]. The above mentioned techniques are proposed almost based on traditional linguistic features. These features are abstracted by low-level descriptors (LLD) such as Mel Frequency cepstral coefficients (MFCCs), F0, ZCR, etc. Yet, the recognition performance of SER systems based on these features is often unstable and unsatisfactory.

Different researchers may record different emotion databases according to their own research. The cross-corpus defined in this paper are mainly from the following several aspects: recording equipment, record environments, personnel, record contents and languages, etc. Therefore, when SER is performed, generally a particular speech emotion database is chosen for training and recognition, and the rest of the database is excluded. However, human perception of the ear to different emotions is not affected by external conditions. Taking various languages for example, although we may not be able to understand the specific meanings of each statement, the emotions can be judged by perception not depending on semantics. Therefore, in our research, the process of cross-databases recognition is as follows: A model is trained on a single database (or a database based hybrid corpus), and a feature vector is obtained from the training set, and then testing is performed on another database. This setting in the ideal situation is expected to achieve or approach the effect of an individual corpus.

Features based on the auditory spectrum describe the slow temporal evolution of speech. Moreover, these features emulate the perception ability of the human auditory system. Earlier studies reported that the visual saliency components from the spectrogram contain important linguistic information[5-6]. This specific observation forms the basis for the proposed feature abstractor to detect the emotional conditions of the speakers.

1 Spectrogram-Based Auditory Attention Method

A novel spectrogram feature-based auditory attention

Received 2016-05-25.

Biographies: Zhang Xinran (1980—), male, graduate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 61273266, 61231002, 61301219, 61375028), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20110092130004), the Natural Science Foundation of Shandong Province (No. ZR2014FQ016).

Citation: Zhang Xinran, Song Peng, Zha Cheng, et al. Auditory attention model based on Chirplet for cross-corpus speech emotion recognition [J]. Journal of Southeast University (English Edition), 2016, 32(4): 402 - 407. DOI: 10.3969/j.issn.1003-7985.2016.04.002.

scheme is presented. Then, an efficient selective attention cue via feature extraction from cross-corpora is adopted. The model is biologically inspired and mimics the processing stages in the human auditory system. First, the spectrogram of the input speech is computed based on the early stages of the human auditory system, which consists of cochlear filtering and inner hair cells. Then, the center surround differences of auditory salient stages are computed, which mimics the process from the basilar membrane to the cochlear nucleus in the auditory system.

Fig. 1 shows a block diagram of the proposed SER system using auditory attention spectrogram features. The block diagram computes the proposed spectrogram features which are available to determine the emotional status of a speaker's speech from diverse databases. Our motivation for the proposed method is as follows: In a speech spectrogram, we can usually find traits of gray scaled images and local discontinuities around phoneme boundaries, particularly around vowels since they exhibit high energy and a clear format structure.

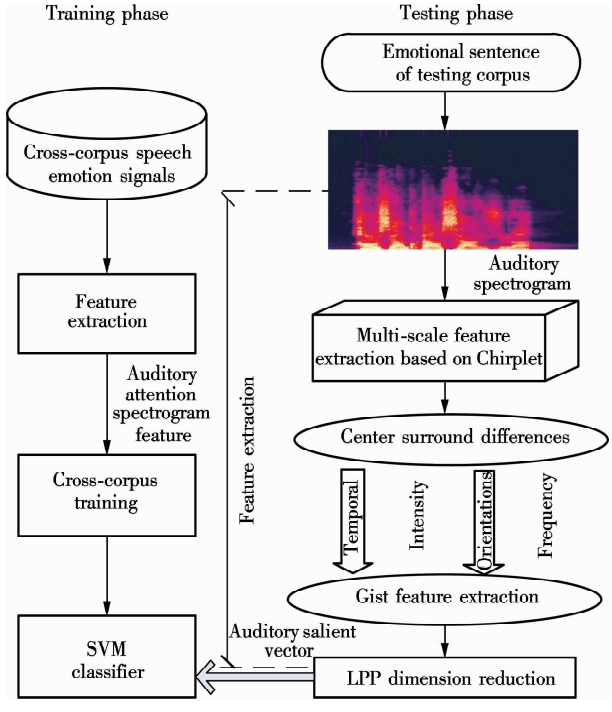

Fig. 1 Diagram of the proposed auditory attention-based SER system

In the auditory attention model, the spectrogram is analogous to an image of a scene in vision and contrast features are extracted from the spectrum in multi-scales using spectro-temporal Chirplet filters. In the feature extraction pre-processing part, first, the training set and testing set of spectrogram samples are calculated. Then, the model based on auditory attention mechanism preprocesses the spectrogram through multi-scale time-frequency atom filters. Next, the spectrogram signals are filtered and decomposed. After that, the frequency spectrum information of filters are calculated. Therein, the local orientation (direction), temporal information (time contrast feature), intensity and frequency contrast feature are united as the emotion feature vector. Through cascading these features, the pre-processing local acoustic salient vector is obtained eventually. These pre-processing features can be used to detect the formants of voices and capture them.

Next, low level auditory gist features are obtained and a locality preserving projection (LPP)[7] approach is used to discover the relevant oriented channels, which can learn the mappings between the auditory salient gist features and emotional categories from different databases. In order to make the SER system insensitive to various databases, cross-corpus training is also required in the training phase as shown in Fig. 1.

The spectrograms of speech signals are time-frequency representations, which contain abundant information from the speakers. Through the changes of tone, intonation and speed, speakers can change the form of the voices, so as to express different emotions. Fig. 2 shows the contrast of spectrograms with different speeds, tones and intonations. The figure reveals that the certain differences exist in the spectrograms with various speech expressions. For example, when speakers are angry or happy, the moods are strong, while the deep color parts in spectrograms take up larger areas. When speakers are surprised, the voice tones change more frequently and then the transverse stripe waves in corresponding spectrograms increase. When speakers feel sad, the speeds are generally slow, and then the energy spacing in each part of corresponding spectrograms may increase. Therefore, these emotional changes can cause the change in speech expression, which may consequentially result in the changes in spectrograms.

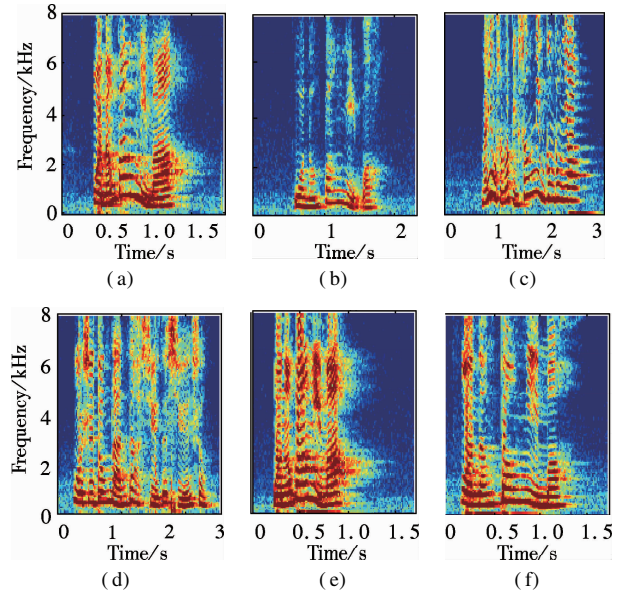

Fig. 2 Contrast of spectrograms with different speech expressions. (a) Strong tone; (b) Polite tone; (c) More intonation; (d) Less intonation; (e) Fast tempo; (f) Slow tempo

The establishment of the spectrogram emotion feature database can be realized through the stretch, modulation and translation of window function $g(t)$ ($g(t) \in L^2(R)$). This function should satisfy the following conditions:

1) $\| g \| = 1$ and it is continuously differentiable;

2) The window function $g(t)$ is a real function which satisfies $g(t) \in O((1/t^2 + 1))$;

3) $\int g(t)\,\mathrm{d}t \neq 0$ and $g(0) \neq 0$.

These characteristics make the window functions have good local features in the time domain[8], which can be described as the Gauss function: $g(t) = 2^{1/4} \mathrm{e}^{-\pi t^2}$. The Chirplet database meeting the above conditions is the collection of functions, on which the displacement, stretch and modulation are carried out. The collection can be represented as

$$g_\varphi(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) \mathrm{e}^{\mathrm{j}\xi t} \tag{1}$$

Then, the feature decomposition graph is realized by the multi-scale Chirplet filter, which is a function modulating the three-parameter Gabor for adding the frequency modulation parameters[9]. Here, u is the center of the energy concentration in the time domain and ξ is the phase shift of the wave. Benefiting from the multi-scale characters of Chirplet, this method is superior to the traditional sparse signal decomposition method wherein only a single scale is adopted. The form of the modified method is as follows:

$$g_\varphi(t) = \frac{1}{\sqrt{s}} g\left(\frac{\lambda}{s}\right) \exp[\mathrm{i}(\xi(\lambda) + 0.5c\lambda^2)] \tag{2}$$

where $\lambda = t - u$ and $\varphi = (s, u, \xi, c)$ are the frequency parameters consisting of u, ξ, the spread of pulse s and the frequency modulation slope c. Thus, compared with Gabor, the multi-scale Chirplets obtain the advantage for matching the frequency characteristic of signals by frequency parameters. Next, the decomposed feature graphs are extracted through two-dimensional direction Chirplet filters. Then, the decomposition graphs on the directional channels are obtained through the convolution of the filter and image with corresponding scale. The directional feature graphs with different scales and angles can be calculated as

$$P_\theta(\sigma) = |P_{\mathrm{I}}(\sigma) * G_0(\theta)| + |P_{\mathrm{I}}(\sigma) * G_{\pi/2}(\theta)| \tag{3}$$

In our research, the Chirplet parameters are set to be 4 scales and 6 directions. Then, the Gabor-Gauss generated process is carried out for the convolution operation with the gray images of the spectrogram. Accordingly, 24 Chirplet spectrograms are acquired and shown in Fig. 3.

In the next stage, four categories of multi-scale features

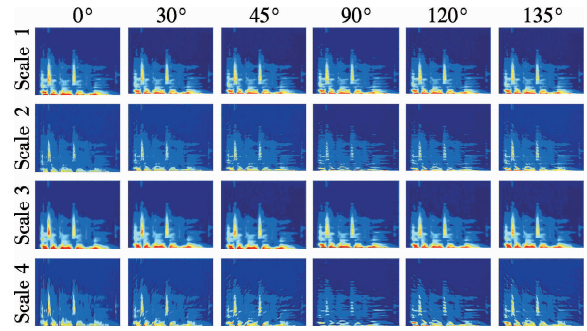

Fig. 3 Chirplet spectrograms with 4 scales and 6 directions

are extracted from the Chirplet spectrograms. This process simulates the corresponding stages of the central auditory system[10-11]. The types of features include: intensity (I), frequency (F), temporal (T), and orientation (O) with $\theta = \{0°, 30°, 45°, 90°, 120°, 135°\}$. These features are extracted by using Chirplet filters mimicking the analysis stages in the primary auditory cortex. The multi-scale emotional features reveal linguistic traits from cross-corpus. For example, F filter corresponds to the receptive fields in the primary auditory cortex with an excitatory phase and simultaneous symmetric inhibitory side bands. Each of these Chirplet filters is capable of detecting and capturing certain changes in signal characteristics. For example, F can detect and capture changes along the spectral axis, whereas O is capable of capturing and detecting moving ripples (i. e. raising and falling curves) in the salient gray scaled images. It is important that the distinction features are computed in the model, which is crucial for changed points detection and cross-corpus generalization. Next, the "auditory gist" vectors are acquired from the feature maps with I, F, T and O. After extracting a salient gist vector from each feature map, we can obtain the cumulative gist vectors by augmenting them. Finally, LPP is used to remove redundancy and to reduce the dimension before the recognition of the corpora.

2 Experiment

2.1 Experimental setup

The demonstration shown in Fig. 4 reveals that spectrogram features are suitable for emotion recognition. The normalization scatter plot compares the classification performances of visual-based intensity and pitch for sadness vs. disgust emotions. It is observed that the intensity feature which is visual-based can classify sadness and disgust emotions accurately. However, the pitch, which is an established and distinguishable feature for emotions on a single corpus, has a huge overlap between the two emotions.

Since the speech signal is non-stationary in nature, the socalled prosodic features such as pitch and energy are the local (frame-level) features which only describe a single or a few aspects of speech traits[12]. Spectrograms are

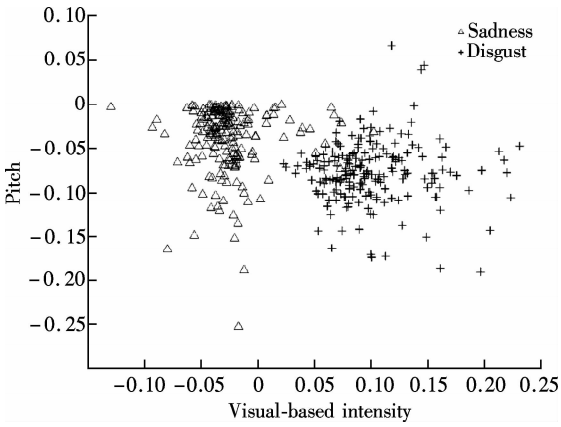

Fig. 4 Normalization scatter plot of visual-based intensity vs. pitch

union representation of time-frequency in speech signals. The analysis and the integration of spectrograms realize a global method for speech signal processing. Hence, global (utterance level) features, as the visual-based intensity of spectrograms, are calculated as statistics of all local features extracted from an utterance[13]. In addition, the auditory attention method is a type of perception mechanism based on salient cues. To this end, it is adopted in our research to improve the effectiveness of emotional features by removing invalid spectral components.

Three speech emotion datasets, including the Berlin database (EMO-DB)[14], the eNTERFACE database[15] and a Chinese emotional speech database (CNDB), are employed in the evaluation experiments. With respect to the CNDB, the speech utterances are induced using an imaging technique, noise eliciting, video clips watching and computer games. In order to maintain the quality of voice, the hearing screening experiments are conducted on students to eliminate statements with obvious quality problems. After being filtered, the speech samples are kept as 4 617 different emotional statements. For comparison with auditory attention spectrogram features, a standard feature set defined in Interspeech 2010 is employed[16]. Additionally, the openSMILE Toolkit[17] is adopted to extract the features. The Toolkit consists of 1 582 components which include 34 low-level descriptors and their corresponding first-order delta coefficients.

For cross-corpus SER, training and testing samples are required to contain accordant emotion categories. In the three databases, five kinds of emotions, i. e. anger, disgust, fear, joy, and sadness, are chosen as experimental emotional categories while the rest are eliminated. The three emotional databases follow a cross-corpus leave-one-corpus out (LOCO) strategy, i. e., one corpus is used as test set while the remaining two are used for (supervised or unsupervised) training. This strategy has shown superior classification performance in comparison with decision fusion for cross-corpus LOCO evaluation with SVM[18].

2.2 Results and analysis

Tab. 1 gives the recognition results on different databases with two feature extraction methods. Three cases are considered. In Case 1, the eNTERFACE and CNDB are used for training, whereas EMO-DB is introduced for testing. By contrast, in Case 2, the eNTERFACE dataset is used for training solely, whereas the Berlin dataset is introduced for testing. In Case 3, SER experiments are independently conducted on each single corpus for training and testing. Although certain corpora alone are insufficient to cover all cross-corpus situations, the method of experiment used in this paper can reflect the robustness of new extracted feature vectors from different sample sources, which represents the promotional performance of SER.

It can be seen from Tab. 1 that the proposed feature extraction method achieves much better performance than the traditional approach in both cross-corpus cases. The average recognition rates are 59.2% and 51.9%. Meanwhile, it can also be found that compared with the single training approach, the Case 1 experimental group with spectrogram features can improve the recognition rates significantly.

Tab. 1 Overall recognition rates of different database types with two feature extraction methods %

Feature extraction method	Cross-corpus		Single corpus (Case 3)		
	Case 1	Case 2	EMO-DB	eNTER-FACE	CNDB
Spectrogram	59.2	51.9	87.6	78.3	72.1
Traditional	49.6	43.8	88.4	73.5	70.5

From Tab. 1, it is shown that the spectrogram approach outperforms five traditional method experimental groups in three cases, which may possibly be attributed to the more sufficient robustness of cross-corpus distinction recognition than traditional acoustic features.

Fig. 5 shows the confusion matrices with the cross-corpus approach in Case 1. The SER rates of all categories of emotions are given, including disgust (D), anger (A), happiness (H), sadness (S) and fear (F). As revealed from the confusion matrix, the disgust and sadness

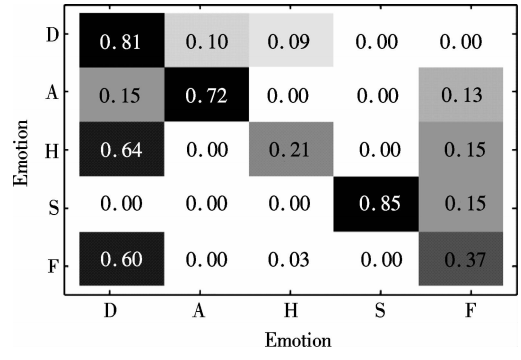

Fig. 5 Confusion matrix of cross-corpus SER in Case 1

emotions achieve the highest recognition rates. Meanwhile, the ambiguity in the classification of happiness vs. fear is responsible for major part of error in the proposed feature extraction method. This may be due to the fact that most of the acoustic features extracted for SER are related to arousal[19], but they do not discriminate between valence related emotions such as happiness and fear.

3 Conclusion

In our work, a novel cross-corpus SER approach is presented based on the auditory attention cues. The model can successfully detect salient audio emotional events in a speech scene. The changed salient events which can be captured are perceptually different from their neighbors. Then, an over-complete dictionary of atoms with multi-scale Chirplets is constructed. Accordingly, the auditory attention model is more applicable to the decomposition of non-stationary speech emotion features. The features of cross-corpus speech samples are abstracted with multi-components in which frequencies are time-varying. Experimental results on the selected cross datasets demonstrate that the proposed auditory attention method significantly outperforms the traditional standard method.

References

[1] Song P, Jin Y, Zha C, et al. Speech emotion recognition method based on hidden factor analysis[J]. *Electronics Letters*, 2014, **51**(1): 112 − 114. DOI: 10.1049/el.2014.3339.

[2] Schuller B, Zhang Z, Weninger F, et al. Synthesized speech for model training in cross-corpus recognition of human emotion[J]. *International Journal of Speech Technology*, 2012, **15**(3): 313 − 323. DOI: 10.1007/s10772-012-9158-0.

[3] Deng J, Zhang Z, Marchi E, et al. Sparse autoencoder-based feature transfer learning for speech emotion recognition[C]//*IEEE* 2013 *Humaine Association Conference on Affective Computing and Intelligent Interaction*. Geneva, Switzerland, 2013: 511 − 516. DOI: 10.1109/acii.2013.90.

[4] Jin Y, Song P, Zheng W, et al. Speaker-independent speech emotion recognition based on two-layer multiple-kernel learning[J]. *IEICE Transactions on Information and Systems*, 2013, **96**(10): 2286 − 2289. DOI: 10.1587/transinf.e96.d.2286.

[5] Kalinli O, Narayanan S. Prominence detection using auditory attention cues and task-dependent high level information[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, **17**(5): 1009 − 1024. DOI: 10.1109/tasl.2009.2014795.

[6] Kalinli O. Syllable segmentation of continuous speech using auditory attention cues[C]//*International Speech and Communication Association*. Florence, Italy, 2011: 425 − 428.

[7] Wong W K, Zhao H T. Supervised optimal locality preserving projection[J]. *Pattern Recognition*, 2012, **45**(1): 186 − 197. DOI: 10.1016/j.patcog.2011.05.014.

[8] Yin Q, Qian S, Feng A. A fast refinement for adaptive Gaussian chirplet decomposition[J]. *IEEE Transactions on Signal Processing*, 2002, **50**(6): 1298 − 1306. DOI: 10.1109/tsp.2002.1003055.

[9] Bayram I. An analytic wavelet transform with a flexible time-frequency covering[J]. *IEEE Transactions on Signal Processing*, 2013, **61**(5): 1131 − 1142. DOI: 10.1109/tsp.2012.2232655.

[10] Noriega G. A neural model to study sensory abnormalities and multisensory effects in autism[J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2015, **23**(2): 199 − 209. DOI: 10.1109/TNSRE.2014.2363775.

[11] Khoubrouy S A, Panahi I M S, Hansen J H L. Howling detection in hearing aids based on generalized teager-kaiser operator[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(1): 154 − 161. DOI: 10.1109/taslp.2014.2377575.

[12] Ali S A, Khan A, Bashir N. Analyzing the impact of prosodic feature (pitch) on learning classifiers for speech emotion corpus[J]. *International Journal of Information Technology and Computer Science*, 2015, **7**(2): 54 − 59. DOI: 10.5815/ijitcs.2015.02.07.

[13] Ajmera P K, Jadhav D V, Holambe R S. Text-independent speaker identification using radon and discrete cosine transforms based features from speech spectrogram[J]. *Pattern Recognition*, 2011, **44**(10): 2749 − 2759. DOI: 10.1016/j.patcog.2011.04.009.

[14] Burkhardt F, Paeschke A, Rolfes M, et al. A database of german emotional speech[C]//*International Speech and Communication Association*. Lisbon, Portugal, 2005: 1517 − 1520.

[15] Martin O, Kotsia I, Macq B, et al. The enterface'05 audio-visual emotion database[C]//*IEEE* 22*nd International Conference on Data Engineering Workshops*. San Francisco, CA, USA, 2006: 8 − 10. DOI: 10.1109/icdew.2006.145.

[16] Schuller B, Steidl S, Batliner A, et al. The interspeech 2010 paralinguistic challenge: Deception, sincerity and native language[C]//I*nternational Speech and Communication Association*. Chiba, Japan, 2010: 2794 − 2797. DOI: 10.21437/interspeech.2016-129.

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

[18] Moustakidis S, Mallinis G, Koutsias N, et al. SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2012, **50**(1): 149 − 169. DOI: 10.1109/TGRS.2011.2159726.

[19] Kim E H, Hyun K H, Kim S H, et al. Improved emotion recognition with a novel speaker-independent feature[J]. *IEEE/ASME Transactions on Mechatronics*, 2009, **14**(3): 317 − 325.

用于跨库语音情感识别的时频原子听觉注意模型

张昕然[1] 宋 鹏[2] 查 诚[1] 陶华伟[1] 赵 力[1]

([1] 东南大学水声信号处理教育部重点实验室, 南京 210096)
([2] 烟台大学计算机与控制工程学院, 烟台 264005)

摘要: 为解决跨数据库语音情感识别领域中实验数据集特征不匹配的问题,提出一种基于时频原子的听觉注意特征提取模型. 首先,为了提取频谱特征,引入听觉注意模型对多类情感特征进行有效的探测. 然后,利用选择注意机制改进了提取的语谱图特征,其中包含的显著性信息与跨库识别性能有紧密联系. 再引入 Chirplet 时频原子,通过形成的过完备原子库提高语谱图特征的信息量. 来自多个数据库的样本具有多成分分布的特征,据此所提模型中的 Chirplet 扩大了特征向量在时频域上的尺度. 实验结果显示,相比传统特征模型,所提方法性能有显著提升. 此外,该方法在训练集和测试集来源不一致情况下具有更好的鲁棒性.

关键词: 语音情感识别;选择性注意机制;语谱图特征;跨数据库

中图分类号: TN912. 34