Evaluation of Speech Emotion Recognition Based on Support Vector Machines with Gaussian Mixture Model Super Vectors[1]

Maria Art Antonette D. Clariño* and Mark M. Alfonso
Institute of Computer Science
University of the Philippines Los Baños
College, Laguna, Philippines 4031
(049) 536-2302
**mdclarino@up.edu.ph*

Abstract: This paper presents multi-class speech emotion recognition developed using Support Vector Machines (SVM) with Gaussian Mixture Model (GMM) super vectors. Input to the system is in the form of speech utterances. Seven emotions (Happiness, Sadness, Anger, Fear, Surprise, Disgust, and Neutral) were considered in this study. For each of these emotions, feature extraction and normalisation were implemented using Praat Scripting Language (PSL) to compute for the average pitch and intensity and to perform batch processing of the training data. The processed features were then subjected to SVM-GMM for modelling and classification. A data set comprising of 175 speech utterances collected from selected individuals were divided into 140 training and 35 test data. To measure the performance of the developed SVM-GMM classifier, performance measures such as precision, recall, F score, accuracy and error rate were calculated. Two levels of averages were computed: micro- and macro-averaging. The initial data set resulted to macro-averages of 54.29% recall, 48.06% precision, 50.99% F Score, 86.94% accuracy, and 13.06% error rate. A second run was implemented using labelled speech data from the website of Center for Empathic Human-Computer Interaction of De La Salle University increasing the training data to 1750 and test data of 175 speech utterances. This resulted to higher macro-averages of 96% recall, 96.66% precision, 96.33% F Score, 98.86% accuracy and lower error rate of 1.14%.

Key Words: Speech Emotion Recognition; Gaussian Mixture Model; Support Vector Machine; Performance Measures

[1] The paper is a continuation of a research work presented in the 11th National Natural Language Processing Research Symposium entitled "Speech Emotion Recognition Using Support Vector Machines and Gaussian Mixture Model Super Vectors".

1. INTRODUCTION

As the field of Artificial Intelligence (AI) and Machine Learning progress, computers and computer-based systems are made to recognize, behave and understand closer to how humans do. Computer vision, for example, enables the computer to "see" and process relevant information from an image. With this, it serves as an eye and brain for the computer since it does not only capture the image but also extract information resulting to interpretations. Machines are trained by feeding them information coming from features selected a priori and constructed description vectors pertaining to each class or category. These features serve as labels that are used for classifying incoming data (Clariño et al., 2012). This approach is a supervised machine learning where the feature set and its descriptors where classification will be based are determined before the actual classification. It has two stages: training and testing. The same approach can be made to speech emotion recognition.

The processing of speech is a necessary step before emotion can be correctly detected and identified. One of the reasons of unsatisfactory performance of emotion recognition systems from previous studies is the difficulty with the appropriate identification of a feature space to be used in classification. This improvement in the human-computer interaction leads to better computing. Through speech and voice recognition, extracted features from the speech will be used as inputs in detecting emotion.

Emotion is a very important factor in human-computer interaction (HCI). An important feature of HCI is to identify emotional states as prescribed by the signals (Fragopanagos & Taylor, 2005). If a machine can be taught to infer the user's emotion, it can produce useful applications such as in call centers (Yacoub et al., 2003). It can improve the services provided to the clients as a reaction to their emotional state. Improving HCI with emotion detection can play a major role in increasing the accuracy of the decision making of the machine or of another human interacting with the machine when dealing with the needs of the user. Speech is the most natural method of interaction between humans in which emotion can be expressed (El Ayadi et al., 2011). Many studies and experimentations have been done to improve the accuracy of speech emotion recognition. Different modelling methods, mathematical computations and emotion classifiers have been used to detect emotion from speech.

After training and testing, evaluation has to be performed to properly assess the success of the developed classifier. This stage is also necessary to serve as benchmarks for improvement by succeeding studies. Performance measures, to properly assess the classifier, have to be tailored according to the type of classification task performed. The results of the classification task are reflected to a confusion matrix from which measures such as precision and recall can be computed (Sokolova & Lapalme, 2009). It is not only important to know how much of the data set is correctly classified but also to know which data are misclassified or falsely classified and what could be causing the confusion.

This study aims to present evaluation of SVM-GMM based speech emotion recognition. Specifically, a stand-alone program/system must be developed to process human speech, to extract features from the recorded utterances, to model each selected emotion with these features using Gaussian Mixture Model (GMM) Super Vectors, to implement actual classification using Support Vector Machines, and to subject testing data sets to a multi-class classification evaluator

generating performance measures.

2. REVIEW OF LITERATURE

Several studies in speech emotion recognition utilized varying feature set as emotion descriptors, classification methodology (different methods under supervised machine learning), set of emotions considered, and performance evaluation of the system developed. A comparative study on classifiers (Iliou & Anagnostopoulos, 2010) was presented in 2010. A year after, a survey on speech emotion recognition covering the features, classification schemes, and databases was conducted (El Ayadi et al., 2011).

Using pitch and energy features, emotion recognition was implemented using Hidden Markov Models (HMM) (Nogueiras et al., 2001). The overall accuracy across all emotions was computed resulting to an accuracy exceeding 80%. Few years after, another study on speech emotion recognition using HMM was presented (Nwe et al., 2003). Short time log frequency power coefficients (LFPC) were used as speech descriptors. This yielded to an average accuracy of 78% in classifying six emotions. Anger, Disgust, Fear, Joy, Sadness and Surprise were considered as emotion set (Nogueiras et al., 2001; Nwe et al., 2003) but Neutral could also be added as an emotion class (Nogueiras et al., 2001).

In 2007, two novel aspects to speech emotion recognition were presented (Cichosz & Slot, 2007). The first aspect is the selection of emotional speech descriptors. According to Cichosz & Slot (2007), speech characteristics can be classified into three groups: frequency characteristics (e.g. pitch and pitch-derived measures), energy descriptors (energy of utterance), and temporal features (e.g. utterance duration and pauses). These features were used to describe six emotions: joy, anger, boredom, sadness, fear, and neutral. In a more recent study, El Ayadi et al. (2011) presented 4 speech feature categories: continuous, qualitative, spectral, and TEO (Teager energy operator)-based. The other novel aspect presented by Cichosz & Slot (2007) is the use of a binary decision tree utilizing the three element vector at every node and employing an exhaustive search. For evaluation, recognition accuracy was computed.

Metze et al. (2010) presented two approaches to identifying emotions from speech: emotional salience classifier and bag of words classifier. Emotional salience recognizes that certain words are strongly associated to a specific emotion and less to others (Lee & Narayanan, 2005; Metze et al., 2010). A set of most salient uni-grams were presented stating that most of them are related to negative emotions (like grudge). The second approach is more quantitative representing emotion in numeric feature space. The acquired features are subjected to Support Vector Machines (SVM) and Discriminative Multinomial Naïve Bayes (DMNB). Weighted and unweighted recalls were used to evaluate and were compared to baseline on references.

One good application of speech emotion recognition is in interactive voice response systems used in call centers (Yacoub et al., 2003). Because of the nature of the system, utterance-level features related to pitch, loudness, and segments were calculated. Feature extraction takes place in the signal level without considering the data obtained from the recognition of the speech. For the validation of the result, a 10-fold cross validation technique was used where the training data were divided into ten sets randomly. Artificial Neural Networks, Support Vector Machine, 3-Nearest Neighbors and decision trees were used to

compare the results. Weka toolkit was used in the experiments. The developed system by Yacoub et al. (2003) resulted to more than 90% accuracy distinguishing between hot anger and neutral utterances. These two emotions are important in call centers.

A more recent study proposed improvements to Gaussian Mixture Models (GMM) based features improving the classification abilities of relative wavelet packet energy and entropy features most especially in multi-class classification (Muthusamy et al., 2015). A paired t-test was performed to compare emotion recognition rates for comparing raw and enhanced features resulting to improved classification rates with the enhanced features.

The previous studies presented show how important feature selection, method/s used in classification, and evaluation of performance measure supporting claim of improvement achieved or success in satisfying the objective of the study.

3. METHODOLOGY

Five major steps are implemented in this study to implement speech emotion recognition as shown in Figure 1. Recorded speech is subjected to the following processes to detect the emotion involved: (a) feature extraction, (b) feature normalisation, (c) emotion modelling and (d) classification. An additional but necessary step (e) is included to measure the performance of the proposed algorithm.

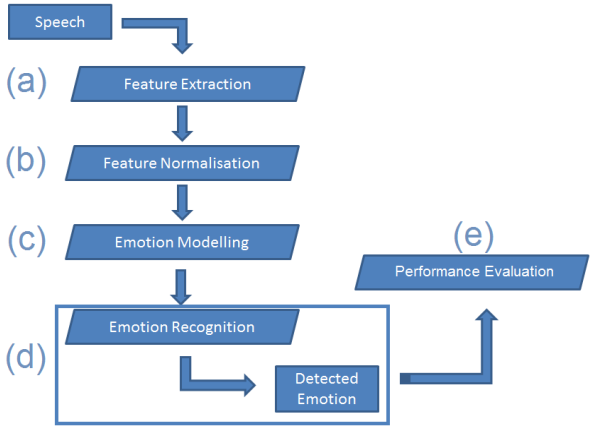

Fig. 1. Flow chart of the overall process

3.1 Emotion Selection and Data Gathering

Before utterances were recorded, the set of emotions must first be finalized. Figure 2 shows the locations of the selected emotions in the activation-valence space (Busso, Bulut, & Narayanan, 2013; Cowie & Cornelius, 2003; Cowie et al., 2001). At least one representative emotion per quadrant was considered except for the 4th quadrant that covers the positive-passive area. Emotions that belong to this category include calm and content (Roesch et al., 2006). Most studies use the "big six" emotions-(happy, sad, angry, fear, surprise, and disgust) (Cowie & Cornelius, 2003; Cornelius, 1996; Nwe et al., 2003), which are found in the first 3 quadrants. Other studies (Shahzadi et al., 2015) considered 5 out of the six commonly used emotions. In place of Surprise, Boredom, a 3rd quadrant emotion (Busso, Bulut, & Narayanan, 2013; Shahzadi et al., 2015), was used and Neutral was also added. The Disgust** in Figure 2 is classified with the emotion Sadness (Shahzadi et al., 2015). A smaller set (Angry, Happy, Sad, and Neutral)

could also be considered (Gomes & El-Sharkawy, 2015). In this study, the "big six" and Neutral are selected to comprise the emotion set (Nogueiras et al., 2001).

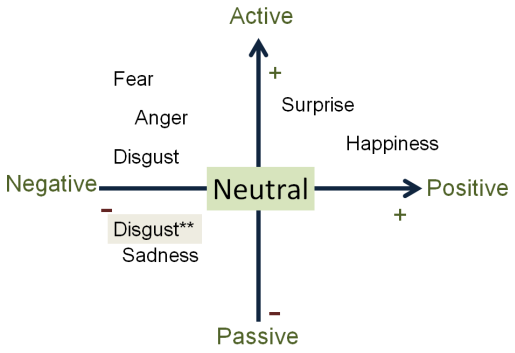

Fig. 2. Activation-Valence Space

After considering all these studies, the utterances for the selected emotions were recorded using the microphone in a HS-5P A4Tech headset. A total of 175 (.wav) sound files were recorded from 25 selected individuals for each of the 7 emotions. Out of the 175 files, 140 were used as training data and the remaining 35 as test data.

3.2 Feature Extraction

An important aspect in any classification problem is the selection of which feature/s to use (El Ayadi et al., 2011). For extracting features from the recorded utterances, Praat script (Boersma, 2002) was used to extract the following classes of features: a) vocal intensity, b) vocal frequency, c) vocal quality and d) vocal resonance. The speech features extracted are then measured according to amplitude or envelope of signal, periodicity of signal and spectral energy distribution. These audio features are generally embodied in the pitch and intensity. The average pitch and average intensity are used as the main features for this study.

3.3 Feature Normalisation

Normalisation is a necessary step to reduce phonetic variability (due to linguistic dependence of emotion information) and speaker identity (attributes of the source) resulting to variability in the speech (Busso, Mariooryad, et al., 2013). There can be anomalies that could cause confusion in modelling. It is done by using warping, a module available in Praat (Boersma, 2002), where the signal is being elongated to get rid of the targeted variability.

3.4 Emotion Modeling and Classification

The numerical values from feature extraction and normalisation were patterned in the emotion model and computed using Support Vector Machine (SVM) with Gaussian Mixture Model (GMM) Super Vectors. The combination is illustrated in Figure 3 (Epps et al., 2010).

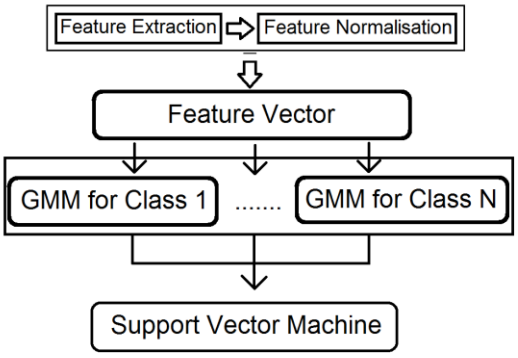

Fig. 3. SVM-GMM Classifier

The formula is provided (Epps et al., 2010):

$$p(X)=\sum_{m=1}^{M} w_m \frac{1}{(2\pi)^{K/2}|C_m|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_m)^T C_m^{-1}(x-\mu_m)\right)$$

where w_m, μ_m, and C_m are the weight, mean vector, and covariance matrix of the m-th mixture respectively.

Through the GMM, a class-dependent mixture for each feature is used for modelling each emotion X. A generative model for the discriminative classifier capability of the SVM is created. As prescribed by Epps et al. (2010), SVM-GMM performs better than the individual performances of SVM and GMM.

3.5 Performance Evaluation

Confusion matrix (Table 1) is used in this study to present properly the classified data. Possible trends in misclassification, closely related emotions, and some deviances may be discovered using the matrix.

Table 1. Sample confusion matrix

		PREDICTED	
		POSITIVE	NEGATIVE
ACTUAL	POSITIVE	True Positive	False Negative
	NEGATIVE	False Positive	True Negative

From the confusion matrix, true positive (correctly classified/detected), true negative (correctly not detected), false positive (incorrectly classified/detected) and false negative (incorrectly not detected) can be obtained to compute for recall, precision, F score, accuracy and error rate. These performance measures were obtained relative to each emotion.

The values of *tp* (true positive), *tn* (true negative), *fp* (false positive), and *fn* (false negative) are counted per j-th emotion/class from the total C=7 classes. Accuracy measures the effectiveness of a classifier in terms of detections in agreement with the actual classifications (Sokolova & Lapalme, 2009). Precision considers false detections. It is given by the number of correct detections over all detections. High precision means having little wrong detection, i.e. detections are done only when they should be made (Clariño et al., 2012). Recall gives the effectiveness of identifying labels per class (Sokolova & Lapalme, 2009). F Score is the harmonic mean of precision and recall while Error Rate measures the overall classification error. The formula for these performance measures are shown in Table 2 (Sokolova & Lapalme, 2009).

Table 2. Set of performance measures for multi-class classification (Sokolova & Lapalme, 2009)

Measure	Macro-level	Micro-level
Recall	$Recall_M = \frac{\sum_{j=1}^{C} \frac{tp_j}{(tp_j + fn_j)}}{C}$	$Recall_\mu = \frac{\sum_{j=1}^{C} tp_j}{\sum_{j=1}^{C} (tp_j + fn_j)}$
Precision	$Precision_M = \frac{\sum_{j=1}^{C} \frac{tp_j}{(tp_j + fp_j)}}{C}$	$Precision_\mu = \frac{\sum_{j=1}^{C} tp_j}{\sum_{j=1}^{C} (tp_j + fp_j)}$
F Score	$\frac{(\beta^2+1) Precision_M Recall_M}{\beta^2 Precision_M + Recall_M}$	$\frac{(\beta^2+1) Precision_\mu Recall_\mu}{\beta^2 Precision_\mu + Recall_\mu}$
Ave. Accuracy	$\frac{\sum_{j=1}^{C} \frac{tp_j + tn_j}{tp_j + fn_j + fp_j + tn_j}}{C}$	

Measure	Macro-level	Micro-level
Error Rate	$\frac{\sum_{j=1}^{C} \frac{fp_j + fn_j}{tp_j + fn_j + fp_j + tn_j}}{C}$	

As shown in Table 2 (Sokolova & Lapalme, 2009), there are two levels of averaging for assessing a multi-class classification. Micro-averaging is the cumulative sum of counts then calculating the performance measure as compared to macro-averaging, which is the average of the measures per class.

4. RESULTS AND DISCUSSION

The 140 utterances were labelled to generate the emotion model. The other 35 utterances where tested to generate the confusion matrix of the emotion detection shown in Table 3. The data set for testing consists of 5 expected occurrences of each of the 7 emotions as shown by the total of each row. Correct classification is found in the main diagonal of the matrix. The summary of results from the 35 sample data is shown in Table 4.

Table 3. Confusion Matrix of 35 samples
Legend: **HA**-*Happy* **SA**-*Sad* **AN**-*Angry* **FE**-*Fear* **SU**-*Surprise* **DI**-*Disgust* **NE**-*Neutral*

		PREDICTED							
	Emotion	HA	SA	AN	FE	SU	DI	NE	TOTAL
ACTUAL	HA	2	0	0	0	2	0	1	5
	SA	1	2	0	0	0	0	2	5
	AN	1	0	4	0	0	0	0	5
	FE	0	0	0	5	0	0	0	5
	SU	0	0	0	0	5	0	0	5
	DI	1	2	2	0	0	0	0	5
	NE	0	2	0	0	0	2	1	5
TOTAL		5	6	6	5	7	2	5	35

Table 4. Summary of micro- and macro-level evaluation from the initial 35 test data set
Legend: **TP**-*True Positive* **TN**-*True Negative* **FP**-*False Positive* **FN**-*False Negative*

Emotion	TP	TN	FP	FN	Recall	F Score	Precision	Accuracy	Error Rate
Happy	2	27	3	3	0.4000	0.4000	0.4000	0.8286	0.1714
Sad	2	26	4	3	0.4000	0.3636	0.3333	0.8000	0.2000
Angry	4	28	2	1	0.8000	0.7273	0.6667	0.9143	0.0857
Fear	5	30	0	0	1.0000	1.0000	1.0000	1.0000	0.0000
Surprise	5	28	2	0	1.0000	0.8333	0.7143	0.9429	0.0571
Disgust	0	28	2	5	0.0000	0.0000	0.0000	0.8000	0.2000
Neutral	1	27	3	4	0.2000	0.2222	0.2500	0.8000	0.2000
Micro-level Average:					**0.5429**	**0.5429**	**0.5429**	-	-
Macro-level Average:					**0.5429**	**0.5099**	**0.4806**	**0.8694**	**0.1306**

Classification and modelling requires ample amount of data to generate better results. In this study, 1750 labelled speech data from the website of Center for Empathic Human-Computer Interaction, De La Salle University (DLSU) were also used to test the accuracy of the study if ample amount of data is available. The speech data from DLSU were no longer subjected to feature extraction and normalisation since they have already been labelled. Another 175 utterances, subjected to the methods in this study including feature extraction and normalisation, were tested to generate a second confusion matrix with increased training data. Results are shown in Table 5 and Table 6. The samples consisted of 25 expected occurrences of each of the 7 emotions included in the study.

Table 5. Confusion matrix of 175 samples from the second run

Legend: **HA**-*Happy* **SA**-*Sad* **AN**-*Angry* **FE**-*Fear* **SU**-*Surprise* **DI**-*Disgust* **NE**-*Neutral*

		PREDICTED							
	Emotion	**HA**	**SA**	**AN**	**FE**	**SU**	**DI**	**NE**	**TOTAL**
ACTUAL	**HA**	**24**	1	0	0	0	0	0	**25**
	SA	0	**25**	0	0	0	0	0	**25**
	AN	0	0	**25**	0	0	0	0	**25**
	FE	0	0	0	**25**	0	0	0	**25**
	SU	0	0	0	0	**25**	0	0	**25**
	DI	0	5	0	0	0	**20**	0	**25**
	NE	1	0	0	0	0	0	**24**	**25**
	TOTAL	**25**	**31**	**25**	**25**	**25**	**20**	**24**	**175**

Table 6. Summary of micro- and macro-level evaluation from the 175 test data set

Emotion	TP	TN	FP	FN	Recall	F Score	Precision	Accuracy	Error Rate
Happy	24	149	1	1	0.9600	0.9600	0.9600	0.9886	0.0114
Sad	25	144	6	0	1.0000	0.8929	0.8065	0.9657	0.0343
Angry	25	150	0	0	1.0000	1.0000	1.0000	1.0000	0.0000
Fear	25	150	0	0	1.0000	1.0000	1.0000	1.0000	0.0000
Surprise	25	150	0	0	1.0000	1.0000	1.0000	1.0000	0.0000
Disgust	20	150	0	5	0.8000	0.8889	1.0000	0.9714	0.0286
Neutral	24	150	0	1	0.9600	0.9796	1.0000	0.9943	0.0057
Micro-level Average:					**0.9600**	**0.9600**	**0.9600**	-	-
Macro-level Average:					**0.9600**	**0.9633**	**0.9666**	**0.9886**	**0.0114**

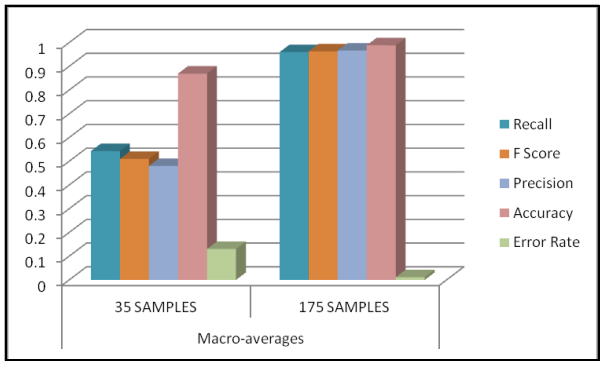

Fig. 4. Comparison of macro-level performance measures from the two test data sets

As the size of training data set increased, subjecting the second data set (175 samples) resulted to the improvement of all performance measures as shown in Figure 4. Precision and recall values increased showing that false detections and misdetections were both decreased.

5. CONCLUSION

The usage of Praat scripting language is effective in extracting features for the emotion modelling. The use of SVM-GMM resulted to macro-averages of 54.29% recall, 48.06% precision, 50.99% F Score, 86.94% accuracy and 13.06% error rate across the 7 selected emotions using 140 training data and 35 test data. All these measures (96% recall, 96.66% precision, 96.33% F Score, 98.86% accuracy and 1.14% error rate) improved when a larger training data of 1750 and test data of 175 were used. Thus, the objective of improving results using machine learning has been achieved. Proper account of the improvement was achieved by the use of appropriate performance measures for a multi-class classifier as what has been presented in this study.

The use of more advanced classification algorithm together with additional audio features may be explored. The use of speech databases may also be considered for further extension of this study.

6. ACKNOWLEDGMENTS

The authors would like to thank selected UPLB students for their participation in data gathering of speech utterances and the Center for Empathic Human Computer Interaction, De La Salle University for making their speech data accessible to public via their website as supplementary data for this study.

7. REFERENCES

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341-345.

Busso, C., Bulut, M., & Narayanan, S. S. (2013). Toward effective automatic recognition systems of emotion in speech. *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, 110-127.

Busso, C., Mariooryad, S., Metallinou, A., & Narayanan, S. (2013). Iterative feature normalization scheme for automatic emotion detection from speech. *IEEE transactions on Affective computing*, *4*(4), 386-397.

Cichosz, J., & Slot, K. (2007). Emotion recognition in speech signal using emotion-extracting binary decision trees. *Proceedings of Affective Computing and Intelligent Interaction*.

Clariño, M.A.A.D.C., Mariano, V.Y., & Albacea, E.A. (2012). Bi-level classification using naive bayes classifier and gaussian line-seeker method on rice leaf images. *Philippine Information Technology Journal*, *5*(2):19–24.

Cornelius, R. R. (1996). *The science of emotion: Research and tradition in the psychology of emotions*. Prentice-Hall, Inc.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech communication, 40*(1), 5-32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine,18*(1), 32-80.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572-587.

Epps, J., Chen, F., & Yin, B. (2010). Emotion Recognition and Cognitive Load Measurement from Speech. *APSIPA Annual Summit and Conference*. NICTA, Australia.

Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks, 18*(4), 389-405.

Gomes, J., & El-Sharkawy, M. (2015, December). i-vector Algorithm with Gaussian Mixture Model for Efficient Speech Emotion Recognition. In *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*(pp. 476-480). IEEE.

Iliou, T., & Anagnostopoulos, C. N. (2010). Classification on speech emotion recognition-a comparative study. *International Journal on Advances in Life Sciences Volume 2, Number 1 & 2, 2010, 4, 5*.

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing,13*(2), 293-303.

Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., & Steidl, S. (2010). Emotion recognition using imperfect speech recognition. *ISCA*.

Muthusamy, H., Polat, K., & Yaacob, S. (2015). Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. *Mathematical Problems in Engineering, 2015*.

Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. In *INTERSPEECH* (pp. 2679-2682).

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech communication, 41*(4), 603-623.

Roesch, E., Fontaine, J., & Scherer, K. (2006). The world of emotion is two-dimensional–or is it. *Presentation at the HUMAINE Summer school, Genoa, Italy*.

Shahzadi, A., Ahmadyfard, A., Harimi, A., & Yaghmaie, K. (2015). Speech emotion recognition using nonlinear dynamics features. *Turkish Journal of Electrical Engineering & Computer Sciences, 23*(Sup. 1), 2056-2073.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management,45*(4), 427-437.

Yacoub, S. M., Simske, S. J., Lin, X., & Burns, J. (2003). Recognition of emotions in interactive voice response systems. In *INTERSPEECH*.