I.J. Image, Graphics and Signal Processing, **2016, 4, 39-45**
Published Online April 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2016.04.05

Speech Emotion Recognition based on SVM as Both Feature Selector and Classifier

Amirreza Shirani
Department of Computer Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
Email: shiraniamirreza@gmail.com / a.shirani@eng.ui.ac.ir

Ahmad Reza Naghsh Nilchi
Department of Computer Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
Email: nilchi@eng.ui.ac.ir

Abstract—The aim of this paper is to utilize Support Vector Machine (SVM) as feature selection and classification techniques for audio signals to identify human emotional states. One of the major bottlenecks of common speech emotion recognition techniques is to use a huge number of features per utterance which could significantly slow down the learning process, and it might cause the problem known as "the curse of dimensionality". Consequently, to ease this challenge this paper aims to achieve high accuracy system with a minimum set of features. The proposed model uses two methods, namely "SVM features selection" and the common "Correlation-based Feature Subset Selection (CFS)" for the feature dimensions reduction part. In addition, two different classifiers, one Support Vector Machine and the other Neural Network are separately adopted to identify the six emotional states of anger, disgust, fear, happiness, sadness and neutral. The method has been verified using Persian (Persian ESD) and German (EMO-DB) emotional speech databases, which yield high recognition rates in both databases. The results show that SVM feature selection method provides better emotional speech-recognition performance compared to CFS and baseline feature set. Moreover, the new system is able to achieve a recognition rate of (99.44%) on the Persian ESD and (87.21%) on Berlin Emotion Database for speaker-dependent classification. Besides, promising result (76.12%) is obtained for speaker-independent classification case; which is among the best-known accuracies reported on the mentioned database relative to its little number of features.

Index Terms—Emotion recognition, speech analysis, feature selection, support vector machine.

I. Introduction

Emotion recognition is one of the newest challenges in Human-Computer Interaction (HCI), particularly when the recognition is relied on speaker's voice, which is considered as the basic means of human communication. In order to address many operational needs in people's daily life, many systems are proposed to automatically identify human's emotional states out of human voice. Thanks to the developed emotion recognition systems, a machine could provide users with more adaptive and personalized services which can be regarded as a huge leap for human-machine interaction.

In literature, many potential applications for emotion recognition from speech are proposed in many systems including car board system, E-tutoring [1], automatic translation system [2], call center and mobile communication [3].

The identification of emotion-related speech features is one of the extreme challenges in speech analysis. In spite of the fact that many audio features are explored, there is still no general agreement on a fixed set of features. Also in the majority of works, various feature selection methods; also known as feature reduction, help to improve the performance of recognition systems.

Because of their flexibility, computational efficiency and capacity to handle high dimensional data, SVMs were extensively used as a classification method in many previous works; however, little attention is paid so far to utilize SVM as a feature selection method.

Language and cultural differences, on the other hand, are considered as other challenges in Speech Emotion Recognition (SER). Even though emotions themselves are universal phenomena, how these emotions are experienced, expressed, perceived, and regulated may vary from one culture to another. Since it is believed that the expression of a specific certain emotion depends on the speaker's cultural background and that exploring cultural and linguistic backgrounds are the keys to understanding emotions, many previous studies tried to analyze various aspects of emotion in different languages to understand the differences. Unlike the wide range of examinations on many languages like English and German, speech emotion recognition in the Persian language is not yet fully investigated.

As a result, this paper discusses the impact of using SVM attribute selection method on extracted feature set and make a comparison between other methods in three separate experiments. Moreover, the paper reports the systematic evaluation of the technique on the recently developed Persian emotional database introduced in [4]. Because of Persian language, one could expect some emotions different that of in German, English as well as

other languages. Accordingly, the famous Berlin database [5] is also employed to have a better evaluation of the proposed model, therefore, speaker-dependent and speaker-independent evaluation are performed on this database.

The rest of the paper is organized as follows: Section II discusses the previous works about emotion speech recognition, especially works based on SVM approaches. Section III introduces emotional databases used in this paper. Section IV lists the details of baseline feature set. The feature selection methods are discussed in Section V. Section VI presents experimental results of the system and Section VII concludes the paper.

II. Related Work

In order to boost recognition rate of SER systems, various techniques in feature extraction, feature selection, and classification were used in previous works. In order to obtain high accuracy system, many former works tried different sets of audio features such as energy, pitch, formants and Mel-frequency cepstral coefficients (MFCCs). These features are employed to differentiate emotions in spoken utterance.

Furthermore, various feature selection or reduction techniques are considered to identify those features with the highest potential. One popular technique for selecting those features is Principal Component Analysis (PCA) which is extensively used in [6-8]. Correlation-based Sub Set Evaluators is the other famous method which improved many recognition systems such as in [9, 10]. Other feature selection algorithms such as Mutual Information (MI) [7, 11], Canonical Correlation Analysis (CCA) [12] and Sequential Floating Forward Selection (SFFS) algorithm [13].

A wide range of classifiers is used for classification task in SER. Some with promising results are GMM [14, 15], HMM [16], ANN [17, 18] and Support Vector Machines (SVMs) [19-23]. SVM is used more recently for emotion recognition in speech. For example, SVM classifier is used to recognize three emotional states and its evaluation is carried out on Chinese and German databases in [23]. Similarly, with the help of histogram equalization as a data normalization method, SVM classifier is used to build up a multi-lingual system in [20]. They performed their evaluation on a Mandarin database and EMO-DB. In another work, both GMM and SVM classifiers are used to recognize five basic emotional states; and its evaluation is performed on their own database [24].

As stated in [25], unfortunately in the literature, there is a lack of uniformity in the way these methods are evaluated (different test sets, feature vectors and evaluation frameworks). Therefore, to have a relative comparison, table 1 reports part of the comparison made in [25] which shows some promising attempts on EMO-DB database with different classification methods (SVM, GMM, HMM, ANN, C4.5, RF and the combination of SVM, K-NN, Naïve Bayes, C4.5, ANN) along with their recognition rates.

Table 1. A part of the overview of classification performance in EMO-DB database reported in [25].

SVM- (EMO-DB)	Reference
87.5%	Schuller et al. (2005) [19]
90%	Vlasenko et al. (2007) [21]
78%	Luengo et al. (2010) [22]
88.6%	Wu et al. (2009) [26]

GMM- (EMO-DB)	Reference
74.6% (speaker independent)	Lugger and Yang (2007) [14]
63%	Mishra and Sekhar (2009) [15]

HMM- (EMO-DB)	Reference
89%	Yun and Yoo (2009) [27]
78.4% (speaker independent)	Fu et al. (2008) [16]

ANN- (EMO-DB)	Reference
63.3%	Fu et al. (2008) [17]
47% in Berlin EMO database (speaker dependent but utterance independent)	Anagnostopoulos and Vovoli (2010) [28]
83.2 and 55% (speaker dependent and independent)	Iliou and Anagnostopoulos (2009) [18]

C4.5- (EMO-DB)	Reference
61.5%	Schuller et al. (2005a) [19]

RF- (EMO-DB)	Reference
77.2 and 48% (speaker dependent and independent)	Iliou and Anagnostopoulos (2009) [18]

Combination of SVM, K-NN, Naïve Bayes, C4.5, ANN- (EMO-DB)	Reference
80.5%	Schuller et al. (2005a) [19]

In order to address the variety of expressing emotions in different languages, different vocal expression databases of the basic emotions are established in several languages. One of the most widely-used emotional databases is the famous Berlin EMO database which is in German [5]. The list of common and widely-used emotional databases can be found in [2]. However, only a few validated emotional databases could be found in the Persian language. In fact, the lack of a comprehensive and officially available Persian database makes emotion recognition in this language difficult. The earlier studies in Persian are articulated by native speakers with no expertise in acting to express emotional utterances. Furthermore, those works only carried on a limited number of emotions and/or audio features [4].

Just recently, however, a Persian emotional speech database (Persian ESD) is created which is for colloquial Persian and may be considered as one of the best, officially available emotional database in this language [4]. We used both Persian ESD and Berlin database to evaluate our model.

III. Emotional Speech Corpus

The efficiency of any SER system is highly dependent on the emotional speech database samples used. Therefore, it is necessary to use well-made databases to have a proper evaluation of a system. In this study, the Persian Emotional Speech Database (Persian ESD) [4] is used to evaluate the proposed model. Persian ESD is an emotional database which is validated by a group of 1,126 native Persian speakers. It is the collection of actor-based simulated audio emotion database in the Persian language which is used to train and test our model. The database contains a set of 90 validated novel Persian sentences classified in five basic emotional categories (anger, disgust, fear, happiness, and sadness), as well as a neutral category (6 emotions totally). These sentences are articulated by two native Persian speakers, one male, and one female, in three conditions: (1) congruent (emotional lexical content articulated in a congruent emotional voice), (2) incongruent (neutral sentences articulated in an emotional voice), and (3) baseline (all emotional and neutral sentences articulated in neutral voice). Since the congruent part is articulated in the emotional voice, it is chosen.

To date, few serious works are done regarding automatic emotion recognition on this database. However, the human perception test in [4] could represent a proper measurement to have a statistical comparison. The percentages of accurate responses are presented in table 2. This table represents the percentage of native listeners who accurately categorized the target emotion expressed in each sentence and was computed for each item and speaker.

Table 2. Distribution (as percentages) of the responses given to each of the intended expressions in congruent condition reported in [4]

	Ang.	Dist.	Fea.	Sad.	Happ.	Neut.	Non-of Above
Anger	97.55	0.7	0.5			0.1	1.15
Disgust	0.35	95.65	0.2	1.3	0.55	0.45	1.8
Fear	0.1		97.7	1.2			1
Sad	0.6		1.05	98.35			
Happy		0.55		0.65	97.7		1.1
Neutral						100	

As it is shown in Table 2, the recognition rate of human test for 5 emotions is almost less than 98 percent which indicates that the emotion recognition from speech is a rather difficult problem to solve even for human. Therefore, it is beyond expectation that machines can easily differentiate between the expressed emotions. Interestingly, the most difficult emotion to recognize was disgust with 95.65 % in the congruent condition.

In addition to Persian ESD database, for a better evaluation of our method, “Berlin Database of Emotional Speech” (EMO-DB) [5] is also used to train and test the model. The database is widely used in emotion classification studies. The Berlin Database consists of 535 speech samples, which contain German utterances relevant to emotions such as anger, disgust, fear, joy, sadness, surprise and neutral, acted by five males and five females. Of the seven mentioned emotions, six are chosen in this experiment (except surprise) and a total number of 438 utterances are used in this work. Since the surprise emotion is not considered as "universal six emotions", and it is not also included in Persian ESD database, we did not take it into account in our study.

IV. Feature Extraction

It is believed that a proper selection of features can have a significant impact on the classification performance. Many diverse audio features are assessed in the literature to boost up recognition rate in SER. However, necessarily not all of them have a positive impact on emotion recognition. In fact, having too many features can reduce the performance and/or increase the computing time. As a result, only a set of most significant features is considered in this study.

Pitch, Energy and Intensity are traditional but important prosodic features of speech which provide valuable information to differentiate emotional states. The resonant frequencies, on the other hand, are produced in the vocal tract referred to as formants in several forms, each at a different frequency, occurring at roughly 1000Hz intervals. The first three formants convey valuable information and are used in our experiments. Mel-frequency cepstral coefficients (MFCCs) are other common features that are used in fields like speech and

Table 3. Extracted audio features (Baseline set)

Time-Min	MedianF2	Std-HNR	Mean-MFCC1	Std-MFCC3
Time-Max	MedianF3	Jitter	Mean-MFCC2	Std-MFCC4
Mean-Intensity	MinF1	Mean-jitter	Mean-MFCC3	Std-MFCC5
Min-Intensity	MinF2	Std-jitter	Mean-MFCC4	Std-MFCC6
Max-Intensity	MinF3	Shimmer	Mean-MFCC5	Std-MFCC7
Std-Intensity	MaxF1	Energy	Mean-MFCC6	Std-MFCC8
Mean-pitch	MaxF2	Energy-Air	Mean-MFCC7	Std-MFCC9
Min-pitch	MaxF3	Power	Mean-MFCC8	Std-MFCC10
Max-pitch	StdF1	Power-Air	Mean-MFCC9	Std-MFCC11
Std-pitch	StdF2	Mean-Amplitude	Mean-MFCC10	Std-MFCC12
MeanF1	StdF3	Root-Mean-Square - Amplitude	Mean-MFCC11	ZCR
MeanF2	Mean-HNR	Std-Amplitude	Mean-MFCC12	
MeanF3	Min-HNR	Min-Amplitude	Std-MFCC1	
MedianF1	Max-HNR	Max-Amplitude	Std-MFCC2	

gender recognition, music information retrieval and recently used extensively in SER. The 12 MFCC coefficients are used in our experiments. For pitch contour configuration, the standard range 75 to 500 hertz is considered, which means that the pitch analysis method will only find values between 75 and 500 Hz.

Using Praat software [29], 68 sound or speech features are extracted from utterances representing information such as Duration, Pitch, Intensity, the first three Formants, Amplitude, Harmonicity or Harmonics-to-Noise Ratio (HNR), Jitter, Shimmer, Energy, Energy-Air, Power, Zero-cross-rate (ZCR) and the first 12 MFCCs. Table 3 indicates the set of 68 acoustic features which form our baseline feature set. As it is shown in table 3, we adopt several statistical parameters such as Minimum, Maximum, Median, Root-Mean-Square and standard deviation of explained features. Next section explains how applying various feature selection algorithms could impact the accuracy of the system.

V. Feature Selection

The main reason behind using feature selection is to eliminate features which are irreverent and redundant. Advantages of using feature selection algorithms are considered twofold: training time could significantly be reduced and it also helps minimize the problem of over-fitting. In this study, in order to form our "golden set" of sound features, two powerful feature selection methods, SVM attribute evaluation and Correlation-based Feature Subset Selection (CFS) are applied to the extracted feature sets and results are compared to each other. A brief description of used feature selection algorithms is presented as follows.

A. SVM Attribute Evaluation

SVM attribute evaluation method uses an SVM classifier in order to evaluate the worth and thus rank each attribute. All attributes are ranked by the square of the weight assigned by SVM. In the case of multiclass problems, one-vs.-all method is separately used for each class. More description is presented in [30].

B. Correlation-based Feature Subset Selection (CFS):

CFS is a fully automatic algorithm which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. It tries to create a feature subset that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. More description is presented in [31].

In this study, the Best-First feature search method is used to search the space of attribute subsets. Different cases with different feature sets along with their corresponding results are presented in the following experiments.

VI. Experiment And Result

After all computations of diverse features and selection of optimal set, they are fed into classifiers. The first used classifier is a support vector machine (SVM) with polynomial kernels based on Sequential Minimal Optimization (SMO). SMO algorithm breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the need for a numerical optimization algorithm. This implementation globally replaces all the missing values and transforms nominal attributes into binary ones [32].

SVMs, generally, introduce specific advantages over GMM and HMM which entail the global optimality of the training algorithm and the existence of excellent data-dependent generalization bounds. [15]

As a second classifier, we chose an Artificial Neural Network (ANN) following the Multilayer-Perceptron architecture. ANNs have been used in various pattern recognition problems. They are known to be more effective in modeling non-linear mappings compared to other classifiers like HMM and GMM. They usually achieve better accuracy when training data is relatively low [25]. The following sections explain the procedure and results.

A. Experiment 1: Speaker-dependent Approach On Persian ESD

In the first step, all the necessary features which were explained above are extracted. The detailed list of features is presented in table 3. Prior to applying the feature selection algorithms, we first evaluate the baseline performance of the acoustic models trained on the congruent part of Persian ESD database (emotional lexical content articulated in a congruent emotional voice). Then we examined the same by applying the two mentioned attribute selection algorithms: SVM attribute evaluation and Correlation-based Feature Subset Selection (CFS). In order to have a better evaluation, 10-fold-cross validation, as well as randomly selected train-test set (66% train set- 34% test set), are employed over the data set. Results are presented in table 4.

Table 4. The results of experiment 1 on congruent part of Persian ESD

Model	Classifier	Feature Selection	Selected features	Recognition Rate	
				Cross val.	Train - Test
M1	SVM	Baseline	68	98.89%	95.08%
M2	MP	Baseline	68	98.33%	98.36%
M3	SVM	CFS	20	98.33%	95.08%
M4	MP	CFS	20	96.67%	96.72%
M5	SVM	SVM	45	99.44%	98.36%
M6	MP	SVM	45	98.89%	98.36%

As shown in table 4, the baseline feature set yields excellent result, however using feature selection, does have a significant impact on the recognition performance. Because of eliminating many vital features, applying CFS caused accuracy reduction in comparison with the baseline set. However, with the help of SVM Attribute Evaluation accuracies in both classifiers got higher or remained as it was. In this case, we have observed that by selecting 45 features the best recognition rate is achieved and adding more irrelevant features or removing more essential features would decrease the recognition rate.

To sum up our first experiment, we reach 99.44% for the correctly classified rate with the help of SVM Attribute Evaluation and SVM classifier (M5). This rate is nearly 3 percent higher than human perception test. The selected features set for M5 model is provided in table 5 and confusion matrix is shown in Table 6.

Table 5. Selected feature set by SVM (M5)

Std-MFCC3	Mean-MFCC7	Std-MFCC6	Std-MFCC7	Mean-MFCC8
Max-Intensity	Mean-MFCC11	Std-MFCC5	Mean-MFCC6	Std-Pitch
Std-F1	Median-F3	Std-MFCC10	Mean-MFCC2	Jitter
Mean-Jitter	Std-MFCC8	Mean-MFCC12	Std-MFCC12	Mean-MFCC5
Std-MFCC1	Energy-Air	Std-MFCC9	Min-Pitch	Mean-Pitch
Std-Intensity	Shimmer	Mean-F3	Max-Amplitude	Std-MFCC2
ZCR	Mean-MFCC1	Std-MFCC4	Mean-MFCC10	Power
Power-Air	Mean-MFCC9	Mean-F1	Std-HTN	Std-F3
Max-F3	Mean-MFCC4	Min-HTN	Mean-MFCC3	Std-F2

Table 6. Confusion Matrix of M5 using SVM classifier

A	B	C	D	E	F	Classified as	Class
28	0	0	0	0	0	A	Sad (100%)
0	30	0	0	0	0	B	Disgusting (100%)
0	0	30	0	0	0	C	Happy (100%)
0	0	0	34	0	0	D	Angry (100%)
0	0	0	0	28	0	E	Neutral (100%)
1	0	0	0	0	29	F	Frightened (96.66%)

B. Experiment 2: Speaker-dependent Approach On EMO-DB

In order to compare the presented method with other works and evaluate it in other languages, EMO-DB is employed and 3 separate evaluations are performed on it. 10-fold-cross validation and randomly selected train-test validation (66% train set- 34% test set) form the first two experiments and speaker-independent evaluation builds our third part which is explained in the next section. Table 7 shows the results of 10 Fold-Cross-Validation and randomly selected train-test set evaluation on Berlin database.

Table 7. The result of experiment 2 on EMO-DB

Model	Classifier	Feature Selection	Selected features	Recognition Rate	
				Cross val.	Train - Test
M1	SVM	Baseline	68	86.53%	82.55%
M2	MP	Baseline	68	84.70%	81.88%
M3	SVM	CFS	20	78.08%	77.85%
M4	MP	CFS	20	78.77%	81.21%
M5	SVM	SVM	61	87.21%	84.56%
M6	MP	SVM	61	85.16%	85.23%

The results presented in table 7 show that similar to our previous experiment, applying CFS brings about accuracy reduction and applying SVM Attribute Evaluation raises the accuracy compared to the baseline set. The highest recognition rate is achieved by applying SVM Attribute Evaluation as feature selection and using SVM as the classifier. Selected feature set and confusion matrix for this setting (M5) are presented in table 8 and 9 respectively.

Table 8. Selected feature set by SVM

Energy	Min-Pitch	Mean-MFCC6	Min-F3	Max-F1
Mean-MFCC4	Mean-MFCC8	Mean-Amplitude	Std-Pitch	Std-Amplitude
Mean-MFCC1	Mean-Jitter	Shimmer	Power	Mean-Intensity
Min-Intensity	Std-MFCC11	Std-MFCC7	Max-F2	Min-F1
Std-MFCC2	Mean-Pitch	Std-MFCC4	Min-F2	Root-Mean-Square-Amplitude
Mean-F1	Mean-MFCC2	Std-Intensity	Max-Amplitude	
Mean-HTN	Std-F3	Mean-F3	Mean-MFCC10	
Mean-MFCC5	Mean-MFCC12	Mean-MFCC3	Max-Pitch	
Max-Intensity	Mean-MFCC7	Std-MFCC3	Std-MFCC8	
Std-F1	Std-Jitter	Std-MFCC1	Max-F3	
Median-F1	Std-MFCC10	ZCR	Median-F2	
Mean-MFCC11	Min-HTN	Std-MFCC5	Std-MFCC12	
Std-F2	Jitter	Std-HTN	Energy-Air	
Mean-MFCC9	Min-Amplitude	Median-F3	Std-MFCC6	

Table 9. Confusion Matrix of M5 using SVM classifier

A	B	C	D	E	F	Classified as	Class
45	0	0	15	5	3	A	Happy (66.17%)
0	57	1	0	0	0	B	Sad (98.27%)
0	1	73	0	3	1	C	Neutral (93.58%)
8	0	0	118	0	0	D	Angry (93.65%)
3	2	2	2	53	0	E	Frightened(85.48%)
3	0	3	1	3	36	F	Disgusting(78.26%)

The overall experimental results reveal that applying SVM attribute selection method and using SVM classifier obtains highest accuracy rate on both Persian emotional database (99.44%) and Berlin emotional database (87.21%) for speaker-dependent classification. We notice that the contributing features in the two databases are different. It is suggested that cultural differences might be the reason behind it [23]. Furthermore, to reach the highest possible accuracy, the number of features used in Berlin database is relatively higher than Persian ESD. The possible reason behind that might be the diversity of speakers in Berlin database.

C. Experiment 3: Speaker-independent Approach

Since Persian ESD was recorded only by two persons (one man and one woman), performing Leave-One-Speaker-Out scheme in order to a have speaker-independent approach is impossible. In order to evaluate our model in speaker-independent approach, the common Leave-One-Speaker-Out scheme is used to evaluate the performance of Berlin database. Nine speakers' speech data is used for training the models and remaining one speaker's speech data is used for validating the trained models. In this experiment, SVM is chosen as classifier because of its noticeable performance in former experiences. Table 10 shows the results of speaker-independent classification on Berlin database.

Table 10. Correctly Classified rates of Leave-One-Speaker-Out scheme – (EMO-DB)

Classifier	SVM	
Feature Selection Method	Baseline	SVM
Number of Features	68	62
Recognition Rate (10-fold-cross validation)	75.81%	76.12%

As it is shown in Table 10, the proposed recognizer yields noticeable performance in Leave-One-Speaker-Out scheme compared to existing systems and by applying SVM Attribute Evaluation as feature selection, the performance reached the highest possible rate.

VII. Conclusion And Prospects

In this paper, we show that SVM feature selection method is more effective to increase system accuracy by reducing the number of the feature set. Other feature selection algorithms such as CFS are used as well, and results are compared analytically. The experimental results show that meaningful improvement is achieved in the recognition performance using SVM feature selection method.

Audio features such as energy, pitch, ZCR, formants, jitter, power, shimmer and MFCC are used to design our emotion recognizer. Also, SVM and Neural Network classifiers are employed to classify six universal emotions. The performance of the proposed model is evaluated on both Persian and German databases. Experimental results and comparison reveal a noticeable performance of the proposed recognizer. The best-achieved result is obtained when SVM algorithm is selected as the feature selection method and as a classifier.

The performance of an emotion recognizer is highly dependent on various factors such as the type of emotional database, the number and even the language of articulated utterances and also the evaluation framework. The speaker-independent and dependent frameworks are two types of common testing frameworks. However, the independent frameworks provide more reliable and natural evaluation [25]. For Persian language model, the system would be more robust if Persian ESD database had more utterances and were articulated by more than two speakers. In this way, speaker-independent evaluation could be done in this language as well as Berlin database; and indeed, the results would be more similar to the real world situation.

In future work, we plan to invoke linguistic features to enhance the accuracy of the system. This requires speech recognition system to be integrated within the model. We will also try to make our system more robust by combining databases together. In this way, we can improve the performance of the system with data fusion. Moreover, other perceptible emotions, besides the six common universal emotions, can be taken into account. For instance, the recognition of emotions like stress, jealousy, love, and pride could play an important role in many today's applications.

References

[1] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, 2004, pp. I-577-80 vol. 1.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.

[3] H. Jin, L. T. Yang, and J. J.-P. Tsai, *Ubiquitous Intelligence and Computing: Third International Conference, UIC 2006, Wuhan, China, September 3-6, 2006, Proceedings* vol. 4159: Springer, 2006.

[4] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klann-Delius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," *Behavior research methods*, vol. 47, pp. 275-294, 2015.

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

[6] Y. Zhou, Y. Sun, L. Yang, and Y. Yan, "Applying articulatory features to speech emotion recognition," in *Research Challenges in Computer Science, 2009. ICRCCS'09. International Conference on*, 2009, pp. 73-76.

[7] B. Schuller, R. J. Villar, G. Rigoll, and M. K. Lang, "Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition," in *ICASSP (1)*, 2005, pp. 325-328.

[8] S. Wang, X. Ling, F. Zhang, and J. Tong, "Speech emotion recognition based on principal component analysis and back propagation neural network," in *Measuring*

Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on, 2010, pp. 437-440.
[9] T. Kostoulas, T. Ganchev, A. Lazaridis, and N. Fakotakis, "Enhancing emotion recognition from speech through feature selection," in *Text, speech and dialogue*, 2010, pp. 338-344.
[10] C. N. Anagnostopoulos and E. Vovoli, "Sound processing features for speaker-dependent and phrase-independent emotion recognition in Berlin Database," in *Information systems development*, ed: Springer, 2010, pp. 413-421.
[11] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. ii/1085-ii/1088 Vol. 2.
[12] X. M. Cheng, P. Y. Cheng, and L. Zhao, "A study on emotional feature analysis and recognition in speech signal," in *Measuring Technology and Mechatronics Automation, 2009. ICMTMA'09. International Conference on*, 2009, pp. 418-420.
[13] H. Atassi and A. Esposito, "A speaker independent approach to the classification of emotional vocal expressions," in *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, 2008, pp. 147-152.
[14] M. Lugger and B. Yang, "An incremental analysis of different feature groups in speaker independent emotion recognition," in *16th Int. Congress of Phonetic Sciences*, 2007.
[15] H. K. Mishra and C. C. Sekhar, "Variational Gaussian mixture models for speech emotion recognition," in *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*, 2009, pp. 183-186.
[16] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition using hmms fusion system with relative features," in *Intelligent Networks and Intelligent Systems, 2008. ICINIS'08. First International Conference on*, 2008, pp. 608-611.
[17] L. Fu, X. Mao, and L. Chen, "Relative speech emotion recognition based artificial neural network," in *Computational Intelligence and Industrial Application, 2008. PACIIA'08. Pacific-Asia Workshop on*, 2008, pp. 140-144.
[18] T. Iliou and C.-N. Anagnostopoulos, "Comparison of different classifiers for emotion recognition," in *Informatics, 2009. PCI'09. 13th Panhellenic Conference on*, 2009, pp. 102-106.
[19] B. Schuller, R. Müller, M. K. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *INTERSPEECH*, 2005, pp. 805-808.
[20] B.-C. Chiou and C.-P. Chen, "Speech Emotion Recognition with Cross-lingual Databases," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
[21] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing," in *Affective Computing and Intelligent Interaction*, ed: Springer, 2007, pp. 139-147.
[22] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *Multimedia, IEEE Transactions on*, vol. 12, pp. 490-501, 2010.
[23] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, pp. 101-107, 2012.
[24] A. S. Utane and S. Nalbalwar, "Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine," *emotion*, vol. 2, p. 8, 2013.
[25] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, pp. 155-177, 2015.
[26] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic recognition of speech emotion using long-term spectro-temporal features," in *Digital Signal Processing, 2009 16th International Conference on*, 2009, pp. 1-6.
[27] S. Yun and C. D. Yoo, "Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4169-4172.
[28] C. N. Anagnostopoulos and E. Vovoli, "Sound processing features for speaker-dependent and phrase-independent emotion recognition in Berlin Database," in *Information systems development*, ed: Springer, 2009, pp. 413-421.
[29] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2001.
[30] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389-422, 2002.
[31] M. Hall, "Correlation-based Feature Subset Selection for Machine Learning, 1998," *Hamilton, New Zealand*.
[32] J. Platt, "Fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods—support vector learning*, vol. 3, 1999.

Authors' Profiles

Amirreza Shirani is a graduate student at the University of Isfahan who majors in software engineering. He earned his B.S. in computer engineering from Shahid-Beheshti University (National University of Iran). His area of interests includes Machine Learning, Pattern Recognition, Information Retrieval and Data Mining. He is currently working on his thesis in the area of Speech Emotion Recognition.

Ahmad R. Naghsh Nilchi is an Associate Professor of Artificial Intelligence and Multimedia Engineering at the University of Isfahan, Iran. He received his B.S., M.S., and Ph.D. degrees from Electrical and Computer Engineering Department in 1988, 1989, and 1996, respectively, all from the University of Utah, Salt Lake City, Utah, USA. He is the Chairman of the Artificial Intelligence and Multimedia Engineering at the University of Isfahan. He has been awarded several research grants from distinguished research institutions and has completed a number of research projects for Iranian industries. He is the author and co-author of several journal articles and conference papers. In addition, he has collaborated with internationally known institutions and peers and was a Research Scholar with the National University of Ireland Maynooth, Ireland, in 2011, and with the University of California, Irvine, in 2012. He also is the chief editor of the Journal of Computing and Security. His research interests include image and signal processing, data hiding, as well as intensive computing.