RECOGNITION OF EMOTION USING RECONSTRUCTED PHASE SPACE OF SPEECH

Ali Harimi[1], Hasan Shaygan Fakhr[2], Ali Bakhshi[3]

[1, 2, 3] Department of Electrical Engineering, Shahrood Branch, Islamic Azad University, Shahrood, Iran
[2] Department of Electrical Engineering, Shahrood Science and Research Branch, Islamic Azad University, Shahrood, Iran

Email: [1*] a.harimi@gmail.com, [2] hasanshaygan1362@gmail.com, [3] a.bakhshi.iau@gmail.com

ABSTRACT

In recent years, automatic recognition of human's emotion from speech has become one of the most important research areas, which can improve man-machine interaction. In this study, we proposed new features derived from reconstructed phase space (RPS) of speech. To this end, the RPS is uniformly divided into non-overlapping discrete cells and the number of points included in each cell is counted to form the proposed feature vector. Then multiple classifiers were examined to classify speech samples according to their emotional states. Our experimental results have demonstrated the potential and promise of proposed RPS based features as a useful combination for standard prosodic and spectral features. The best average recognition rate of 89.34% was obtained for classifying seven emotion categories in the Berlin database using a support vector machine with both radial basis function and polynomial kernels.

Keywords: speech emotion recognition; reconstructed phase space.

1.0 INTRODUCTION

Speaking is the fastest and one of the most important communication means of human beings. Now a days, automatic speech recognition (ASR) systems are extensively used in man-machine interaction. However, human speech is generally embedded with emotions to convey the intended message. From this fact arises a new multidisciplinary research area known as Speech emotion recognition (SER). Despite of widespread efforts done in SER, there are many challenging problems that need to be solve in order to improve the performance of SER systems [1]. Based on contradictory reports on the effect of emotions on some of acoustic attributes, specifying effective features is still the major unsolved problem.

To solve the above problem, many prosodic and spectral features have been presented. Prosodic features appear when sounds are put together in connected speech and they are mainly deal with intonation, stress and rhythm of speech. It has been shown that prosodic features, which are widely used in SER, suggest important emotional cues of the speaker [1-7]. In theliterature, pitch, duration, energy and their derivatives are widely used to representprosodic features [1,8]. Features extracted from the spectrum of speech are generally called spectral. These features convey the frequency contents of the signal and provide complementary information for prosodic features [1].Formants related features [9-13], Mell Frequency Cepstral Coefficients (MFCCs) [9,13-16], Linear Prediction Coding (LPC) [17,18], Sub-Band features [17,18] and Perceptual Linear Prediction (PLP) [15] have been reported as effective spectral features in SER.

Both prosodic and spectral features are generally computed by the traditional linear source-filter model of the human speech production system [19]. Unfortunately, such a model cannot convey nonlinear 3D fluid dynamics phenomena of speech [20,21]. In order to fill the existing gap between this ideal linear deterministic model and real strongly unpredictable speech production process, non-linear processing techniques can be used [22,23,24,25].In recent years, reconstructed phase space (RPS) of speech has been used for speech recognition [26,27], speech enhancement [26,27]and detecting sleepiness [30]. Moreover, nonlinear dynamics features extracted from RPS of speech has been employed in SER.It has been shown that geometrical properties of RPScontain important emotional cues of speaker

[24,25]. Our contribution in this work is to propose new features extracted from RPS of speech. In this method, distribution of points in PSR introduced as effective features utilized for our proposed SER system. By conveying the non-linear dynamics of speech, these features have been shown to be effective in SER.
In supervised learning problems like SER, there are a large number of classifiers that can be employed. There is no general report about which one is the best for all applications, but trends on SER can be categorized in two groups: (a) complex classifier like SVM used with very high dimensional feature space. (b) Straightforward classifiers like Bayesian–GMM and a moderate number of meaningful features [7]. In 1990s, most SER systems designed based on Bayesian learning and Linear Discriminant Analysis (LDA) [7]. Around 2000, systems based on Neural Network (NN) became popular [3]. Since 2002, Support Vector Machine (SVM) [4,7,15]as an extension of LDA with a high-dimensional feature space [7] and Hidden Markov Model (HMM) [31,32] to capture temporalstate transitions [7]have received more attention.Researchers are still trying to find the better solution [7].

The reminder of paper isorganized as follows. Section 2details the feature extraction consists of RPS based features, as well as prosodic and spectral features extracted for comparison purposes. Experimental results are representedand discussed in Section 3. The paper finally ends with conclusion remarks in section 4.

2.0 FEATURE EXTRACTION

In this section, we detail the process of extracting features from the Reconstructed Phase Space (RPS) of speech signal. Prosodic and spectral features considered in our experiments are also described here. These features are employed here as a benchmark, and more importantly, to verify whether the proposed features can serve as useful additions to the conventional prosodic and spectral features.

2.1. Reconstructed Phase Space Of Speech

In order to reconstruct the phase space of a time series x_n , n = 1,2 ,3,4 ….N, the vector $\overline{x_n}$ defined as:

$$\overline{x_n} = [\, x_n, x_{n+\tau}, x_{n+2\tau}, ..., x_{n+(d-1)\tau} \,] \qquad (1)$$

Where d and τ are the embedding dimension and chosen time lag values, respectively. In fact, the row vector $\overline{x_n}$ is a single point in the RPSrepresents the d sequential values of the time series in the intervals of τ. Finally, the RPS formed by these points:

$$X = \begin{bmatrix} x_1 & x_1+\tau & x_1+2\tau & \dots & x_1+(d-1)\tau \\ x_2 & x_2+\tau & x_2+2\tau & \dots & x_2+(d-1)\tau \\ x_3 & x_3+\tau & x_3+2\tau & \dots & x_3+(d-1)\tau \\ . & & & & \\ . & & & & \\ . & & & & \\ x_N & x_N+\tau & x_N+2\tau & \dots & x_N+(d-1)\tau \end{bmatrix} \qquad (2)$$

According to Taken's theorem, the trajectory matrix, X, completely defines the dynamics of the system produced time series x_n. Each row of X represents a single point in the d dimensional RPS.

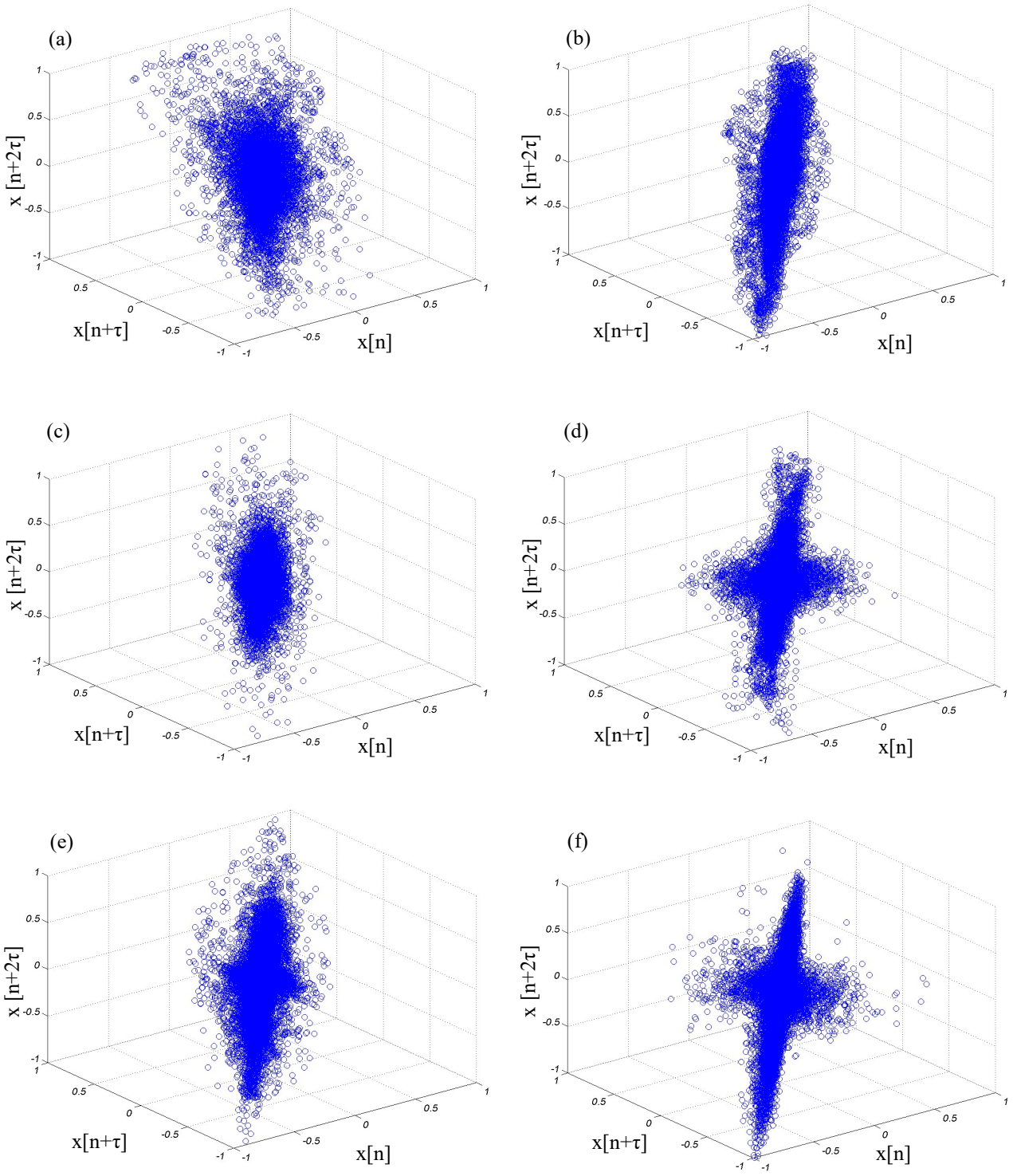

Fig.1. RPS of a sentence uttered in 6 different emotions: (a) anger, (b) boredom, (c) disgust, (d) fear, (e) joy, and (f) sadness.(d=3, and τ=1).

According to previous works [22, 23] and our experiments, embedding dimensiond=3 and time lag τ=1 are good choices to determine the RPS of speech signal. Fig.1 shows the RPSs of six speech signals with different emotions: anger, boredom, disgust, fear, joy, and sadness.

As can be seen in Fig.1, the distribution of points in the RPS relates to the emotional state of the speaker. For instance for an anger speech signal the point distributed in the RPS while for boredom they squeezed along the identity line. For disgust points placed near the center and for fear they have shaped a cross. From these observations arises the idea of deriving features from the RPS. To this end, here the RPS is uniformly divided into non-overlapping discrete cells as shown in Fig.2. The number of points included in each cell is counted and the corresponding feature vector is formed. The number of cells in RPS determines the size of feature vector. For instance, if the RPS divided into 64 cells, as shown in Fig.2, we would have a 64 dimensional feature vector; each feature corresponds to the number of points placed in the corresponding cell.

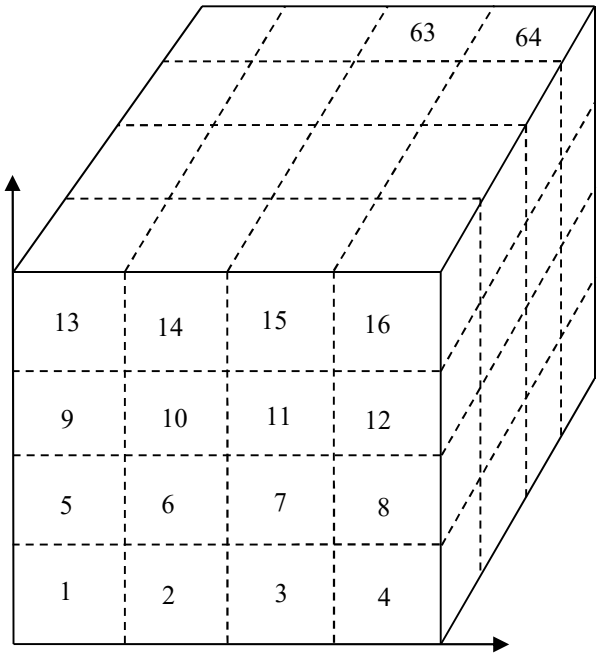

Fig.2. dividing the RPS into 64 non-overlapping cells.

It is clear that,thoughthe more cells can better describe the distribution of RPS points,theyincrease the size of feature vector and computational complexity. In this study, we perform our experiments by dividing the RPS into 125, 500 and 1000 uniform cells and their corresponding three feature vectors.

2.2. ProsodicFeatures

In linguistic literature, prosody of speech is defined as the supra-segmental or long-term features of speech. Speech features extracted from longer speech segments like syllables, words and sentences are known as prosodic features [1]. These features are the most widely used features in SER and treated as major correlates of vocal emotions.They are commonly based on pitch/F0, energy/intensity and speaking rate/rhythm. The statistics of pitch and energy tracking contours are shown to offer important emotional cues of the speaker [15]. In this work, 14 statistical functions include: minimum, maximum, range (max-min), mean, median, variance, standard deviation, skewness, kurtosis, quartiles, the difference between quartiles, linear & quadratic regression coefficients, regression error (RMSE) applied to pitch and energy contours and their first and second derivatives. Furthermore, mean and standard deviation of vowels duration, and voice to unvoice time duration ratio are employed. Mean of zero crossing ratio

and Teager Energy Operator [15] of the speech signal are also examined here. Totally, 89 prosodic features are extracted in this work.

2.3. Spectral Features

In this study, the MFCCs and PLPs are employed as two types of spectral features. These features are successfully applied to automatic speech recognition and also reported as effective spectral features for emotion recognition [15]. The first 13 MFCCs and 6PLPs are extracted from speech. After that, their contours are formed and then the mean and standard deviation ofthese contours and their first and second derivatives are calculated to form 114 spectral features in total.

3.0 EXPERIMENTS

In this section, the results of experimental evaluation are presented. SER, like most computer science problems is a language-base problem [33]. However, since most of the developed emotional speech databases are not available for public use [1], the proposed SER system is evaluated on the well-known Berlin database [34]. This German database consists of 535 utterances with 10 different contexts, which are expressed by 10 professional actors (5 male and 5 female) in seven emotions: “anger, boredom, disgust, fear, joy, neutral and sadness”.

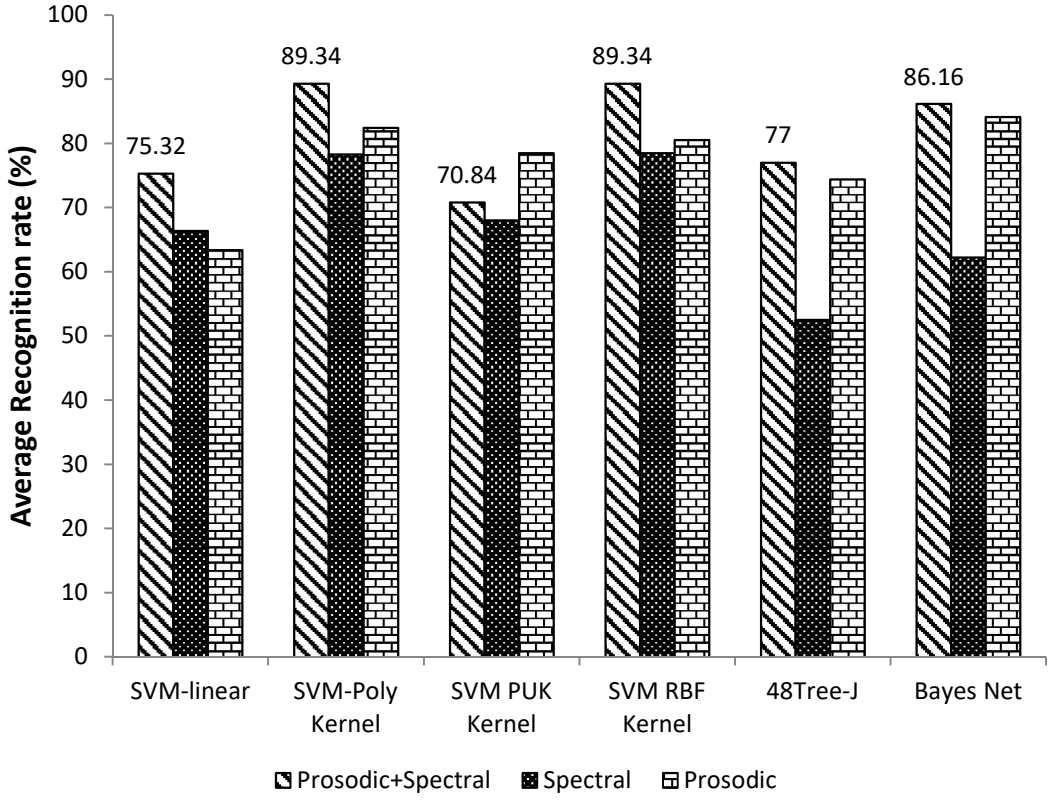

Fig.3. Average recognition rate achieved by different classifiers using prosodic and spectral features.

In order to overcome the small sample size problem which can influence the train and test reliability, all the experiments performed under 10-fold cross validation. Generally, in K-fold cross validation, increasing the number

of folds, K, make the training and testing procedures more reliablewith the price of computational complexity.Since many SER researches do their experiments with 10-fold cross validation, we also set K to 10. This will guarantee the results reliability and the execution time in experiments is acceptable. Features from training data are linearly scaled to [-1, 1] before applying the classifier. As suggested in [15], features from test data are also scaled using the trained linear mapping function. We evaluate the proposed features by 3 classifiers: byes net, tree-j48 and SVM (linear, polynomial, PUK and RBF kernels) using Weka software [35], which contains a collection of machine learning algorithms for data mining tasks. Fig.3 shows the results of each classifier using conventional prosodic and spectral features.

According to Fig.3, the best average recognition rate of 89.34% obtained by SVM-Poly and SVM-RBF using combination of prosodic and spectral features.It is coincident with the fact that RBF kernel is usually the best choice for SVM [15]. Thus, next experiments performed by the best classifier, SVM-RBF. Also, this figure shows that while prosodic features yields better results than spectral features, augmenting spectral features to prosodic features can improve the classification accuracy.

As it is described in section 2.1, we divided the RPS into 125, 512 and 1000 bins to construct three feature vectors PSR125, PSR512 and PSR1000, respectively. Fig.4 shows the results achieved by adding these feature vectors to the combined prosodic and spectral ones.

According to Fig.4, augmenting PSR125 to the prosodic and spectral features improve the recognition rate by 1.87%. Also, adding PSR512 and PSR1000 unexpectedly decreases the classification performance. This shows that increasing the number of bins in RPS does not necessarily lead to improve the accuracy.

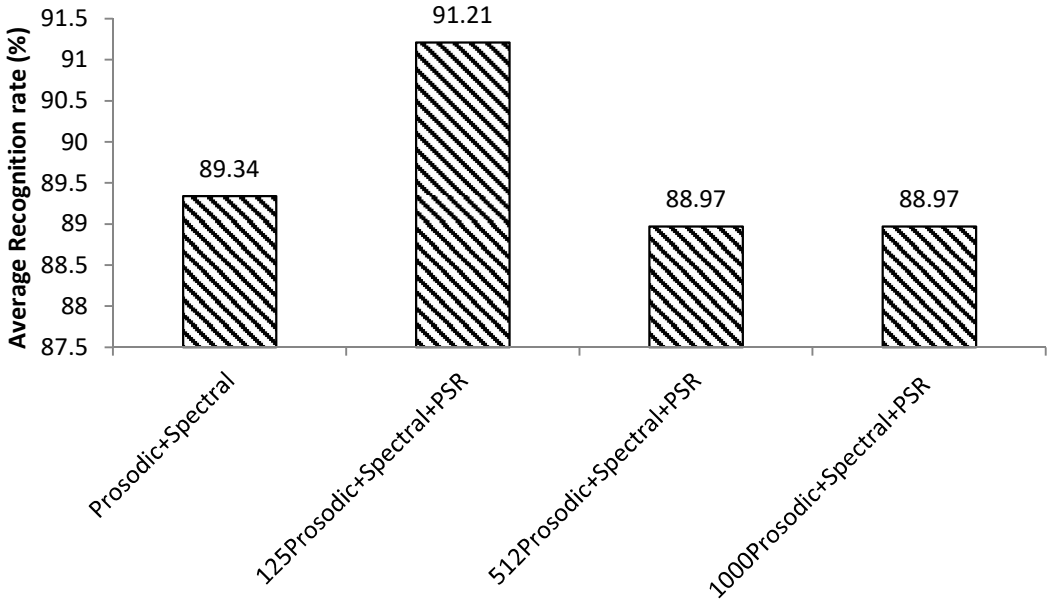

Fig.4. Average recognition rate achieved by SVM-RBF using prosodic and spectral features combined with proposed PSR feature vectors.

Table.1 shows the confusion matrix of best result achieved by combination of prosodic and spectral features with PSR125. According to Table.1, the worst recognition rate of 78.87% obtained for joy samples. The confusion of joy and anger samples (nine joy samples misclassified to anger and six anger samples misclassified to joy)are responsible for major part of classification error. Ambiguity in valence related emotions, joy and anger has been reported as a challenging problem in SER [15]. It is also interesting to see that the sadness samples are successfully

classified with the accuracy of 100%. The average recognition rate of 91.21% is determined for classification of all 7 emotions. In this study, the personality traits and sex differences was not considered for emotion recognition. However, it has been shown such information can improve the performance of SER systems [36].

Table 1. Confusion matrix using combination of prosodic and spectral features with PSR125 by SVM-RBF classifier.

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate(%)
Anger	**121**	0	0	0	6	0	0	95.28
Boredom	0	**73**	1	0	0	5	2	90.12
Disgust	0	1	**41**	3	0	0	1	89.13
Fear	1	0	0	**60**	4	1	3	86.96
Joy	9	0	1	3	**56**	2	0	78.87
Neutral	0	**3**	0	1	0	**75**	0	94.94
Sadness	0	0	**0**	0	0	0	**62**	**100**
Precision (%)	92.37	94.81	**95.35**	89.55	84.85	90.36	**91.18**	

It is also useful to briefly review performance figures reported on the Berlin database by other works. Although the numbers cannot be directly compared due to factors such as different data partitioning, employed classifier and testing strategies, they are still useful for general benchmarking. The maximum average recognition rate is reported 84.6% in [7] using a new feature selection technique. The best accuracy of 71.75% achieved in [14] using class-level spectral features. In [15], the authors proposed modulation spectral featureswhich resulted in best accuracy of 85.6%.In [17] by boosting selection of speech related features the performance of multi-class SVMs in emotion detection is improved to 84.8%. Class-specific multiple classifiers schemeproposed in [37] yields recognition rate of 83.73%. In [38],the authors proposed weighted spectral features based on local Hu moments and they achieved the best classification accuracy of 81.74%. The best recognition rate of 86.9% was achieved in [39] using spectral patterns. Also, in [40] by using variogram based features the best classification accuracy of 86.82% and 90.43% was obtained for only 6 emotions of females and males, respectively. In this work, we achieved the average recognition rate of 91.21%.

4.0 CONCLUSION

The aim of this study was to evaluate new features extracted from reconstructed phase space of speech for the recognition of human's emotions.These features were also compared to traditional prosodic and spectral features. This paper has demonstrated the potential and promise of RPS based features for emotion recognition. The following conclusions can be drawn from the present study.

The first major finding of this paper was that, for speech signal, distribution of points in the RPS depends on the emotional state of the speaker. For instance, for an anger speech signal the point scattered in the RPS while for boredom they squeezed along the identity line. For disgust points placed near the center and for fear they have shaped a cross. Our experiments showed that the proposed features extracted from RPS of speech are effectively complements for widely used prosodic and spectral features.

The second major finding was that, the choice of proper classifier astonishingly affects the result of classification. According to our results, the SVM based classifier with the polynomial and radial basis function outperforms other classifiers. It is interesting to see that in all cases the prosodic features have shown better results than spectral features, except linear SVM.

The third major finding was that, since most of the prosodic and spectral features extracted based on traditional linear source-filter models, extracting features from RPS can be efficient to fill the gap by conveying nonlinear

information of speech. According to our experiments these features can provide useful complementary attributes for prosodic and spectral features. Augmenting the proposed RPS base features to the conventional prosodic and spectral features can improve the classification accuracy. Whereas PSR125 features upgrade the average recognition rate by about 2%, PSR512 and PSR1000 decreases accuracy by about 0.4%. This indicates the importance of cell size in feature extraction procedure. Actually, while small cells can precisely describe the distribution of points in RPS, on the other hand, they increase the number of features which yields curse of dimensionality.

With possible refinement in future works, the performance of extracted features from the RPS may be further improved. Investigating in confusion matrixes reported in most SER systems show that confusing of joy and anger samples is responsible for major part of overall classification error. Thus, further research on finding effective features and strategies to classify valence related emotions such as anger and joy can improve SER systems.

REFERENCES

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

[2] B. Yang, M. Lugger, "Emotion recognition from speech signals using new harmony features", Signal Processing, 90, 2010, 1415–1423.

[3] J. Nicholson, K. Takahashi, R. Nakatsu, "Emotion recognition in speech using neural networks", Neural Comput. Appl. 9, 290–296, 2000.

[4] A. Qazi, H. Fayaz, A. Wadi, R. G. Raj, N.A. Rahim, W. A. Khan, "The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review", Journal of Cleaner Production, Vol. 104, pp. 1-12, 2015, ISSN 0959-6526, http://dx.doi.org/10.1016/j.jclepro.2015.04.041.(http://www.sciencedirect.com/science/article/pii/S0959652615004096).

[5] 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), pp. 829–837,2007.

[6] M. Hariharan, M. P. Paulraj, S. Yaacob, "Time-Domain Features And Probabilistic Neural Network For The Detection Of Vocal Fold Pathology", Malaysian Journal of Computer Science, pp. 60-67, 2010.

[7] J. Rong, G. Li, Y.P. Phoebe Chen, "Acoustic feature selection for automatic emotion recognition from speech". Information Processing and Management, 45, pp. 315–328, 2009.

[8] C.T. Ishi, H. Ishiguro, N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality", Speech Communication, 50, pp. 531–543, 2008.

[9] E. Bozkurt, E. Erzin, C.E. Erdem, A.T. Erdem, "Formant position based weighted spectral features for emotion recognition", Speech Communication, 53, pp. 1186–1197, 2011.

[10] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette,"Fear-type emotion recognition for future audio-based surveillance systems", Speech Communication, 50, pp. 487–503, 2008.

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

[12] M.B. Goudbeek, J.P. Goldman, K. Scherer, "Emotion dimensions and formant position", 10th Annual Conference of the International Speech Communication Association, pp.1575-1578, 2009.

[13] 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, pp. 139–146, 2011.

[14] D. Bitouk, R. Verma, A. Nenkova, "Class-level spectral features for emotion recognition", Speech Communication, 52, pp. 613–625, 2010.

[15] S. Wu, T.H. Falk, W.Y. Chan, "Automatic speech emotion recognition using modulation spectral features", Speech communication, 53, pp. 768–785, 2011.

[16] M. A. Shayegan, S. Aghabozorgi, R. G. Raj, "A Novel Two-Stage Spectrum-Based Approach for Dimensionality Reduction: A Case Study on the Recognition of Handwritten Numerals," Journal of Applied Mathematics, vol. 2014, Article ID 654787, 14 pages, 2014. doi:10.1155/2014/654787.

[17] 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, pp. 8197–8203, 2009.

[18] R.G. Raj and S. Abdul-Kareem. "A Pattern Based Approach for The Derivation Of Base Forms Of Verbs From Participles And Tenses For Flexible NLP". Malaysian Journal of Computer Science, Vol. 24(2), pp 63-72, 2011.

[19] R.G. Raj, S. Abdul-Kareem, "Information Dissemination And Storage For Tele-Text Based Conversational Systems' Learning", Malaysian Journal of Computer Science, Vol. 22(2), pp. 138-159, 2009.

[20] T. Sauer, JA. Yorke, M. Casdagli." Embedology", J Stat Phys,65, pp. 579–616, 1991.

[21] F. Takens."Detecting strange attractors in turbulence", Lect Notes Math, pp. 366–381, 1981.

[22] B. Rehmam, Z. Halim, G. Abbas, T. Muhammad, "Artificial Neural Network-Based Speech Recognition Using DWT Analysis Applied On Isolated Words From Oriental Languages," *Malaysian Journal of Computer Science*, Vol. 28 (3), pp. 242-262, 2015.

[23] Vásquez-Correa, J. C., et al. "Non-linear Dynamics Characterization from Wavelet Packet Transform for Automatic Recognition of Emotional Speech", Recent Advances in Nonlinear Speech Processing. Springer International Publishing, pp.199-207, 2016.

[24] A. Shahzadi, A.R.Ahmadyfard, A.Harimi, K.Yaghmaei, "speech emotion recognition using nonlinear dynamics features", Turkish Journal of Electrical Engineering & Computer Sciences, 23, pp. 2056 – 2073, 2015.

[25] A. Harimi, A.R. Ahmadyfard, A. Shahzadi, K. Yaghmaie, "Anger or Joy? Emotion Recognition Using Nonlinear Dynamics of Speech", Applied Artificial Intelligence, 29, pp. 675–696, 2015.

[26] K.M. Indrebo, R.J. Povinelli, M.T. Johnson. "Sub-banded reconstructed phase spaces for speech recognition", SpeechCommun 48, pp. 760–774, 2006.

[27] L. B. Huang, V. Balakrishnan, R.G. Raj, "Improving the relevancy of document search using the multi-term adjacency keyword-order model." Malaysian Journal of Computer Science, Vol. 25, No. 1, pp. 1-10, 2012.

[28] M.T. Johnson, A.C. Lindgren, R.J. Povinelli, X. Yuan. "Performance of nonlinear speech enhancement using phase spacereconstruction", In: IEEE 2003 International Conference on Acoustics, Speech, and Signal Processing; 6–10 April2003; Hong Kong, China. New York, NY, USA: IEEE. pp. 872–875.

[29] J. Sun, N. Zheng, X. Wang. “Enhancement of Chinese speech based on nonlinear dynamics”, Signal Process 2007; 87:2431–2445.

[30] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, B. Schuller. “Applying multiple classifiers and non-linear dynamicsfeatures for detecting sleepiness from speech”, Neurocomputing, 84, pp. 65–75, 2012.

[31] Schuller, B., Rigoll, G., & Lang, M. (2003). “Hidden markov model-based speech emotion recognition”, In Proceedings of the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP’03) (Vol. 2, pp. 1–4). IEEE Computer Society.

[32] Song, M., Chen, C., & You, M. (2004). “Audio-visual based emotion recognition using tripled hidden markov model”, In Proceedings of IEEE international conference on acoustic, speech and signal processing (ICASSP’04) (Vol. 5, pp. 877–880). IEEE Computer Society.

[33] B. Rehman, Z. Halim, M. Ahmad, "ASCII Based GUI System for Arabic Scripted Languages: A Case Study of Urdu", International Arab Journal of Information Technology, Vol. 11(4), July 2014.

[34] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, “A database of German emotional speech”, Interspeech, pp. 1517–1520, 2005.

[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, Volume 11, Issue 1.

[36] A. Terracciano, M. A. Zonderman, M. Evans, "Personality Traits and Sex Differences in Emotion Recognition Among African Americans and Caucasians", Annals of the New York Academy of Sciences 1000, 2016.

[37] A. Milton, S. Tamil Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals", Computer Speech and Language, vol 28, pp. 727–742, 2014.

[38] Y. Sun, G. Wen, J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition", Biomedical Signal Processing and Control, vol 18, pp. 80–90, 2015.

[39] A. Shahzadi, A. R. Ahmadyfard, K. Yaghmaie, A. Harimi, “Recognition Of Emotion In Speech Using Spectral Patterns”, Malaysian Journal of Computer Science. Vol. 26(2), pp. 140-158, 2013.

[40] Z. Esmaileyan, H. Marvi, “Recognition Of Emotion In Speech Using Variogram Based Features”, *Malaysian Journal of Computer Science, Vol. 27(3), pp.*156-170, *2014.*