J Inf Process Syst, Vol.14, No.1, pp.218~227, February 2018
https://doi.org/10.3745/JIPS.04.0063

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Vocal Effort Detection Based on Spectral Information Entropy Feature and Model Fusion

Hao Chao*, Bao-Yun Lu*, Yong-Li Liu*, and Hui-Lai Zhi*

Abstract

Vocal effort detection is important for both robust speech recognition and speaker recognition. In this paper, the spectral information entropy feature which contains more salient information regarding the vocal effort level is firstly proposed. Then, the model fusion method based on complementary model is presented to recognize vocal effort level. Experiments are conducted on isolated words test set, and the results show the spectral information entropy has the best performance among the three kinds of features. Meanwhile, the recognition accuracy of all vocal effort levels reaches 81.6%. Thus, potential of the proposed method is demonstrated.

Keywords

Gaussian Mixture Model, Model Fusion, Multilayer Perceptron, Spectral Information Entropy, Support Vector Machine, Vocal Effort

1. Introduction

Vocal effort (VE) was characterized as "the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance" [1]. Generally, there are five different VE levels: whispered, soft, normal, loud, and shouted. Changes in VE result in a fundamental change in speech production and then cause the change of acoustic characteristics, which will reduce the accuracy of speech recognition system [2,3]. Therefore, accurate VE detection can enlarge the application range of speech recognition technology, and will promote the practicability of speech recognition. In addition, it also has a positive effect on speaker recognition and speech synthesis [4-7].

It is important for VE detection to find salient information regarding the VE level, and obtain the features which are sensitive to VE change. Because the vocal cords are almost not vibrating when pronouncing, the whispered speech is obviously different in speech production mechanism and acoustic characteristic from the other VE levels. Therefore, as a typical representative of VE, related studies of whisper have been conducted since the 1960s, and the accuracy of whisper detection is satisfactory [8,9]. The average energy ratio between high-energy segment and low-energy segment of low-frequency band is acquired, and the ratio is used as a basis for the judgment of a whisper speech or

Manuscript received March 3, 2017; first revision April 21, 2017; accepted May 29, 2017.
Corresponding Author: Hao Chao (chaohao1981@163.com)
* School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China (chaohao1981@163.com, {Lubaoyun, liuyongli, zhihuilai}@hpu.edu.cn)

a normal voice [10]. Zhang and Hansen [11] proposed a detection method of vocal effort change points. In [12], a whisper detection algorithm is proposed using features obtained from waveform energy and wave period. In addition, an accurate whispered speech detection method, which uses auditory-inspired modulation spectral-based features to separate speech from environment-based components, is proposed [13].

For the remaining four VE levels, there are no significant differences in the way of pronunciation, and no significant changes have been reflected in the spectrum. For the two adjacent VE levels, it is even more so. Therefore, just a few studies collectively consider detection of all five speech levels, and only limited performances are provided. In [2,14], spectrum features including sound intensity level, sentence duration, frame energy distribution and spectral tilt are extracted to recognize all VE levels. Experimental results show that the spectrum features have a strong ability to distinguish whisper mode, but have a bad performance when detecting the other VE modes. In [3], a VE classification method using Support Vector Machine (SVM) is proposed based on the Mel-frequency cepstral coefficients (MFCC), and achieves. In addition, a detection method integrating spectrum features and MFCCs is proposed for the identification of VE levels in robust speech recognition [15]. Compared with the spectral features, MFCCs show a stronger ability to distinguish all VE levels. Nevertheless, MFCCs are, after all, proposed specially for speech recognition because they contain salient information regarding speech content rather than regarding the VE level. Thus, MFCCs have limited potential in VE detection.

In order to further improve the detection accuracy of all five VE levels, this paper proposes the spectral information entropy (SIE) which shows a stronger ability in distinguish all VE levels than MFCC and spectral features. On the basis that the spectrum features, MFCCs and SIE describe the speech signal from different aspects and the salient information regarding the VE level they have is not completely overlapping, the three features should be complementary in VE detection. Therefore, this paper proposes a VE detection method based on model fusion, and the method integrates the three features effectively.

This paper is organized as follows. In Section 2, the introduction of spectral information entropy is given. Section 3 introduces the model fusion method based on complementary models. The performance of the proposed method is reported in Section 4. The last Section 5 briefly concludes the work.

2. Spectral Information Entropy

It is important for VE detection to find salient information regarding the VE level, and obtain the features which are sensitive to VE change. In view of the presented drawbacks of the global spectrum features, frame based features, which are able to capture small differences of acoustic properties, are introduced. In addition, MFCC feature is proposed for speech recognition, so this feature mainly reflects acoustic properties caused by different pronunciation instead of VE change. To accurately detect all VE levels, the spectral information entropy feature which contains more salient information regarding the VE level is proposed.

2.1 Feature Extraction

For each frame, the spectrum obtained from FFT can be viewed as a vector of coefficients in an

orthonormal basis. Hence, the probability density function (pdf) can be estimated by the normalization over all frequency components. The spectral information entropy can be obtained from this estimated pdf.

Each frame is evenly divided into 6 sub-frames. The SIE of each sub-frame is calculated to form a 6-dimension SIE feature for each frame. In fact, the 6 dimensions of the feature are the spectral information entropy of the 6 sub-bands evenly divided over the frequency range 0–4,000 Hz, respectively. The 6 bands and their range of frequency domain are shown in Table 1.

Table 1. Six bands and their range of frequency domain

Sub-band	Frequency (kHz)
1	0.0–0.8
2	0.6–1.5
3	1.2–2.0
4	1.8–2.6
5	2.4–3.2
6	3.0–4.0

For each sub-band, the spectral information entropy can be obtained as follows:

Assuming $X(k)$ is the power spectrum of speech frame $x(n)$, k varies from k_1 to k_M in a sub-band; then that portion of the frequency content in k band versus the entire response is written as,

$$p(k) = \frac{|X(k)|^2}{\sum_{j=k_1}^{k_M} |X(j)|^2}, \quad k = k_1, \ldots, k_M \tag{1}$$

Since $\sum_{k=k_1}^{k_M} p(k) = 1$, $p(k)$ has the property of probability. The spectral information entropy for the sub-band can be calculated as,

$$H = -\sum_{k=k_1}^{k_M} p(k) \cdot \log p(k) \tag{2}$$

Using the power spectrum of each frame, the above calculation is performed for each of 6 sub-bands, so the 6-D SIE over the frequency domain is obtained.

2.2 Salient Information Analysis of SIE

From the perspective of speech perception, speech signals are composed of vowels, consonants and silent segments. Obviously, silent segment does not contain salient information regarding the VE level. So we only need to know which contains more salient information between vowel and consonant.

In order to facilitate the analysis, it can be assumed that the speech signal, which shows greater spectrum change when VE level changes, contain more salient information regarding the VE level. For this purpose, an Euclidean distance-based cepstral distance measure D_C is used.

$$D_C = \sqrt{\sum_{i=1}^{N} (c_p^{VE(1)}(i) - c_p^{VE(2)}(i))^2} \tag{3}$$

where N is the number of SIEs and $c_p^{VE(j)}(i)$ represents i^{th} SIE coefficient belonging to a phoneme p at vocal effort level VEj. An average distance between all pairs of VE levels for a given phoneme was then computed. After normalization, the obtained average distances for all phonemes are documented in Fig. 1 (sorted in descending order). The highest average distances were obtained for the set of 5 vowels (/a/, /e/, /o/, /i/, /u/) and consonants /j/, /g/, and /y/. These consonants appear less frequently in words, hence the vowels are the best candidates for VE classification.

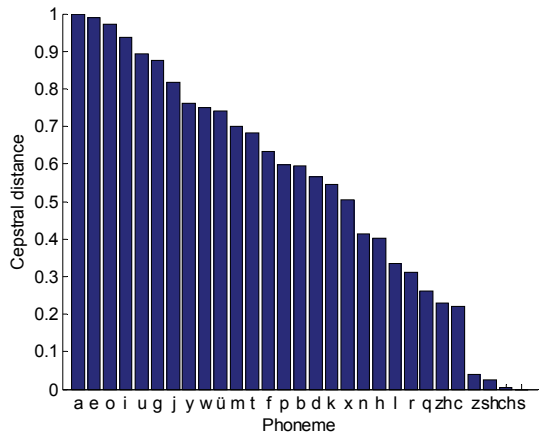

Fig. 1. Sorted average cepstral distances among the 5 VE levels for all phonemes.

The average cepstral distances using MFCC features are also acquired in this paper, and the average cepstral distances are compared with the average cepstral distances of SIE. After the analysis above, we only compare the five Chinese vowels: /a/, /e/, /o/, /i/, and /u/. As shown in Fig. 2, when using SIE features, the average spectral distance of each vowel is higher than that of the MFCC feature. This seems to indicate that SIE contain more salient information regarding the VE level.

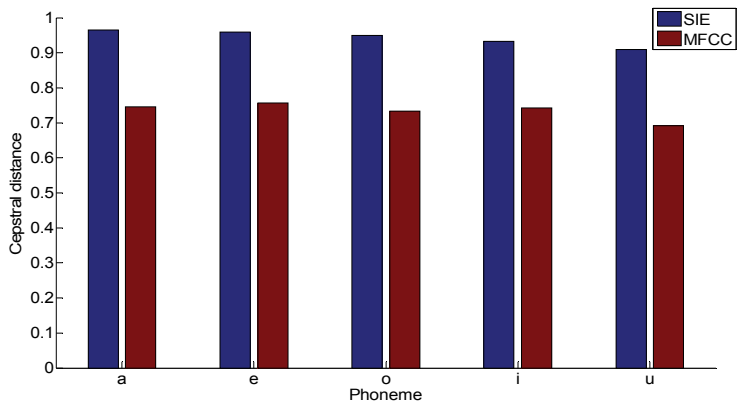

Fig. 2. Comparison of average cepstral distances between MFCC and SIE for the five vowels.

It is important to keep in mind that the speech samples grouped into the individual VE levels are not results of some artificial signal classification but a genuine representation of what the speakers considered to be "whispering", "soft speech", "normal speech", etc. The histograms represent a measurable physical quantitywhich can be used to measure the difference between different VE levels.

3. Vocal Effort Detection Based on Model Fusion

Section 2.2 has shown that the simple vowels (/a/, /e/, /o/, /i/, /u/) contain more salient information regarding VE level than other phonemes, so they are extracted from speech signal for VE level detection. The simple vowels can be obtained by manual segmentation or vowel endpoint detection. If the simple vowel sequence in sentence S is $\{v_1,v_2,...,v_n\}$, Eq. (4) can be obtained,

$$S^{VE} = \{v_1^{VE}, v_2^{VE}, \cdots, v_n^{VE}\} \tag{4}$$

where S^{VE} represents that the vocal effort level of S is VE, and v_i^{VE} represents that the VE level of the i^{th} simple vowel is VE. The Eq. (4) means that the VE levels of all simple vowels in S are also VE if the VE level of S is VE, and the converse is also true. Thus, the most likely VE level of S is $S^{VE*} = \{v_1^{VE*}, v_2^{VE*}, \cdots, v_n^{VE*}\}$, and

$$S^{VE*} = \arg\max\ p(S^{VE} \mid F, M, I) \tag{5}$$

where $F = \{f_1,f_2,...,f_n\}$ is the sequence of spectrum feature, $M = \{m_1,m_2,...,m_n\}$ is the sequence of MFCC feature, and $I = \{i_1,i_2,...,i_n\}$ is the sequence of SIE feature. The Eq. (5) can be transformed to Eq. (6),

$$\begin{aligned} S^{VE*} &= \arg\max\ p(S^{VE} \mid F, M, I) \\ &= \arg\max\ p(S^{VE} \mid F)\, p(S^{VE} \mid M)\, p(S^{VE} \mid I) \\ &= \arg\max \prod_{t=1}^{n} p(v_t^{VE} \mid f_t)^{\alpha}\, p(v_t^{VE} \mid m_t)^{\beta}\, p(v_t^{VE} \mid i_t) \\ &= \arg\max \alpha \sum_{t=1}^{n} \log\left(p(v_t^{VE} \mid f_t)\right) + \beta \sum_{t=1}^{n} \log\left(p(v_t^{VE} \mid m_t)\right) + \sum_{t=1}^{n} \log\left(p(v_t^{VE} \mid i_t)\right) \end{aligned} \tag{6}$$

where $\log(p(v_t^{VE} \mid f_t))$ is the spectrum-VE model score, $\log(p(v_t^{VE} \mid m_t))$ is the MFCC-VE model score, $\log(p(v_t^{VE} \mid i_t))$ is the SIE-VE model score. α and β are weight coefficients between three models.

Premise condition of the Eq. (6) is that F, M and I are independent. However, the premise condition that the spectrum feature, the MFCC feature and the SIE feature are independent is not established. In order to reduce the computational complexity, $p(S^{VE} \mid F, M, I)$ is generally simplified as $p(S^{VE} \mid F)P(S^{VE} \mid M)P(S^{VE} \mid I)$, but this means sacrificing the detection accuracy to some extent.

Instead of above simplified calculation method, another model fusion method which does not rely on the hypothesis that the spectrum feature, the MFCC feature and the SIE feature are independent is proposed. And the Eq. (5) can be transformed as:

$$\begin{aligned} S^{VE*} &= \arg\max p(S^{VE} \mid F, M, I) \\ &= \arg\max(\lambda \cdot p(S^{VE} \mid F, M, I) + (1-\lambda) \cdot p(S^{VE} \mid F, M, I)) \\ &= \arg\max(\lambda \cdot p_1(S^{VE} \mid F, M, I) + (1-\lambda) \cdot p_2(S^{VE} \mid F, M, I)) \end{aligned} \tag{7}$$

Eq. (7) is only deformation of Eq. (5). $\lambda \cdot p(S^{VE} \mid F, M, I)$ is given a new symbol $\lambda \cdot p_1(S^{VE} \mid F, M, I)$, and $(1-\lambda) \cdot p(S^{VE} \mid F, M, I)$ is given another new symbol $(1-\lambda) \cdot p_2(S^{VE} \mid F, M, I)$. If we use the same method to model $p_1()$ and $p_2()$, the Eq. (7) can be achieved by a traditional machine learning method such as Gaussian mixture model (GMM), SVM, artificial neural network (ANN), etc. If we don't use the same method to model $p_1()$ and $p_2()$, and adopt the hypothesis that the spectrum feature, the MFCC feature and the SIE feature are independent, the Eq. (7) can be written as Eq. (6). If different methods are used to model $p_1()$ and $p_2()$, and abandon the hypothesis that the spectrum feature, the MFCC feature and the SIE feature are independent, a new model fusion method is obtained to recognize VE level. Thus, the Eq. (7) can be transformed into Eq. (8):

$$\begin{aligned} S^{VE*} &= \arg\max p(S^{VE} \mid F, M, I) \\ &= \arg\max(\lambda \cdot p_1(S^{VE} \mid F, M, I) + (1-\lambda) \cdot p_2(S^{VE} \mid F, M, I)) \\ &= \arg\max(\frac{\lambda}{(1-\lambda)} \cdot p_1(S^{VE} \mid F, M, I) + p_2(S^{VE} \mid F, M, I)) \\ &= \arg\max(\gamma \cdot p_1(S^{VE} \mid F, M, I) + p_2(S^{VE} \mid F, M, I)) \end{aligned} \tag{8}$$

Two different machine learning methods can be used to model $p_1()$ and $p_2()$ separately, and the two different models are complementary to some extent. The idea of complement is from the observation that confusion occurs in different systems. The complement means that one model can perform VE detection in one way, and the other can also do it in another way. The distributions of their results are overlapping partially, ant not totally same. Thus, there exists complement of effect between them. By means of the model fusion, the spectrum feature, the MFCC feature and the SIE feature are integrated effectively to detect VE level.

4. Experimental Results and Analysis

4.1 Speech Corpora

The data corpus applied in experiments consists of 25000 Mandarin isolated digits (0–9). Twenty male speakers are employed for train set and test set. In the train set, each VE level contains 4000 digits, and each speaker records the digits (0–9) 20 times. In the test set, each VE level contains 1000 digits, and each speaker records the digits (0–9) 5 times. The data corpus is recorded in the laboratory environment, and is stored using the 16 kHz sampling rate and 16-bit resolution.

4.2 Experimental Setup and Result Analysis

In order to find out which kind of feature has more advantages in VE detection, a single type of feature is employed for VE detection, and the spectrum feature, the MFCC feature and the SIE feature are used in turn. This means that $\log(p(v_t^{VE} \mid f_t))$, $\log(p(v_t^{VE} \mid m_t))$, and $\log(p(v_t^{VE} \mid i_t))$ in Eq. (7) are used separately for VE detection. The spectrum feature includes sound intensity level, vowel duration, frame energy distribution and spectral tilt which are introduced in [2]. GMM, SVM, and multilayer perceptron (MLP) are selected as detection models. The MLP model has one hidden layer, and the number of hidden nodes is 2*N*+1, where *N* is the number of input nodes. LibSVM is used to train the SVM model [16]. The number of mixture components in a GMM is 128, and the diagonal covariance matrices are adopted. The detection results can be seen in Table 2.

Table 2. Two-stage VE detection results

Method	Feature type	Detection result (%)				
		Whisper	Soft	Normal	Loud	Shouted
GMM	Spectrum	92.4	56.7	51.7	57.2	62.4
	MFCC	90.7	72.9	64.6	68.2	74.7
	SIE	92.6	75.1	68.5	71.6	77.5
MLP	Spectrum	93.4	58.4	53.2	59.0	63.9
	MFCC	91.6	73.7	65.8	70.0	76.1
	SIE	93.5	75.8	69.2	72.9	79.3
SVM	Spectrum	94.2	59.7	54.4	59.3	64.7
	MFCC	93.3	75.5	67.2	71.7	77.5
	SIE	94.2	77.2	70.6	74.3	80.4

As can be seen in Table 2, no matter which model is used, the performance when using the SIE feature is the best. And the performance of the spectrum feature is close to the performance of SIE when judging whisper mode. The results indicate the SIE feature proposed is more sensitive to VE change than the spectrum feature and the MFCC feature. Meanwhile, the spectrum feature can provide sufficient salient information regarding whisper level. In addition, the performance of SVM is better than GMM and MLP.

Table 3 shows the performance of various combined models which are integrated according to Eq. (6).The value of α in Eq. (6) ranging from 0.15 to 0.3 and the value of β in Eq. (6) ranging from 0.2 to 0.4 have good effect, and can fuse the detection results of the spectrum-VE model (GMM), MFCC-VE model (MLP), and SIE-VE model (SVM). The GMM model is obtained by the spectrum feature. The MLP model is obtained by the MFCC feature, and The SVM model is obtained by the SIE feature. From Table 3, the combination of different features by the way described in Eq. (6) obtains better performance than each alone for all classifiers.

Table 3. The performance of combined models

	Detection result (%)				
	Whisper	Soft	Normal	Loud	Shouted
GMM/MLP/SVM	94.6	78.6	72.3	76.0	81.5

Finally, the proposed method by weighting combination of two classifiers according to Eq. (8) is used for VE detection, and the performance is shown in Table 4. Different from Table 3, all models in Table 4 (GMM*, MLP*, SVM*) are obtained by using the spectrum feature, the MFCC feature and the SIE feature together. The value of γ in Eq. (8) is 1.

Table 4. The performance of complementary model

	Detection result (%)				
	Whisper	**Soft**	**Normal**	**Loud**	**Shouted**
GMM*/MLP*	94.2	78.4	71.9	75.7	81.4
MLP*/SVM*	94.9	79.0	72.8	76.5	81.9
GMM*/SVM*	95.5	79.7	73.1	77.4	82.3

Table 4 shows that both the complementary model MLP*/SVM* and the complementary model GMM*/SVM* can achieve better performance than the combined models in Table 3. This means the complementary model based model fusion approach described in Eq. (8) can better integrate different features than the model fusion approach described in Eq. (6). Performance of the complementary model GMM*/MLP* is slightly worse than the combined models in Table 3. The possible reason is that SVM, which has shown strong classification ability, is not used.

5. Conclusion

In this paper, the spectral information entropy feature which contains more salient information regarding the VE level is firstly presented. After analyzing the sensitivity of global spectrum features, MFCC and SIE to the change of VE level, we proposed the model fusion method based on complementary model and yield 81.6% average VE detection accuracy rate.

The future research will be focused on a more precise detection of VE level considering real-world situations (i.e., including additive noise).

Acknowledgement

This paper is supported in part by the China National Nature Science Foundation (No. 61502150, 61300124, and 61403128), Foundation for University Key Teacher by Henan Province (No. 2015GGJS-068) and the Fundamental Research Funds for the Universities of Henan Province.

References

[1] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438-3451, 2000.

[2] P. Zelinka and M. Sigmund, "Automatic vocal effort detection for reliable speech recognition," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, Kittila, Finland, 2010, pp. 349-354.

[3] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732-742, 2012.

[4] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proceedings of 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, 2008, pp. 609-612.

[5] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech" in *Proceedings of 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013, pp. 1544-1548.

[6] D. S. Brungart, K. R. Scott, and B. D. Simpson, "The influence of vocal effort on human speaker identification," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001, pp. 747-750.

[7] R. Saeidi, P. Alku, and T. Backstrom, "Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 42-53, 2016.

[8] S. T. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263-274, 2008.

[9] S. Ghaffarzadegan, H. Boril, and J. H. Hansen, "UT-VOCAL EFFORT II: analysis and constrained-lexicon recognition of whispered speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 2544-2548.

[10] S. J. Wenndt, E. J. Cupples, and R. M. Floyd, "A study on the classification of whispered and normally phonated speech," in *Proceedings of 7th International Conference on Spoken Language Processing*, Denver, CO, 2002, pp. 649-652.

[11] C. Zhang and J. H. Hansen, "Advancements in whisper-island detection within normally phonated audio streams," in *Proceedings of 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, 2009, pp. 860-863.

[12] M. A. Carlin, B. Y. Smolenski, and S. J. Wenndt, "Unsupervised detection of whispered speech in the presence of normal phonation," in *Proceedings of 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, 2006, pp. 1-4.

[13] M. Sarria-Paja and T. H. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 783-786, 2013.

[14] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Proceedings of 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, 2007, pp. 2289-2292.

[15] H. Chao, C. Song, and Z. Z. Liu, "Multi-level detection of vocal effort based on vowel template matching," *Journal of Beijing University of Posts and Telecommunications*, vol. 39, no. 4, pp. 98-102, 2016.

[16] C. C. Chang and C. J. Lin, "LibSVM: a Library for Support Vector Machines," 2016 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Hao Chao http://orcid.org/0000-0001-6700-9446

He received his Ph.D. degree in pattern recognition and intelligent system from Institute of Auotmaition, Chinese Academy of Sciences in June 2012. He is currently a lecturer in Henan Polytechinic University. His current research interests include speech signal processing and data mining.

Bao-Yun Lu

Shereceived her Ph.D. degree in pattern recognition and intelligent system from Institute of Auotmaition, Chinese Academy of Sciences in June 2011. She is currently a lecturer in Henan Polytechinic University. Her current research interests include speech signal processing and data mining.

Yong-Li Liu

He received his Ph.D. degree in computer science and engineering from Beihang Unversity in 2010. He is currently an associate professor in Henan Polytechnic University. His current research interests include data mining and information retrieval..

Hui-Lai Zhi

He received his Ph.D. degree in computer application technology from Shanghai University in June 2010. He is currently a lecturer in Henan Polytechinic University. He current research interests are in knowledge representation and processing and signal processing.