JAIDM
Journal of AI and Data Mining
Vol. 2, No .1, 2014, 53-61.

Classification of emotional speech through spectral pattern features

A. Harimi[1*], A. Shahzadi[1], A.R. Ahmadyfard[2] and Kh.Yaghmaie[1]

1. Faculty of Electrical & Computer Engineering, Semnan University, Iran.
2. Department of Electrical Engineering and Robotics, Shahrood University of technology, Iran.

Received 01 October 2012; accepted 09 February 2013
**Corresponding author: a.harimi@gmail.com (A. Harimi).*

Abstract
Speech Emotion Recognition (SER) is a new and challenging research area with a wide range of applications in man-machine interactions. The aim of a SER system is to recognize human emotion by analyzing the acoustics of speech sound. In this study, Spectral Pattern features (SPs) and Harmonic Energy features (HEs) for emotion recognition are proposed. These features extracted from the spectrogram of speech signal using image processing techniques. For this purpose, details in the spectrogram image are firstly highlighted using histogram equalization technique. Then, directional filters are applied to decompose the image into 6 directional components. Finally, binary masking approach is employed to extract SPs from sub-banded images. The proposed HEs are also extracted by implementing the band pass filters on the spectrogram image. The extracted features are reduced in dimensions using a filtering feature selection algorithm based on fisher discriminant ratio. The classification accuracy of the proposed SER system has been evaluated using the 10-fold cross-validation technique on the Berlin database. The average recognition rate of 88.37% and 85.04% were achieved for females and males, respectively. By considering the total number of males and females samples, the overall recognition rate of 86.91% was obtained.

Keywords: *Speech emotion recognition, spectral pattern features, harmonic energy features, cross validation.*

1. Introduction

Speaking is the fastest and most natural method of communication among human beings [1]. This fact has motivated researchers to use speech as the primary mode in human computer interaction. In order to make a natural interaction, the machine should be intelligent enough to recognize speaker's emotion by analyzing the acoustics of his or her voice. This has introduced a relatively *new* and challenging research area with a wide range of applications in man-machine interaction, known as Speech Emotion Recognition (SER). SER can improve the performance of automatic speech recognition systems [1]. It is also useful in e-learning, computer games, medicine, psychology and in-car boards [2-4].

SER is commonly treated as a pattern recognition problem which includes three main stages: feature extraction, feature reduction and classification. Despite of widespread efforts, finding effective features is still one of the main challenges in SER [1]. Most acoustic features used in SER can be listed in two main categories: prosodic features and spectral features. Prosodic features, which are widely used in SER, have been shown to offer important emotional cues of the speaker [1, 5]. These features are usually derived from statistics of pitch and energy contours [5]. Spectral features, on the other hand, have received increased attention in recent years. These features which are generally obtained from the speech spectrum can improve the rate of recognition by providing complementary information for prosodic features [5].

Figures 1 (a) to (c) show the spectrograms of an utterance expressed by a woman in 3 different emotions; anger, neutral and boredom, respectively. In time-frequency representation of speech signal, when voicing is present, horizontal

bands at the harmonics of the fundamental frequency of the vocal fold vibration will characterized the resulting spectrogram (pitch) [6].

As can be seen from Figure 1, harmonics, their position, stability and evolution are mostly related to the emotional state of the speaker. As it is reported by [1], in high arousal emotions such as anger, the resultant speech would be loud and fast with a higher pith average and wider pitch range. These conditions also induce the presence of strong high-frequency energy in the corresponding speech signal [1]. In low arousal emotions such as boredom, on the other hand, the resultant speech would be slow, low pitched and with little high-frequency energy [1]. These facts can be confirmed by Figures 1 (a) to (c).

The main contribution of this study is to propose Spectral Pattern features (SPs) and Harmonic Energy features (HEs) extracted from the speech spectrogram using image processing techniques. This scheme appears to be effective to fill the existing gap between time analysis methods and frequency analysis methods. The proposed SP and HE features are employed for classifying 7 emotions using a linear Support Vector Machine (SVM).

The reminder of this paper is organized as follows: Section 2 details the SPs and HEs proposed in this work, as well as prosodic and spectral features extracted for comparison purposes. Section 3 introduces the databases employed. Experimental results are presented and discussed in Section 4. The paper finally ends with conclusion remarks in section 5.

2. Feature extraction

In this section, we detail the proposed SPs and HEs extracted from the spectrogram of speech. Prosodic and spectral features considered in our experiments are also described. The prosodic and spectral features calculated in this study are by no means exhaustive, but serve as a representative sampling of the essential features. These features are used here as a benchmark, and also, to verify whether the SPs and HEs can serve useful additions to the widely used prosodic and spectral features.

2.1. Time-Frequency Representation Of Speech

Here, the silent part of speech signal is firstly discarded by a Voice Activity Detection (VAD) algorithm [7], and then the speech signal is passed through a pre-emphasize filter:

$$H(z) = 1 - \alpha z^{-1}, 0.9 \le \alpha \le 1 \quad (1)$$

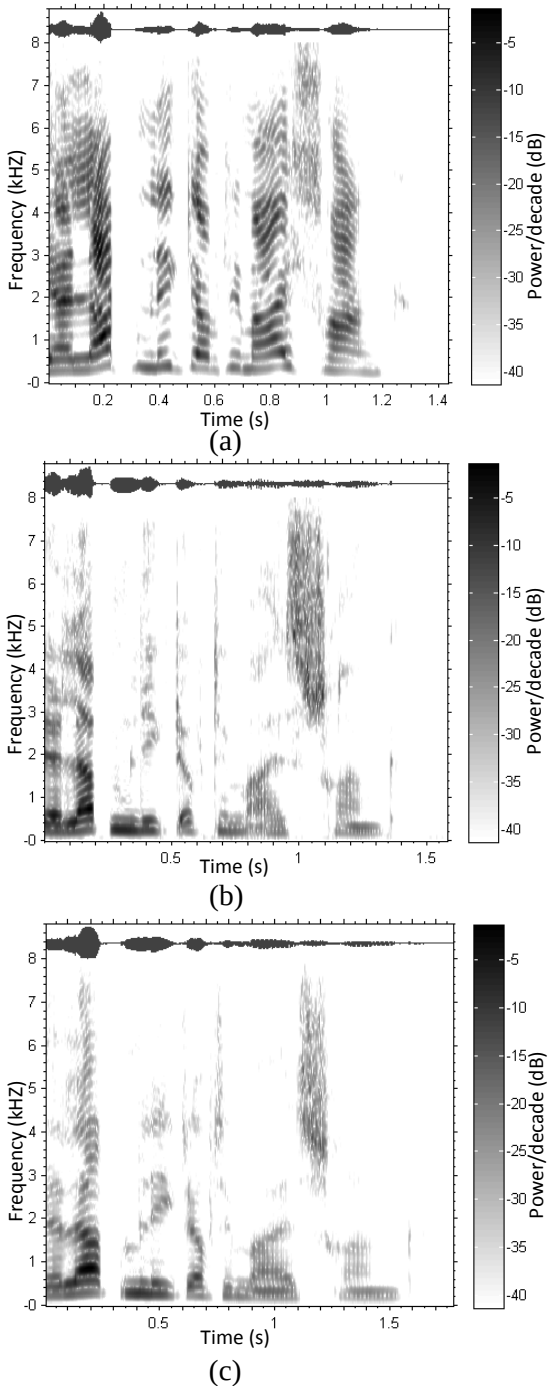

Figure 1. Spectrogram of an utterance expressed by a woman in 3 different emotions: (a) anger, (b) neutral, and (c) boredom.

As suggested by [8], α is set to 0.95 here. Although the speech signal is non-stationary in nature, it is assumed to remain stationary over a short duration of 20–30 ms. In this work, the pre-emphasized signal is segmented to frames of 320 samples length (20ms) with the shift of 160 samples (10ms) between two consecutive frames to retain a good quality of the signal and to avoid loss of information [8,9]. In order to reduce the

edge effects at the two ends of the frames, Hamming window is multiplied with each frame. Then, N=512 length DFT of a windowed frame is computed to obtain the logarithmic power spectrum as [10]:

$$S_i(k) = \log_{10}((\mathrm{Re}\{X_i(k)\})^2 + (\mathrm{Im}\{X_i(k)\})^2), \quad k = 0,1,...,N-1, i = 1,2,...,M \tag{2}$$

Where M is the total number of frames and $\tilde{X}_i(k)$ is the k[th] component of DFT of $\tilde{x}_i(n)$ (i[th] windowed frame). Re{…} and Im{…} indicate real and imaginary parts, respectively. The spectrums of these frames, $S_i(k)$, are concatenated row-wise to construct the speech spectrogram, f(k,i), as [10]:

$$f(k,i) = \begin{bmatrix} S_1(0) & ... & S_M(0) \\ . & . & . \\ . & . & . \\ S_1(N-1) & ... & S_M(N-1) \end{bmatrix} \tag{3}$$

2.2. Spectral pattern features

In order to highlight the details of spectrogram image, the contrast of the image is firstly increased using histogram equalization [11]. Then, the image is decomposed to eight components by applying eight directional filters, H1 to H8, which have been shown in Figure 2 (a). The eight resultant images are binarized, and then the morphological operators "cleaning" and "removing" are applied to remove the isolated pixels and interior pixels, respectively. Finally, the desired patterns are detected using binary masking technique. The employed binary masks are shown in Figure 2 (b).

In order to unify the patterns that represent a same direction (H5 and H6 & also H7 and H8), the logical "OR" operator is used. Thus, we would have six different binary images representing six different directional patterns. In these images, the pixels with the value of one indicate the presence of the corresponding pattern, and in the contrary, the pixels with the value of zero indicate the absence of the corresponding pattern.

Since, the behavior of the energy bands vary over different frequency ranges, we decompose the *images in*to several sub bands. The simplest method is to simply divide the bandwidth equally. However, this does not seem appropriate, as it does not correspond to the human ear [12]. The spectral resolution of the human ear varies logarithmically along the frequency, with better resolution at lower frequencies [13]. The Mel-scale and the Bark-scale which are empirically determined using human subjects are two potential choices. We decompose the images into 17 sub bands according to Bark-scale with non-overlapping filters. So, 102 sub-banded images will be obtained from the 6 binary images.

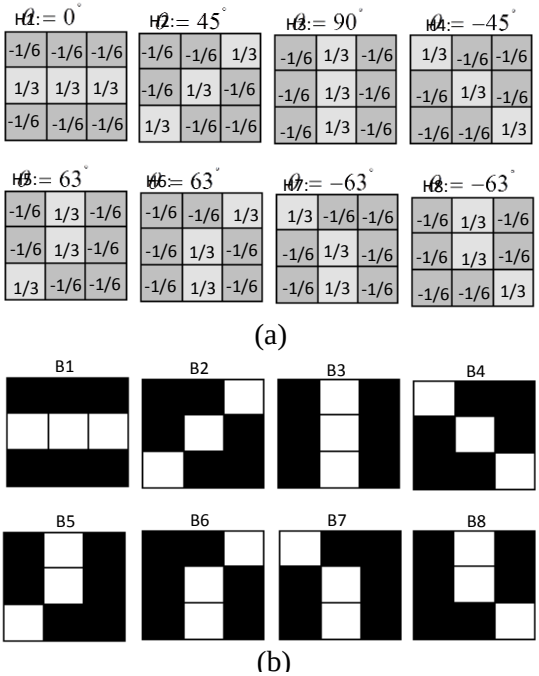

Figure 2. (a) Eight filters which are used to highlight eight directional patterns, (b) Eight binary masks which are used to detect eight directional patterns.

Two different types of features are extracted from each *sub-banded image:*

1) *The average number of patterns per frame in each sub band;* these features form the first 102 SP features of the SP Feature Vector, *SP_FV*. It could be formulated as follows:

$$SP\_FV(1:102) = \frac{1}{M}\sum_{r=1}^{R_q}\sum_{c=1}^{M} SI_{q,p}(r,c), \quad 1 \le q \le 17, 1 \le p \le 6 \tag{4}$$

Where $SI_{q,p}$ *is the* q^{th} *sub-band* of the p^{th} *image*. R_q is the number of rows in the $SI_{q,p}$ and M is the number of columns in the $SI_{q,p}$.

2) *The relative number of each pattern in each sub band*; these features form the second 102 features of the *SP_FV* and they could be determined as:

$$SP\_FV(103:204)=\frac{\sum_{r=1}^{R_q}\sum_{c=1}^{M} SI_{q,p}(r,c)}{\sum_{i=1}^{17}\sum_{r=1}^{R_i}\sum_{c=1}^{M} SI_{i,p}(r,c)}, \quad (5)$$

$1 \le q \le 17, 1 \le p \le 6$

The process of extracting SPs is schematically shown in Figure 3.

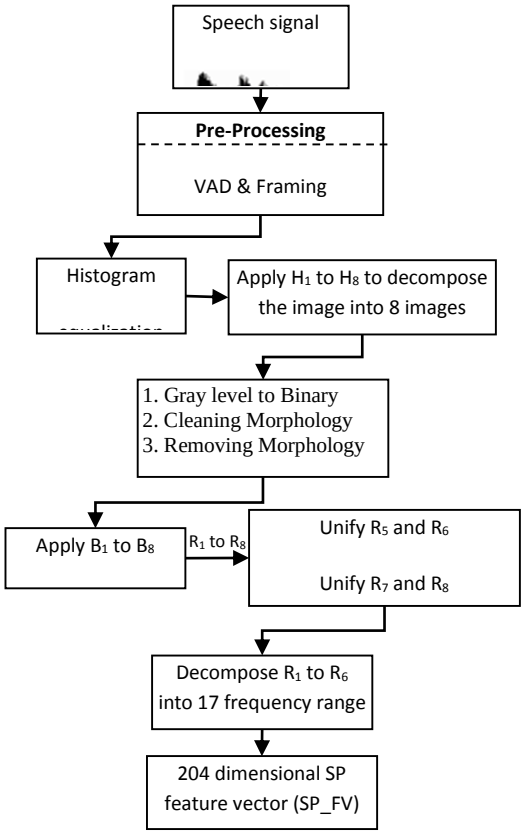

Figure 3. The process of extracting SPs.

2.3. Harmonic energy features

Harmonics of a speech signal are multiples of its fundamental frequency, F0, which is created by the vibration of the vocal folds. Since F0 varies over time, a filter bank consists of time varying band-pass filters is required to extract the harmonics of speech. In this filter bank, the central frequency of the h^{th} band-pass filter at a certain time is the h^{th} multiple of the fundamental frequency at that time, and the bandwidth is equal to the fundamental frequency [14].

In order to reduce the computational cost, we propose to implement the filter bank on the spectrogram image. In this method, the central frequency and cutoff frequencies of each of the sub-band filters on the image is determined as following. Firstly, the frequency range which is covered by each pixel in the vertical direction determined as:

$$FR = f_s / 2R \quad (6)$$

Where *fs* and *R* are the sampling rate and the number of rows in the image, respectively. The position of central frequency for the h^{th} filter in the i^{th} column (i^{th} frame) could be determined as:

$$Fc_h(i) = h \times F0(i) / FR, 1 \le i \le M \quad (7)$$

Where *F0(i)* is the fundamental frequency of the i^{th} frame which is computed using the autocorrelation-based pitch tracking algorithm [15]. The locations of the first and second cutoff frequencies on the image could be determined as follows:

$$Fs_{h,1}(i) = Fc_h(i) - F0(i) / 2FR \quad (8)$$

and

$$Fs_{h,2}(i) = Fc_h(i) + F0(i) / 2FR \quad (9)$$

Where $Fs_{h,1}(i)$ and $Fs_{h,2}(i)$ are the lower and upper cutoff frequencies of the h^{th} sub-band filter in the i^{th} column of image, respectively. The obtained values should be rounded to be used for digital images.

The energy of the h^{th} harmonic in the i^{th} frame could be determined as:

$$E_h(i) = \sum_{k=Fs_{h,1}(i)}^{Fs_{h,2}(i)} f(k,i), 1 \le i \le M \quad (10)$$

Where *f(k,i)* is the gray level of the pixel located in the k^{th} row and i^{th} column of the spectrogram image determined by equation (3). Figure 4 shows the 1^{th}, 5^{th}, 9^{th} and 13^{th} sub-band filters on the spectrogram image.

In this figure, the *F0* contour and its 5^{th}, 9^{th} and 13^{th} harmonics are indicated with black dashed lines. Each of these curves could be considered as the center frequency of the corresponding sub-band filter. The cutoff frequencies of each filter are also depicted around the center frequencies by solid lines.

In summary, the log energy contour of each harmonic could be computed by adding the gray level of the pixels within the corresponding pass band.

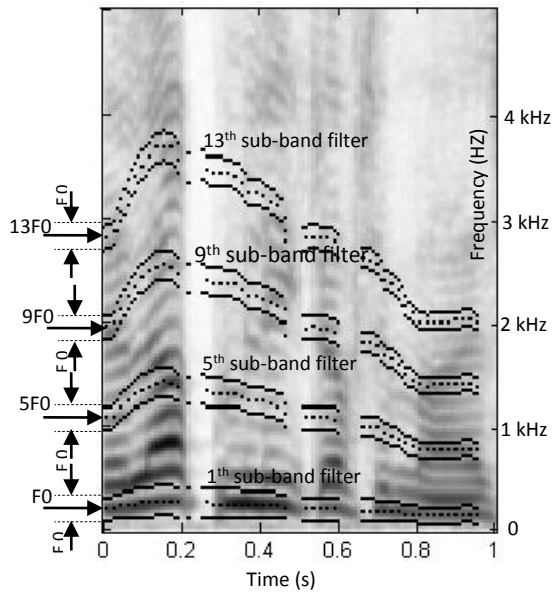

Figure 4. Implementation of the band pass filters on spectrogram image to determine harmonic energy contours.

In this work, we determine the energy contours of the 13 first harmonics. Then, we apply 20 statistical functions to these contours to extract HEs. The statistical functions are also applied to their first and second derivatives (velocity and acceleration) to capture the dynamical information of the curves. These functions include: min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentile, interquartile range, mean average deviation, standard deviation, skewness and kurtosis [16,17]. So, in total, $3\times 20\times 13=780$ HE features have been extracted here.

2.4. Prosodic features

Prosodic features form the most widely used features in SER [1,5]. In order to extract these types of features, statistical properties of pitch and energy tracking contours are commonly used. Here, 20 time domain functions are applied to capture the statistical information of pitch and energy tracking contours. These functions include: min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentile, interquartile range, average deviation, standard deviation, skewness and kurtosis [16,17]. These functions are also applied to the first and second derivatives of the contours as a common practice [5]. So, we have in total 60 pitch-based features and 60 energy-based features.

The widely used Zero-Crossing Rate (ZCR) and the Teager Energy Operator (TEO) [18] of the speech signal are also examined here. These features do not directly relate to prosody but in this work we evaluate their performance along with prosodic features. TEO conveys information about the nonlinear airflow structure of speech production [19]. The TEO for a discrete-time signal x_n is defined as:

$$TEO(x_n)=x_n^2-x_{n-1}x_{n+1} \tag{8}$$

In order to extract ZCR and TEO related features, we apply the 20 statistical functions to ZCR and TEO curves and their deltas and double deltas. Finally, we have 240 prosodic features.

2.5. Spectral features

We employ two types of spectral features: The Mel-Frequency Cepstral Coefficients (MFCCs) and formants are reported as effective spectral features for emotion recognition [20-23]. Here, the first 12 MFCCs and 4 formants are extracted from 20 ms Hamming-windowed speech frames every 10 ms, and so their contours are formed. Finally, the 20 functions described in section 2.3, are applied to extract spectral features from the extracted contours and their first and second derivatives. In total, 960 spectral features are extracted here.

3. Emotional speech data

The Berlin database of German emotional speech [24] is a well-known public database. The performance of many SER systems has been evaluated using this database [5,25-28]. This database includes 535 utterances with 10 different contexts expressed by ten professional actors (5 males and 5 females) in 7 emotions. Table 1 lists the numbers of samples for the emotion categories.

Table 1. Number of samples in the Berlin database.

Emotion	Female	Male
Anger	67	60
Joy	44	27
Boredom	46	35
Neutral	40	39
Disgust	35	11
Fear	32	37
Sadness	37	25
All	301	234

4. Experimental results and discussion

In this study, it is assumed that a gender classifier with perfect classification accuracy, which is proposed by [29], is employed in the first stage, so

the system is implemented completely separate for males and females. Features from training data are linearly scaled to [-1, 1] before applying linear Support Vector Machine (SVM). Features from test data are also scaled using the trained linear mapping function [5]. In order to avoid the curse of dimensionality [30], a filter-based feature selection scheme based on the Fisher Discriminant Ratio (FDR) is employed to remove irrelevant features. In this method, the FDR evaluates individual features by means of measuring the inter-classes distance against the intra-class similarity as [5]:

$$FDR(u) = \frac{2}{C(C-1)} \sum_{c_1} \sum_{c_2} \frac{(\mu_{c_1,u} - \mu_{c_2,u})^2}{\sigma^2_{c_1,u} + \sigma^2_{c_2,u}} \quad (8)$$

$$,1 \le c_1 < c_2 \le C$$

where $\mu_{c_i,u}$ and $\sigma^2_{c_i,u}$ represent the mean and variance of the u^{th} feature of the i^{th} class, respectively. $i = 1, 2, \ldots, C$, and C is the total number of classes. Features with little discrimination ratio can then be removed by a thresholding process.

The features proposed here are first compared to prosodic and spectral features, using FDR criterion before applied to the classifier. To this end, the features are ranked by their FDR values using all samples in the Berlin database and then, FDR values averaged over the top N_{fdr} FDR-ranked features. Figures 5 and 6 show the average FDR curves of prosodic, spectral and proposed SP and HE features as a function of N_{fdr} for females and males, respectively. Since there are only 240 prosodic features, all the curves are depicted for 240 features. These curves roughly illustrate discrimination power of features regardless of the utilized classifier.

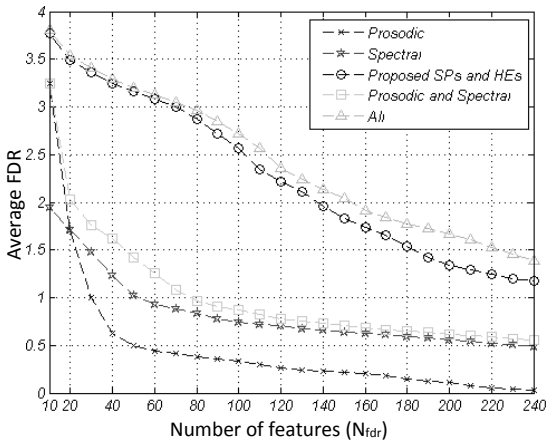

Figure 5. average FDR curves different types of features (females).

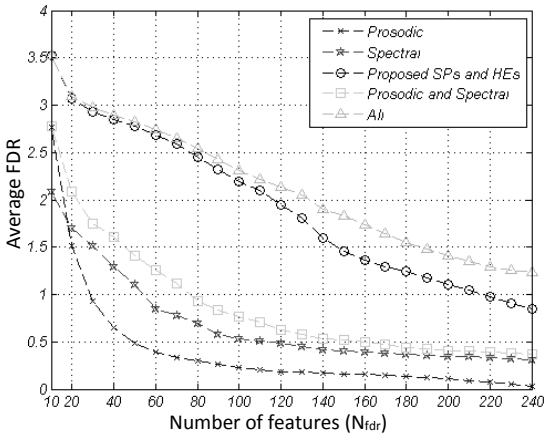

Figure 6. Average FDR curves different types of features (males).

As can be seen from Figures 5 and 6, the proposed SPs and HEs consistently exhibit considerably better discrimination power than the conventional prosodic and spectral features and their combination. However, combining the prosodic and spectral features to the SPs and HEs can slightly upgrade the discrimination power of features. Interestingly, for all types of features, the average FDR of females are higher than males. This shows that the extracted features are more discriminative for females' emotions than males' emotions.

As mentioned earlier FDR can only evaluate discriminative power of each feature individually. So, in order to evaluate power of a feature set in classification, we use directly classification accuracy as a criterion. The classification accuracy represents the performance of the employed classifier as a function of N_{top} features, which are chosen by the FDR-based feature selection algorithm. The classification accuracy is determined as the number of samples correctly recognized divided by the total number of samples. As a common practice for small sample size problems [31], results are produced using 10-fold cross-validation here. In this technique, in each class, samples have been randomly divided into 10 non-overlapping subsets approximately equal in size. In each validation trial, nine subsets from each class are taken for training, and the remaining one kept unseen until the testing phase. The overall recognition rate is achievable by averaging over the results of the 10 validation trials. In this work, we determine the accuracy curves for different types of features using a linear multi-class SVM. The computed curves are depicted in Figures 7 and 8 for females and males, respectively.

According to Figures 7 and 8, for both females and males, spectral features are superior to prosodic and proposed SP and HE features. However, by combining the proposed features with the prosodic and spectral features, the maximum recognition rates of 88.37% and 85.04% are achievable using 700 and 250 top FDR features, for females and males, respectively. Interestingly, all accuracy curves suggest that females' emotions can be classified more accurately than males' emotions. This may be due to the fact that females are more emotionally perceptive and emotional stimuli than males are [32].

Moreover, Figures 7 and 8 show that the classification accuracy is initially improved by increasing the number of features, but after a critical value further increase of the number of features result in degrading the performance. This can be explained as the curse of dimensionality, overfitting or peaking phenomenon [33], wherein the optimal number of features could be represented as a function of the number of samples and correlation of the features [34].

Tables 2 and 3 represent the confusion matrices of applying the proposed classifier for classifying 7 emotions using the combination of all types of features. In Tables 2 and 3, the left-most column is the true classes and the top row indicates the recognized classes. Furthermore, the rate column shows the average recognition rate for each class, which is determined as the number of samples correctly recognized and divided by the total number of samples in the class. The precision of each class is calculated as the number of samples correctly is classified and divided by the total number of samples assigned to the class.

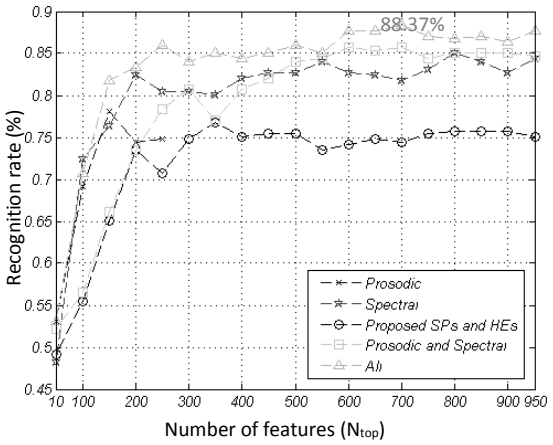

Figure 7. Accuracy curves for different types of features (females).

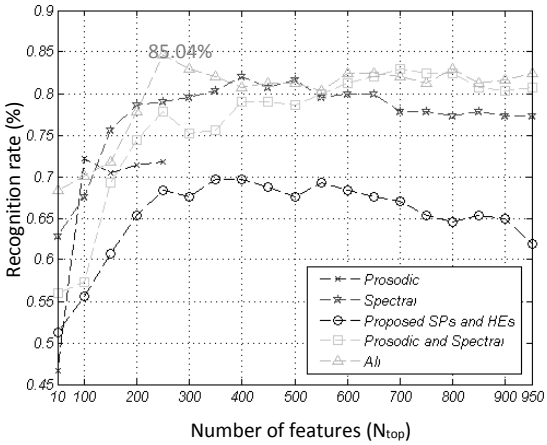

Figure 8. Accuracy curves for different types of features (males).

Table 2. Confusion matrix for classification of 7 emotions using proposed classifier (females).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate (%)
Anger	**56**	0	0	1	10	0	0	83.58
Boredom	0	**42**	0	0	0	4	0	91.30
Disgust	0	0	**34**	1	0	0	0	97.14
Fear	1	0	1	**28**	1	0	1	87.50
Joy	9	0	1	0	**34**	0	0	77.27
Neutral	0	3	0	0	0	**37**	0	92.50
Sadness	0	2	0	0	0	0	**35**	94.56
Precision (%)	84.85	89.36	94.44	93.33	75.56	90.24	97.22	
Overall accuracy: 88.37%								

Table 3. Confusion matrix for classification of 7 emotions using proposed classifier (males).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate (%)
Anger	**56**	0	0	1	3	0	0	93.33
Boredom	0	**30**	0	0	0	4	1	85.71
Disgust	0	0	**8**	1	1	0	1	72.733
Fear	3	0	0	**32**	1	1	0	86.49
Joy	6	0	0	0	**21**	0	0	77.78
Neutral	1	6	1	0	0	**30**	1	76.92
Sadness	0	2	0	1	0	0	**22**	88.00
Precision (%)	84.85	78.95	88.89	91.43	80.77	85.71	88.00	
Overall accuracy: 85.04%								

As can be seen from Tables 2 and 3, the ambiguity in classification of anger vs. joy and also boredom vs. neutral are responsible for major part of error in the proposed classifier. This may be due to the fact that most of the acoustic features employed for SER are related to arousal [35], and so are they not discriminative for valence related emotions such as anger and joy [5,35].

By considering the total number of 301 female and 234 male samples, the overall recognition rate of 86.91% is obtained for the proposed SER system.

It can be also useful to review performance Figures reported on the Berlin database by other works. Although the numbers cannot be fairly compared due to different conditions of experiments such as different data partitioning, they can be useful for general benchmarking. The recognition rate of 86.90% is achieved under 10-fold cross-validation in [25]. The recognition rate of 88.8% is reported by employing a three-stage classification scheme for recognizing six emotions only [26]. In [5], 85.6% accuracy is obtained under 10-fold cross-validation for classifying 7 emotions. In [36], the best average recognition rate of 85.5% is reported using a multi-class SVM classifier.

5. Conclusion

The aim of this study was to evaluate the proposed SPs and HEs for the recognition of human emotions from speech. These features have also been compared to conventional prosodic and spectral features in terms of FDR score and classification accuracy. This paper has demonstrated the potential and promise of the SPs and HEs for emotion recognition. The following conclusions can be drawn from the present study.

First, harmonics, their position, stability and evolution are mostly related to the emotional state of the speaker. This affects the behavior of energy bands and spectral patterns on the spectrogram image.

Second, although the SPs and HEs are superior to conventional prosodic and spectral features in term of FDR score, they are not the best in classification performance. However, these features boost the classification accuracy when used to augment the conventional prosodic and spectral features.

Also, our experiments reveal that females' emotions can be recognized more accurate than males' emotions. Moreover, most acoustic features employed for SER are discriminative for classifying emotions based on arousal level while they are ineffective for classification of valence related emotions [5,35]. This fact results in the ambiguity in classification of anger vs. joy and also boredom vs. neutral which is responsible for major part of error in most SER systems.

In order to improve the performance of current speech emotion recognition systems, the structure of the classifier can be optimized. To this end, tandem classifiers can be employed for classification of valence related emotions. Also, finding effective features for classifying valence related emotions can be a beneficial research focus. Moreover, as ultimate aim of a speech emotion recognition system is to recognize emotions for real work data, evaluating the proposed system under different conditions such as the presence of noise and chatter is useful.

References

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

[2] B. Schuller, G. Rigoll, M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", in: Proceedings of the ICASSP, (1), 577–580, 2004.

[3] D. J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Trans. Biomedical Eng. 47 (7), 829–837, 2007.

[4] J. Hansen, D.C. Icarus, "source generator based real-time recognition of speech in noisy stressful and Lombard effect environments", Speech Commun. 16 (4), 391–422, 1995.

[5] S. Wu, T.H. Falk, W.Y. Chan, "Automatic speech emotion recognition using modulation spectral features", Speech Communication 53, 768‑785, 2011.

[6] P. Gomez Vilda, J.M. Ferrandez Vicente, V. Rodellar Biarge, R. Fernandez Baillo, "Time-frequecy representations in speech perception", Neurocomputing, 72, 820–830, 2009.

[7] J. Sohn, N. S. Kim, and W. Sung. "A statistical model-based voice activity detection", IEEE Signal Processing Lett., 6 (1), 1–3, 1999.

[8] L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] T. Kinnunen, H. Li, "An overview of text-independent speaker recognition: from features to

supervectors", Speech Communication, 52 (1), 12–40, 2010.

[10] P. K. Ajmera, D.V. Jadhav, R.S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram", Pattern Recognition, 44, 2749–2759, 2011.

[11] R. C. Gonzalez, R.E. Woods, "Digital image processing", Pearson Prentice Hall, 2008.

[12] K. M. Indrebo, R. J. Povinelli, M. T. Johnson, "Sub-banded reconstructed phase space for speech recognition", Speech Communication, 48, 750-774, 2006.

[13] B. Gold, N. Morgan, "Speech and Audio Signal Processing". John Wiley and Sons, 2000.

[14] Z. Xiao, E. Dellandrea, W. Dou, L. Chen, "Automatic Hierarchical Classification of Emotional Speech". Ninth IEEE International Symposium on Multimedia Workshops, ISMW '07, 56, 2007.

[15] S, Gonzalez, M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)", Proc EUSIPCO, 2011.

[16] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech". Neurocomputing, 84, 65–75, 2012.

[17] B. Schuller, M. Wimmer, L.M. osenlechner, C. Kern, G. Rigoll, "Brute-forcing hierarchical functional for paralinguistics: A waste of feature space?", Proceedings International Conference on Acoustics, Speech, and Signal Processing, 33, 4501–4504, 2008.

[18] J, Kaiser, "On a simple algorithm to calculate the 'energy' of a signal". Internat. Conf. on Acoustics, Speech and Signal Processing, 1, 381–384, 1990.

[19] G. Zhou, J. Hansen, J. Kaiser, "Nonlinear feature based classification of speech under stress". IEEE Trans. Audio Speech Language Process, 9, 201–216, 2001.

[20] T, Polzehl, A, Schmitt, F, Metze, M, Wagner, "Anger recognition in speech using acoustic and linguistic cues". Speech Communication, 53, 1198–1209, 2011.

[21] C.C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach". Speech Communication, 53, 1162–1171, 2011.

[22] L, He, M. Lech, N.C. Maddage, N.B. Allen, "Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomedical Signal Processing and Control", 6, 139–146, 2011.

[23] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, K. Elenius, "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. Computer Speech and Language", 25, 84–104, 2011.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A database of German emotional speech". Interspeech, 1517‑1520, 2005.

[25] B. Schuller, D. Seppi, A. Batliner, A. Maier, S. Steidl, "Emotion recognition in the noise applying large acoustic feature sets". In: Proc.Speech Prosody, 2006.

[26] M. Lugger, B. Yang, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters". In: Proc. Internat. Conf. on Acoustics, peech and Signal Processing, (4), 4945–4948, 2008.

[27] E.M. Albornoz, D.H. Milone, H.L. Rufiner, "Spoken emotion recognition using hierarchical classifiers". Computer Speech and Language 25, 556–570, 2011.

[28] N. Kamaruddin, A. Wahab, C. Quek, "Cultural dependency analysis for understanding speech emotion". Expert Systems with Applications, 11, 028, 2011.

[29] M. Kotti, C. Kotropoulos, "Gender classification in two Emotional Speech databases", ICPR, 2008.

[30] C. Bishop, "Pattern Recognition and Machine Learning". New York: Springer, 2006.

[31] J. R. Raudays, A.K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners". IEEE T PATTERN ANAL, 13, 252–264, 1991.

[32] S. Whittle, M. Yücel, M.B.H. Yap, N.B. Allen, "Sex differences in the neural correlates of emotion: Evidence from neuroimaging". Biological Psychology. 87, 319– 333, 2011.

[33] S. Theodoridis, K. Koutroumbas, "Pattern recognition". Academic Press, 2008.

[34] C. Sima, E.R. Dougherty, "The peaking phenomenon in the presence of feature-selection", Pattern Recognition Letters, 29, 1667–1674, 2008.

[35] E. Kim, K. Hyun, S. Kim, Y. Kwak, "Speech emotion recognition using eigen-fft in clean and noisy environments", in: The 16th IEEE International Symposium on Robot and HumanInteractive Communication, RO-MAN, 689–694, 2007.

[36] H. Altun, G. Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection", Expert systems with Applications, 36, 8197–8203, 2009.