IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 3 (Nov. - Dec. 2013), PP 48-52
www.iosrjournals.org

Upgrading the Performance of Speech Emotion Recognition at the Segmental Level

Agnes Jacob[1], P.Mythili[2]
[1]*Research Scholar* , [2]*Head of Department*
[1,2] *Division of Electronics, School of Engineering, Cochin University of Science and Technology Kochi, Kerala, India*

Abstract: *This paper presents an efficient approach for maximizing the accuracy of automatic speech emotion recognition in English, using minimal inputs, minimal features, lesser algorithmic complexity and reduced processing time. Whereas the findings reported here are based on the exclusive use of vowel formants, most of the related previous works used tens or even hundreds of other features. In spite of using a greater level of signal processing, the recognition accuracy reported earlier was often lesser than that obtained by our approach. This method is based on vowel utterances and the first step comprises statistical pre-processing of the vowel formants. This is followed by the identification of the best formants using the KMeans, K-nearest neighbor and Naive Bayes classifiers. The Artificial neural network that was used for the final classification gave an accuracy of 95.6% on elicited emotional speech. Nearly 1500 speech files from ten female speakers in the neutral and six basic emotions were used to prove the efficiency of the proposed approach. Such a result has not been reported earlier for English and is of significance to researchers, sociologists and others interested in speech.*

Keywords: *Artificial Neural Networks, Emotions, Formants, Preprocessing,Vowels.*

I. INTRODUCTION

Emotions constitute an essential part of our existence. Neviarouskaya et al., [1] have observed that emotions are often assigned the role of a sensitive catalyst, which fosters lively interactions between human beings. Over the past few decades the focus of researchers on speech emotion recognition (SER) has progressively increased due to the social acknowledgement of the influence of emotions on the physical as well as mental health of people. Plutchik [2] has pointed out that emotional distress-a common phenomenon in the present world, impels people to seek help. The emphasis placed on emotional quotient, a measure of emotional intelligence (EI), in the wholesome development of an individual is yet another reason for the increased focus on speech emotion recognition. Salovey and Mayer [3] have defined emotional intelligence as the ability to monitor and regulate one's own and other's feelings. As reported by N.Naqvi et al., [4] decision making, which plays a vital role in the behavior of any person, is aided by emotions, especially when these have to be made in the face of uncertainty.

This paper is organized as follows: The motivation for this research is presented in Section 1.1 along with the problem definition. Section 1.2 briefly mentions certain recent, relevant research works in this area. Section 1.3 introduces the segmental units of speech which are the focus of this investigation. The features used in this work are reviewed in Section 1.4. Section 2 outlines the method used in this research. The results are presented in Section. 3 and their significance is discussed. Section. 4 concludes this paper, mentioning major contributions and suggesting directions for future work.

1.1 Motivation for this Investigation and Problem Definition

Research literature on SER abounds with results of speech analysis as given by Moataz El Ayadi et al., [5] and applications of emotional speech, as illustrated by Ramakrishnan S. & El Emary [6]. Underlying all such results was the availability of steadily increasing computational power and also the evolution of signal processing algorithms. Therefore, reducing both the number of features and the complexity of the SER was never in the purview of research in emotional speech, which instead focused solely on obtaining good emotion recognition rates. It is in this context that the authors were motivated to focus both on the methods as well as results so as to achieve reliable and quantitatively superior SER using minimal inputs, features and signal processing. The first step in this direction was obviously the judicious choice of the speech segment. Due to the stand alone nature of certain vowel phonemes in English, it was decided to investigate the emotions in a speech data base comprising such vowel phonemes. Moreover these vowel phonemes represent the minimum input utterances on which reliable SER can be based.

Problem Definition : The main objectives of this investigation were three fold. The first was to achieve accurate speech emotion recognition using minimal length utterances, for which we used the stand alone vowels in English. The second was to identify the optimum formant spectral features for speech recognition with minimum algorithmic complexity. This was done based on the performance of the kMeans, Naïve Bayes and K nearest neighbor classifiers for single formant classes. The third and final objective was to implement efficient speech emotion recognition with the selected formant values given to an artificial neural network classifier.

1.2 Previous Studies in Speech Emotion Recognition

A few relevant works in the processing of emotional speech are briefly introduced here for easy comparison of the results of this investigation. Lee et. al., [7] conducted a study of the averaged tongue tip movement velocity for each of four peripheral vowels sounds (/IY/, /AE/, /AA/, /UW/) in American English as a function of four emotions. Results indicated angry speech to be characterized by greater ranges of displacement and velocity, while it was opposite in sad speech. In a recent SER work done by Morales [8] in Mexican Spanish, HMMs were used for the acoustic modeling of consonants and vowels. The emotional status was detected from the spectrum differences in vowels, though there were confusions between anger and happiness. Hassan and Dampe [9] proposed a hierarchical classification technique called Data-Driven Dimensional Emotion Classification (3DEC) which used binary support vector machines (SVMs) for multiclass classification of emotions. The investigation had been done using 6552 features per speech sample extracted from three databases of acted emotional speech (DES, Berlin and Serbian) and a German database of spontaneous speech (FAU Aibo Emotion Corpus). Investigations in speech recognition based on the Mel-frequency cepstral coefficients (MFCCs) were found to yield superior results when compared to many other features as reported by Wagner et. al. [10] and S. Emerich and E. Lupu [11]

1.3 Segmental Units of Speech – Vowels

A phoneme is a language specific, minimal, meaningful unit of sound. Luengo et al.,[12] observed vowels as one of the most stable segments in a speech signal, making them very appropriate for the computation of certain features in speech recognition as well as speech emotion recognition. The peculiar resonance (formant) gives to each several vowels its distinctive character or quality as a sound of speech. Quateri, T., F [13] has pointed out that the perception of vowels in isolation without the co-articulation effects of neighboring phones is based on their steady state spectra, usually interpreted in terms of the location of the first four formants-F1 to F4. Observed differences in the F1 and F2 distributions of certain vowels suggest varied effects of emotion on the formants of the different vowels. From experimental results, S.Yildirim et. al., [14] concluded that the peripheral vowels other than /IY/ are more affected by changes in emotion. All these indicate that vowels can communicate emotion through language.

1.4 Features- Formants

Formants are the resonant frequencies of the vocal tract and are called so by speech scientists since these resonances tend to "form" the overall spectrum. Ververidis [15] described formants as quantitative characteristics of the vocal tract since the location of vocal tract resonances in the frequency domain, depends upon the shape and the physical dimensions of the vocal tract. It has been found that speakers during stress or under depression do not articulate voiced sounds with the same effort as in the neutral emotional state as pointed out by Tolkmitt and Scherer [16]. A strong dependency of spectral characteristics on phonemes and therefore on the phonetic content of an utterance has been verified by Schuler et al, [17]. Vowels are distinguished primarily by the location of the first three formant frequencies as illustrated by Shaughnessy [18]. Hence the approach adopted in this work takes advantage of these facts by using only formants for SER.

II. METHODOLOGY

Fig. 1 shows the block diagram of the proposed low complexity SER system used to classify emotions based on the first four formants of English vowels. The five vowels a, e, i, o, u, of ten female speakers were recorded in the neutral and six basic emotions to obtain the speech database. The elicitation of authentic emotions from speakers is difficult and restricted to non-extreme states for ethical reasons. Here the subjects were instructed to imagine and re-experience apt situations or events corresponding to each of the seven emotions.

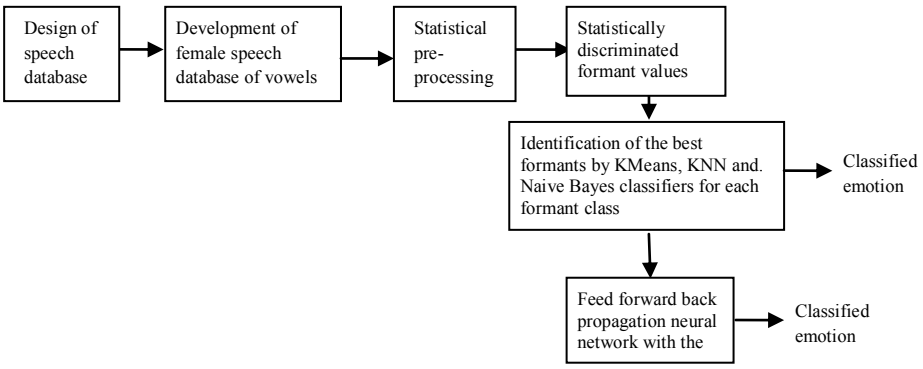

FIG. 1. SCHEMATIC OF THE PROPOSED LOW COMPLEXITY SER SYSTEM

Perceptual listening tests were conducted to ensure the emotional quality of the recorded speech database. The ten voluntary listeners were non-native speakers of English without any known hearing disability. The feedbacks were checked using the identity of the wav files and those files whose emotions were mistaken more than twice were considered unfit and removed from the speech database. The validity of the speech corpus was proven by the overall recognition rate of 92.7% with human listeners.

From each wav file the first four formant values (F1 to F4) for the five English vowels were extracted with the open source speech processing software Praat and tabulated for the seven different emotions . The formants were obtained using the linear prediction method.

Statistical Analysis: Statistical evaluations of the formants were done by analysis of variance (ANOVA).The results of ANOVA indicate the statistical discriminations (at three levels of significance) of the formant values between various emotion classes. Within any specific emotion, the formant values have been reported to change with the specific vowel utterances. Therefore statistical analyses were also carried out to examine and identify significant differences in formant values for the 5 different vowel utterances within the same emotion class. Based on the this, finally only those formant values which reflected the differences in emotions only were used for further classification.

Selection of formants using 3 base classifiers: Separate, formant wise (F1 to F4) classification of emotions using four methods namely Kmeans, Naive Bayes, KNN was done in order to identify the formants that gave the best results in SER. Finally ANN classification was done on the selected formant features. A two-layer feed-forward network, with sigmoid hidden and output neurons and sufficient neurons in its hidden layer was used for the final classification. In this supervised learning technique the network is trained to classify the inputs according to the pre defined targets.

III. Results And Discussion

This section presents the results of the ANOVA, followed by the classification results of the three base classifiers for each formant class. The results of the final classification using ANN are presented in the confusion matrix.

3.1 Results of Statistical Analysis: The formants were examined for statistical difference in feature values by the analysis of variances. Statistical analysis of F1 of vowels revealed anger, sadness and disgust to be better discriminated than the rest of the emotions considered. The repeated measures ANOVA of the logarithm of second formant values showed differences among most of the emotions, at a three star significance level. But happiness could be distinguished from surprise at a low level of significance only. Statistically, anger could not be significantly discriminated from disgust. Neutral and fear were the best discriminated. Statistical analysis of F3 values of vowels revealed sadness to be the best discriminated from all other emotions. Statistical analysis of the logarithm of the fourth formant values of vowels showed two star significance between surprise and disgust. Surprise was better discriminated from the rest of the emotions. When compared with the performance of the first three formants, the statistical discrimination between the seven emotion classes based solely on the F4 values of the vowels was poor. Summarizing the results of ANOVA, all the seven emotions were discriminated well across the four formants, though at different recognition rates.

3.2 Results of Classification of Single Formants: Table 1. below gives the comprehensive results of classification by the Naïve Bayes and the KNN method. With the single formant baseline recognition rate fixed arbitrarily at 20%, each of the four formants recognized almost the same number of cases though for different emotions and with different classifiers. Thus it was inferred that the contribution of each of these four formants is significant and all the four formants were therefore included in the feature set for the final ANN classification. The performance of each of the three classifiers also can be assessed from Table.1, formant wise, for each

emotion. Amongst the three classifiers, the KNN classifier gave the best emotion recognition based on individual vowel formant values and could recognize all the emotions, even though at different rates. This was followed by KMeans classifiers, which could also recognize all emotions. The Naïve Bayes classifier failed to recognize all emotions based on any formant. Based on F1, anger and sadness had high classification rates as obtained with ANOVA. But the low recognition rate of disgust was contradictory to the results of ANOVA. F1 is therefore not recommended for the detection of surprise, fear and disgust.

TABLE 1. SER Rates of various Classifiers for Neutral and Basic Emotions

Formants	Classifier	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
F1	KMeans	30.7%	15.8%	24.5%	21.6%	31.3%	19.5%	6.9%
	NB	37.5%	0%	12.5%	43.8%	87.5%	0%	0%
	KNN	25.7%	36.7%	36.4%	51.5%	9.3%	20.7%	19.4%
F2	KMeans	28.2%	11.7%	22.3%	10.6%	10.6 %	11.7%	22.3%
	NB	70%	0%	50%	0%	0%	0%	0%
	KNN	50%	46.7%	41.7%	30%	40%	28.6%	44.4%
F3	KMeans	30.4%	42.9%	23.2%	16.5%	7.1%	3.6%	33.9%
	NB	0%	50%	0%	10%	0%	30%	10%
	KNN	37.5%	15.4%	15.4%	26.3%	25%	23.5%	33.3%
	KMeans	39.2%	31.4%	13.7%	25.5%	3.9%	23.5%	15.7%
	NB	10%	50%	0%	0%	0%	37.5%	10%
F4	KNN	36.8%	25%	32%	19%	14.3%	31.3%	33.3%

3.3 The final ANN classification based on all four formants

All the formant values were given to the ANN classifier since the results of the various classifiers for single formant classes had revealed that all four formants contributed significantly to SER. Table 2 below presents the recognition rates for various emotions based on all the four formants.

TABLE 2. Confusion Matrix Of The Final Ann Classification Based On The First Four Formants

Emotions	Hap	Surp	Neut	Ang	Sad	Fear	Disg
Happy	100%	0%	0%	0%	0%	0%	0%
Surprise	0%	100%	0%	0%	0%	0%	0%
Neutral	0%	0%	100%	0%	0%	0%	0%
Anger	3.4%	0%	0%	96.6%	0%	0%	0%
Sad	0%	0%	3.6%	0%	*96.4%*	0%	0%
Fear	0%	0%	0%	3.8%	11.5%	84.6%	0%
Disgust	0%	4.2%	0%	4.2%	0%	0%	91.7%

Fear was the least recognized. The overall SER rate was 95.6%. The results of the final ANN classification of the first four formant values of the vowels, validates the consolidated findings of ANOVA as well as the results of classifications by the other three classifiers based on single formant values.

3.4 Comparison with state-of-the-art

The results of this investigation are not directly comparable with state-of-the-art results reported elsewhere in the literature on SER. This is because, even for SER in English itself or based exclusively on vowels, the exact database, feature sets, classifiers differ strikingly. Nevertheless, it is informative to provide certain qualitative comparisons between the current results and the state-of-the-art. Schuller et. al, 2009 [19] have reported a speaker independent SER rate of 88.6% on the Berlin speech database. Hassan and Dampe [9] have reported 79.5% on the Berlin database and 80.1% on the Serbian database using the 3DEC method developed by them. The authors have previously reported an obtained recognition rate of 85.3%, with formant bandwidth features on an English database comprising longer utterances comprising several words [20]. Taking into consideration these cited differences in recognition accuracies; we logically conclude that the vowel formant based approach described here, which achieves 95.6% SER rate, is much superior in precision, robustness and computational efficiency. Thus the final ANN classifier in this class seven speech emotion recognition problem outperformed other classifiers used in this work itself and those reported in English.

IV. CONCLUSION

In this paper, we proposed an efficient system for the automatic detection of seven emotions in English, at the segmental level itself using only the formants of stand-alone vowel utterances. The classification results of the KMeans, Naive Bayes and KNN classifiers for various emotions, based on single formant values

agreed with the results of the repeated measures ANOVA. The specific emotion recognition rates obtained by this approach varied with the emotion considered, order of the formant and the classification method. Very high recognition rate was obtained with the final ANN classifier. This approach can be adapted to implement a truly real time SER system.

Acknowledgement

We gratefully acknowledge all the speakers and listeners who contributed to this investigation.

REFERENCES

[1] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, "Emo Heart: Conveying Emotions in Second Life Based on Affect Sensing from Text*". Advances in Human-Computer Interaction, Vol. 2010*, 13 pages, 2010. [Online] Available: http://dx.doi.org/10.1155/2010/209801.

[2] Robert Plutchik, "The nature of emotions", *American Scientist, Vol. 89,* pp. 344-350, 2001.

[3] Salovey P., & Mayer, J.D. Emotional intelligence. *Imagination, Cognition and Personality, Vol.9*, no.3, pp.185-211, 1990.

[4] Nasir Naqvi, Baba Shiv and Antoine Bechara, The Role of Emotion in Decision Making-A Cognitive Neuroscience Perspective.*Current Directions in Psychological Science, Vol.15*, no.5, pp. 260-264, 2006.

[5] Moataz El Ayadi, Mohamed S.Kamel, Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition, Vol. 44,* pp. 572–587, 2011.

[6] Ramakrishnan S., & Ibrahiem M.M.El Emary, "Speech Emotion Recognition Approaches in Human Computer interaction". *Telecommunication Systems*, *Vol. 52*, *no.3*, pp.1467-1478, 2013. doi:10.1007/S.11235-011-9624.

[7] S. Lee, S. Yildirim, A. Kazemzadeh, S. Narayanan, "An Articulatory study of Emotional Speech Production," *Proceedings of the Eurospeech,* Lisbon, Portugal, pp. 497-500, 2005.

[8] Santiago-Omar Caballero-Morales, "Recognition of Emotions in Mexican Spanish Speech: An Approach Based on Acoustic Modeling of Emotion-Specific Vowels" *The Scientific World Journal*, *Hindawi Publishing Corporation, vol. 2013*, 13 pages. [Online] Available: http://dx.doi.org/10.1155/2013/162093.

[9] A. Hassan, R.I. Dampe, "Classification of emotional speech using 3DEC hierarchical classifier," *Speech Communication, Vol. 54, no.7*, pp. 903–916, 2012.

[10] J. Wagner, T. Vogt, and E. André, "A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech," in *Affective Computing and Intelligent Interaction*, *vol. 4738 of Lecture Notes in Computer Science*, pp. 114–125, Springer, Berlin, Germany, 2007.

[11] S. Emerich and E. Lupu, "Improving speech emotion recognition using frequency and time domain acoustic features," in *Proceedings of the Signal Processing and Applied Mathematics for Electronics and Communications (SPAMEC '11)*, pp. 85–88, Cluj- Napoca, Romania, 2011.

[12] Iker Luengo, Eva Navas, and Inmaculada Hernáez, "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech." *IEEE Transactions on Multimedia*, *Vol. 12, no.6*, October 2010.

[13] Quateri, T., F., *Discrete–Time Speech Signal Processing Principles and Practice,* First Edition, Pearson Education, 2001.

[14] S.Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. An acoustic study of emotions expressed in speech**.** *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* 2004**,** Jeju Island, Korea, pp. 2193-2196, 2004.

[15] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: resources features and methods. *Elsevier. Speech Communication .Vol.48,* pp. 1162–1181, April 2006.

[16] F.J.Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameters," *Journal of Experimental Psychology*, *Vol. 12, no. 3*, pp. 302–313, 1986.

[17] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector Machine - belief network architecture". *International Conference on Acoustics, Speech, and Signal Processing- ICASSP 2004, IEEE ,Vol. 1*, pp. I-577-580, 2004

[18] Douglas O' Shaughnessy, *Speech Communication: Human and machine*. Addison Wesley Publication, 1987.

[19] B. Schuller, F. Vlasenko, B. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances". *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU 2009*), Merano, Italy, pp. 552–557. IEEE, 2009.

[20] Agnes Jacob and P. Mythili, "Minimal feature set based classification of emotional speech" *International Journal of Scientific and Engineering Research*, Vol. 4, Issue11, 2013.(In press).