ISSN: 1813 – 1662 (Print)
E-ISSN: 2415 – 1726 (On Line)

Artificial Neural Network vs. Support Vector Machine For Speech Emotion Recognition

Mohamed. A. Ahmad
Computer Science Dept., Computer Science and Mathematics college, Tikrit University
Mohamed.aktham3@gmail.com

Abstract

Today, the subject of emotion recognition from speech got the attention of many researchers who are interested in the topic of speech recognition and it has engaged in many applications. Furthermore, Speech Emotion Recognition (SER) a pivotal part of influential human interaction and it has been a modern challenge to speech processing. SER has two basic phases; which are, features extraction and emotion classification.

This paper presents a comparison in performances of two popular techniques used for classification that Artificial Neural Network (ANN) as well as Support Vector Machine (SVM). Furthermore, gives a view on the collection of corpus, feature extraction techniques and classification methods, which are regarding to emotion detection from wave speech signal.

I. Introduction

The growing interest in speech emotion recognition (SER) field can be observed through a number of achieved research during the last few years. This is at most encourage by intelligent Human to Machine interaction required for different applications. Besides that, this attention increased due to the progress made in several main areas. The computational power that available in modern computers, like processor based on multi-core, graphical processing units (GPU), increasing capacity of memory and CPU/GPU clusters. All this, and more helped training to became more robust for complex models. SER aims at distinguishing the speaker emotion from his speech signal. Neutral, disgust, happiness, sadness, surprise, anger, and fear are considering basic human's emotions. This state can have effect on speech sound, especially in the supra segmental features, like F0, intensity and temporal characteristics of speech. Since muscle, tension and vocal fold may be raised in several emotional state, thus, may some acoustic features are also affected by the speaker's emotional conditions [1].

Emotion recognition depend on two mainly steps. First step is feature extraction and selected best features, and the second step is classification. many studies achieved to determine features that could identify emotion effectively. There are two main types of features of speech, phonetic and prosodic features [2]. Classification algorithm plays a major role in differentiating between different features.

The major aspects in speech emotion recognition system is choice the suitable features for speech representation like Formant frequencies and spectra temporal features [3]. Along with these features, many state of- the-art derived features like Mel-Frequency Cepstral Coefficients (MFCC) [4], and Linear Prediction Coding are widely used for emotion detection propose [5]. Classification methods have an important role in distinguishing among different features. These are some examples of widely used algorithms for classification process, Artificial Neutral Networks which has been known in speech recognition since late 1980s [6], Gaussian Mixture Model(GMM) [7], Hidden Markov Model (HMM) [8], K- Nearest Neighbor [9], and Support Vector Machine [10]. We have chosen Neural Networks (NN) and Support vector machine for our research work and applied on Berlin Database of Emotional Speech.

II. System of emotion recognition

SER is an in-depth analysis system, where the system can extract essential features, which carry on emotional information from the utterance, then choose proper pattern recognition algorithms to detect emotional states. Typically, the system of SER consist of four major stages, which are speech database, extraction features, features selection and classification. Figure 1 shows the typical steps in SER system.

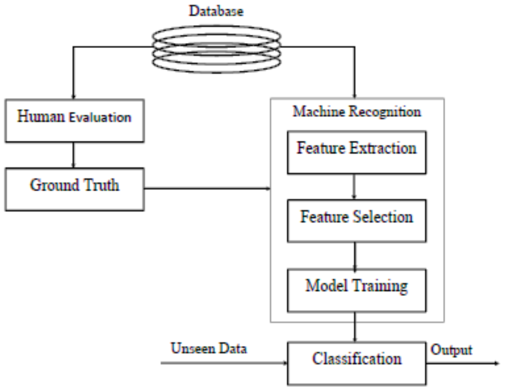

Figure 1: Components of a typical SER system.

III. Speech Emotion Databases

The first step for a SER system is providing the databases of speech signal in order to detect human emotions from it. There are many emotionality speech databases, which have been used by many researchers, related to emotion recognition. Table 1 shows some details of the most currently used databases [6 -19].

Table 1: Summary of popular speech emotion databases

Name	Size	Source	Emotions
Danish emotional speech(1996)	4 speakers, 260 utter.	Nonprofessional actors	Anger, joy, sadness, surprise, neutral
SUSAS (1997)	32 speakers, 16,000 utter.	simulated and actual stress	Stress, fear, anxiety, anger
Interface database (2002)	42 speakers, 2345 utter.	Actors	Anger, disgust, fear, joy, surprise, sadness, neutral
KISMET (2002)	3 speakers, 1002 utter.	Nonprofessional actors	Approval, attention, prohibition, soothing, neutral
BabyEars (2003)	12 speakers, 509 utter	Mothers and fathers	Approval, attention, prohibition
Berlin emotional speech databases(2005)	10 speakers, over 500 utter.	Professional actors	Anger, joy, sadness, fear, disgust, boredom, neutral
FAU Aibo emotion Corpus (2009)	51 speakers, 18,216 utter.	Children	Angry, emphatic, neutral, positive and rest
SAVEE database (2010)	4 speakers, 480 utter.	Nonprofessional actors	Anger, disgust, fear, happiness, sadness, surprise, neutral
SEMAINE database(2010)	20 speakers, 50,350 words	Conversation with human operator	Anger, disgust, amusement, happiness, sadness, contempt

IV. Feature extraction

Feature is a tiny unit, which distinguish classes that are maximally closed to each other. Feature extraction process generally assist to reduction of the amount of dataset. From another side, feature extraction used for extracting certain features from the utterance, these features carry the attribute of the speech signal which are good for identifying the different speech signals, thus, these features will play the main role in speech recognition.

The different sets of acoustic features, which are used by many researchers for developing their SER systems can be classified, depend on two main categories: standard prosodic features and spectral feature.

Prosodic features, like pitch, energy, intensity and rhythm, which are commonly used in SER, have been shown very useful to specify emotional information of the speaker [20, 21]. Spectral features, on the other hand, the spectrum of speech signal may help to obtain emotion state of speaker. These features convey the frequency contents of the signal and provide complementary information for prosodic features [20]. Formant frequencies F1 and F2 are used by authors in [22, 23] Also, Mell Frequency Cepstral Coefficients (MFCCs) [21, 22, 24, 25], Linear Prediction Coding (LPC) [24, 26], Weighted MFCC (WMFCC) and Linear Spectral Frequency (LSF) [25], Sub-Band features [22, 24, 26], Modulation Spectral Features (MSF) and Perceptual Linear Prediction (PLP) [27]. All these have been categorized as an effective technique to extract features depends on spectral in SER system.

V. Selection features

The data size, which is acquired from features extracted process, is almost big, therefore, there is need to attempt reduce size of data by electing the most important features that distinguish the emotion. Selection techniques are applied mainly for two purposes: shorter training times and enhanced generalization by reducing over fitting simplification of models in order to make researchers interpret them simply.

Several methods are serve for feature selection which can belongs to one of the following three methods: Filtering, wrapping and embedding methods [28]. Filter based algorithms characterized by simplicity, where it does not require complex calculations. Besides that, features evaluate separately, and independent of the used classifier. Fisher Discriminant Ratio (FDR) [29] and Information Gain Ratio (IGR) [23] are widely used for selection proper feature. Contrary to filter technique, Wrappers method is considered by complex calculations but it has advantage of considering the collection effects of features and classifier properties the combination features effects and classifier characteristics. In this paper the sequential forward selection methods (SFS) was used to reduced number of features by select which feature convey emotional data [31].

VI. Classification

For classification stage, there are many techniques which are applied to test data from feature vector set, like Support Vector Machine (SVM) [30, 32], K-Nearest Neighbor (KNN) [33, 34], Gaussian Mixture Model (GMM) [35], Hidden Markov Model (HMM) [36], Artificial Neural Networks (ANN) [37, 38], and more. Each one of this classifier has advantages and disadvantages. In this paper, ANN and SVM are used for classification task in order to make a comparison depend on the accuracy of each method depends on Confusion matrix and Receiver Operating Characteristic (ROC) curve.

VII. Experiment

For start Implementation, first we need to provide emotion features vector derived by low-level descriptors (LLDs) from speech waveforms signals. Using 408 utterances from Berlin emotional speech database, categorized as anger (127), neutral (79), fear (69), happiness (71), and sadness (62) as shown in Figure 2. The reason of using Berlin dataset here, because it is considering a most popular database employ for emotion recognition. We extract features from each speech sample using OpenEAR toolkit [39]. Each sample is dividing into several frames of equivalent length then calculate (68) LLD as descript in table 2 (a). Delta and double delta functions are computed for each one of LLD. Also that, 39 statistical functional are applied on each one of LLD, delta and the delta delta functions to provide 7956 features from (68 + 68 + 68) ×39. Table 2 (b), displays the statistical functions.

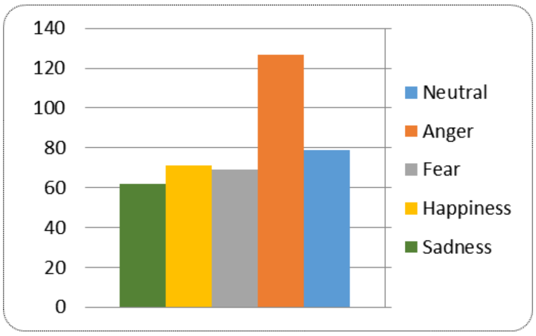

Figure 2: five-emotion class of Berlin dataset.

Table 2 (a): Description of a set of 68 low level descriptors (LLDs).

Feature kind	68 Group of Features
Pitch	Pitch (fo) in Hz and its smoothed contour
Energy	Log Energy per frame Energy in frequency bands 0 – 250 Energy in frequency bands 0 – 650 Energy in frequency bands 250 – 650 Energy in frequency bands 1000 – 4000 Energy in 26 mel-frequency bands
Zerro-crossing rate	Number of zeros crossings and mean ZCR
Cepstrum	13 Mel-frequency cepstrum coefficients
Formants	First three formants and their corresponding bandwidths
Spectral,	Centroid, flux, position of spectral max. and min. peaks, spectral roll of points 90%, 75%, 50% and 25%
Voice Quality	Jitter and shimmer, harmonics to noise ratio, probability of voicing

The sequential forward selection methods (SFS) was used for selection the most important features were obtained from classification stage, which SFS method help to reduce number of features. As result of applying SFS method the number of features becomes 3000. Now the data is ready to be classified.

Table 2 (b): Description of 39 statistical functional derived from each LLD.

Functional (39)	Number
Relative positions of max./min values	2
Range (max − min)	1
Arithmetic and quadratic means	2
Quartile and inter-quartile ranges	6
5 and 85 percentile values	2
Zero crossings and mean crossing rate	2
Number of peaks and mean distance between peaks	2
Arithmetic mean of peaks	1
Overall arithmetic mean	1
Linear regression coefficients and corresponding approx. error	4
Quadratic regression coefficients and corresponding approx. error	5
Centroid of contour	1
Standard deviation, variance, kurtosis, skewness	4
Arithmetic, quadratic and absolute means	3
Arithmetic, quadratic and absolute means of non-zero values	3

Multi-Layer Perceptron Neural Network (MLPNN) is used for the training and testing the 3000 features we have. The MLPNN is feed forward network (FF) depend on back propagation algorithm in trained process. The model is used in this paper consist of input layer, one hidden layer and output layer. The nodes number of each layer are 3000, 122 and 5 for input, hidden and output layer respectively. Randomly divided the 3000 samples to 2100 samples for training, 450 samples for validation and 450 samples for testing. The best validation performance is 0.13256 at epoch 18 as shown in Figure 3.

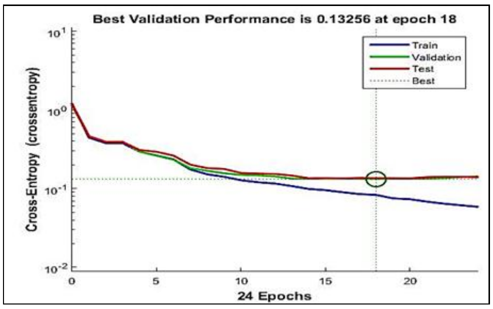

Figure 3: The MLPNN Training performance.

Figure 4 shows the MLPNN model. The model present 91.2% accuracy rate. The confusion matrices by means of MLPNN is shown in table 3. Figure 5, shows ROC curve for each one of five classes.

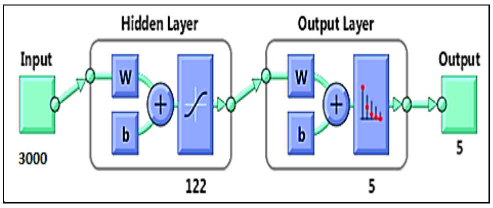

Figure 4: The model of MLPNN classification.

Table 3: Confusion matrix of the MLPNN classifier

	Anger	Fear	Happy	neutral	sadness
Anger	513 37.2%	10 0.7%	21 1.5%	2 0.1%	6 0.4%
Fear	11 0.8%	141 10.2%	2 0.1%	2 0.1%	1 0.1%
Happy	22 1.6%	0 0.0%	188 13.6%	3 0.2%	0 0.0%
neutral	6 0.4%	6 0.4%	2 0.1%	288 20.9%	15 1.1%
sadness	2 0.1%	0 0.0%	3 0.2%	7 0.5%	128 9.3%

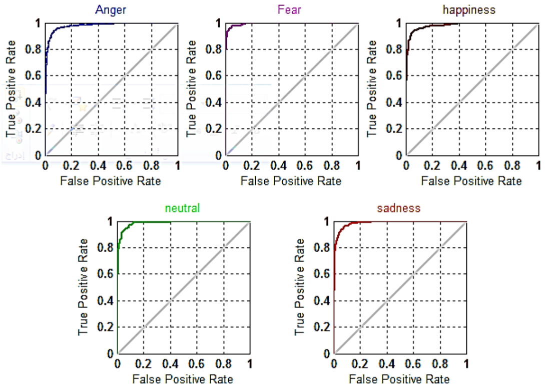

Figure 5: ROC curve of MLPNN classification for five emotion classes.

Another method, which is used for the training and testing the feature vector in this paper, is SVM based on cross validation with 10-fold. The Gaussian kernel functions is chosen to SVM at cost value c equals to 4, with kernel scale σ = 0.95 and experiments are performed using the one-against-all strategy for multi-class classification. It gives recognition rate of 86.3%. Table 4 display the confusion matrices of Gaussian kernel SVM and Figure 6, shows ROC curve for SVM each one of five emotion state are used foe emotion recognition.

Table 4. Confusion matrix of the Gaussian kernel SVM classifier

	Anger	Fear	happiness	neutral	sadness
Anger	503 36.5%	8 0.6%	6 0.4%	11 0.8%	22 1.6%
Fear	11 0.8%	115 8.3%	0 0.0%	23 1.7%	1 0.1%
Happy	15 1.1%	4 0.3%	123 9.6%	8 0.6%	2 0.1%
neutral	15 1.1%	11 0.8%	4 0.3%	268 19.4%	4 0.3%
sadness	35 2.5%	1 0.1%	2 0.1%	4 0.3%	174 12.6%

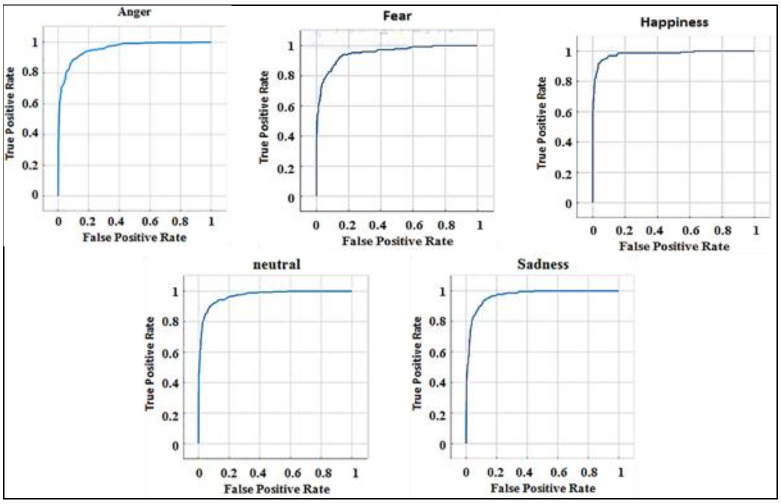

Figure 6. ROC curve of SVM classification for versus five-emotion state.

Conclusion

In this paper, we used the SVM and NN which are most famous classification techniques to distinguish five basic emotions from berlin speech signal and compare their performance.it clearly shows that the ANN surpass on the SVM in field of speech emotion recognition. More precisely, ANN compared to SVM classification, the accuracy of the ANN is 4.9% higher than SVM (ANN gives 91.2%, while SVM gives 86.3%). However, SVM takes shorter executing time than ANN does during learning period. Neutral and anger are distinguished well in both classifiers. While the fear, happiness and sadness got different recognition rate for each method.

References

[1] Katagiri, S., “Handbook of Neural Network for Speech Processing," Artech House Signal Processing Library, October 2000.

[2] Yashaswi, M., Nachamai M. and Joy P., “A Comprehensive Survey on Features and Methods for Speech Emotion Detection,” In Proceedings of International Conference on Electrical, Computer and Communication Technologies (ICECCT) , IEEE, pp. 1-6, 2015.

[3] S. Wu, H. Tiago “Automatic Recognition Of Speech Emotion Using Long-Term Spectro-Temporal Features,” In Proceedings of 16th International Conference on Digital Signal Processing, IEEE, pp. 1-6, 2009.

[4] A. B. Kandali, S. Member, A. Routray, and T. K. Basu, “Emotion recognition from Assamese speeches using MFCC features and GMM classifier,” In Proceedings of TENCON 2008 - 2008 IEEE Region 10 Conference, IEEE, pp. 1-8, 2008.

[5] Ververidis, D., and Kotropoulos, C., “Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections,” In Proceedings of the International Conference on Signal Processing Conference, 4th European, IEEE, 8 Sept. 2006.

[6] Sweeta B., Amita D. “Emotional Hindi Speech: Feature Extraction and Classification,” In Proceedings of International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp. 1865 - 1868, 2015.

[7] A. B. Kandali, A. Routray, and T. K. Basu, “Vocal emotion recognition in five languages of Assam using features based on MFCCs and Eigen Values of Autocorrelation Matrix in presence of babble noise,” Commun. (NCC), 2010 Natl. Conf., 2010.

[8] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” Int. Conf. Multimed. Expo. ICME ’03. Proc. (Cat. No.03TH8698), vol. 1, pp. 1–4, 2003

[9] Y. Pan, P. Shen, and L. Shen, “Speech Emotion Recognition Using Support Vector Machine,” In Proceedings of Electronic and Mechanical Engineering and Information Technology (EMEIT), International Conference on (Volume:2) 621 - 625, 2012.

[10] M. Dumas, “Emotional Expression Recognition using Support Vector Machines.” In Proceedings of International Conference on Computing , IEEE, pp. 1865 - 1868, 2015.

[11] Yi-Lin, L. and Gang, W., “Speech emotion recognition based on HMM and SVM, ” In Proceedings of the International Conference on Machine Learning and Cybernetics Vol.:8, IEEE, 21 Aug. 2005.

[12] Hansen, J. and Bou-Ghazale, S., “Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database,” Proc. EUROSPEECH-97, Rhodes, Greece, Vol. 4, pp. 1743-1746, 1997.

[13] Hozjan, V., Kacic, Z., and Moreno, A., “Interface databases: Design and collection of a multilingual emotional speech database,” In Proceedings of 3rd International Conference on Language Resources and Evaluation, pp. 2024–2028, Canary Islands, Spain 2002.

[14] Breazeal, C. and Aryananda, L. “Recognition of affective communicative intent in robot-directed speech,” Autonomous Robots Vol. 12, Issue 1, pp. 83-104, 2002.

[15] Malcolm, S., and Gerald, M., “BabyEars: A recognition system for affective vocalizations,” Speech Communication Vol. 39, Issues 3–4, pp. 367–384, science direct, February 2003.

[16] Burkhardt, F., Paeschke A., and others, “A Database of German Emotional Speech,” In Proceedings of the Ninth European Conference on InterSpeech, pp.1517–1520, Citeseerx, 2005.

[17] A. Batliner, S. Steidl, E. Noth, “Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus,” In Proceedings of the Corpora for Research on Emotion and Affect Workshop, 2008.

[18] Vinay, G. and Mehra, A., “Gender specific emotion recognition through speech signals,” In Proceedings of the International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, pp. 727 – 733, 20-21 Feb. 2014.

[19] McKeown, G., Valstar, M. , Cowie, R. and more authors, “The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent,” Affective Computing, IEEE Transactions on Vol. 3, Issue 1, IEEE, pp. 5 – 17, 09 April 2012.

[20] ElAyadi, M., Kamel, M. and Karray F., “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, Vol. 44, Issue 344, pp.572–587, 2011.

[21] Rong, J., Li, G., and Phoebe, P., “Acoustic feature selection for automatic emotion recognition from speech,” Journal of Information Processing and Management, Vol. 45, pp.315–328, 2009.

[22] Clavel, C., Vasilescu, I., and others, “Fear-type emotion recognition for future audiobased surveillance systems,” Journal of Speech Communic - ation, Vol. 50, Issue 6, pp. 487–503, 2008.

[23] Polzehl, T., Schmitt, A., Metze, F. and M. Wagner, “Anger recognition in speech using acoustic and linguistic cues,” Journal of Speech Communication, Vol. 53, pp. 1198–1209, 2011.

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

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

[26] Perez-Espinosa, H., Reyes-Garcia, C. and Villasenor-Pineda, L., "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model," Journal of Biomedical Signal Processing and Control, 2011.

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

[28] Guyon, I., Elisseeff A., "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, Vol. 3, pp. 1157-1182, 2003.

[29] Xiao, Z., Dellandrea, E., Dou, W., and Chen, L., "Features extraction and selection for emotional speech classification," In Proceedings of International conference on advanced video and signal based surveillance (AVSS), IEEE, pp. 411–416, 2005.

[30] Yashpalsing, C., Dhore, M., and Pallavi Y., "Speech Emotion Recognition Using Support Vector Machine," International Journal of Computer Applications, Vol. 1, 2010.

[31] A. Jain , D. Zongker, "Feature selection: evaluation, application, and small sample performance," In Proceedings of International Conference on Transactions on Pattern Analysis and Machine Intelligence (Volume:19, Issue: 2), IEEE, pp. 153-158, 2002.

[32] Yixiong, P., Peipei, S. and Liping, S., "Speech Emotion Recognition Using Support Vector Machine," International Journal of Smart Home, Vol. 6, No. 2, April 2012.

[33] Leif E., "K-nearest neighbor," http://scholarpedia.org/article/K-nearest_neighbor, 2009.

[34] Muzaffar, K., Tirupati, G., Mohmmed, N. and Ruhina Q., "Comparison between k-nn and svm method for speech emotion recognition," International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2, Feb 2011.

[35] Yun, S. and Yoo, C., "Loss-scaled large-margin Gaussian mixture models for speech emotion classification," Audio Speech and Language Processing, Transactions on, Vol. 20, no. 2, IEEE, pp. 585–598, 2012.

[36] Schuller, B. and Rigoll, G., "Hidden Markov model-based speech emotion recognition," In Proceedings of the Acoustics, Speech, and Signal Processing International Conference Vol. 2, IEEE, pp. 1-4, 6-10 April 2003.

[37] J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks," Neural Computing and Applications Vol. 9, Issue 4, pp 290-296, Springer, 2000.

[38] M. Bhatti, W. Yongjin, and G. Ling, "A neural network approach for human emotion recognition in speech," In Proceedings of the International Symposium on Circuits and Systems ISCAS'04, Vol. 2, pp. II–181–4, 2004.

[39] F. Eyben, M. Wollmer, and B. Schuller, "Open EAR - Introducing the Munich open-source emotion and affect recognition toolkit," In Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction, ACII'09, IEEE, pp. 1–6, 2009.

مقارنة الشبكات العصبية الاصطناعية مع الة دعم المتجه في تمييز العاطفة من الكلام

محمد اكثم احمد

قسم علوم الحاسوب ، كلية علوم الحاسوب والرياضيات ، جامعة تكريت ، تكريت العراق

Mohamed.aktham3@gmail.com

الملخص

ان موضوع تمييز العاطفة من الكلام في الوقت الحالي حاصل على اهتمام العديد من الباحثين المهتمين بمجال تمييز الكلام، وان تمييز العاطفة متداخل مع العديد من التطبيقات. علاوة على ذلك، تمييز العاطفة في الكلام هو جزء محوري من التفاعل الإنساني المؤثر وقد أصبح يشكل تحديا جديدا في معالجة الكلام. ان نظام تمييز العاطفة يتكون من مرحلتين أساسيتين، هما استخراج الميزات والتصنيف. الهدف من هذا البحث هو اجراء مقارنة لأداء اهم التقنيات التي تستخدم في التصنيف والتي هي: الشبكة العصبية الاصطناعية (ANN) والة دعم المتجه (SVM). بالإضافة الى اعطاء نظرة عامة على مجموعة البيانات، الصفات، وتقنيات التصنيف التي ترتبط في تحديد العاطفة من الكلام.