Audio Features Analysis for Speech Emotion Recognition

Saeeda Begum[*], Sana Ul Haq[†], Muhammad Saeed Shah[‡], Muhammad Kamran[§], Imtiaz Rasool[**]

Abstract

In this paper audio features analysis is performed using two emotional speech databases: SAVEE in English language and EMO-DB in German language. A diverse set of more than 6000 acoustic features were extracted. The extracted features were normalized using z-score and min-max techniques, followed by feature selection using correlation, chi-square, gain ratio, and info gain ratio methods. Finally, classification was performed using various classifiers: support vector machine, Bayes net, meta and trees. The best classification result of 78.5% was achieved for seven emotion classes on Surrey audio-visual expressed emotion database using support vector machine classifier with 3500 features. The best result of 87.1% was obtained for the Berlin emotional speech database using support vector machine classifier with 4000 features. Classification performances comparable to human were obtained for both the databases. The Mel-spectrum, cepstral and spectral features were found most discriminative for audio emotion classification.

***Keywords*: S**peech emotion recognition; Feature selection; Info gain ratio; Chi square; Support Vector Machine.

Introduction

Affective computing plays an important role in making human-computer interaction more natural. In this field, extensive research has been performed on emotion recognition from speech. Research is in progress to develop a system that can recognize and respond to human emotional state by adjusting its behavior. Automatic emotion recognition and synthesis is performed from physiological signals, facial expressions, and speech (Picard, 2000). Affect recognition along with expressive speech synthesis plays a vital role in psychology, education,

[*]Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, saeedabegumjee123@gmail.com
[†]Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, sanaulhaq@uop.edu.pk
[‡]Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, saeedshah@uop.edu.pk
[§]Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, kamranmu@uop.edu.pk
[**] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, imtiazrasoolkhan@uop.edu.pk

entertainment, defense, medicines, and call centers (Burleson & Picard, 2004; Arias, Busso, & Yoma, 2014).

Speech emotion recognition has been performed through different machine learning algorithms. The first step is to acquire high-quality emotional speech data. A few public databases are available because most of the researchers use their personal sets of data which are not available to other researchers (Burleson & Picard, 2004; Arias, Busso, & Yoma, 2014). In recent years, researchers from various disciplines have been invited to solve the challenges in the field of emotion recognition and synthesis (Schuller B. S., 2010; Schuller, et al., 2020).

Researchers have used various speech databases for their analysis (Douglas-Cowie, Campbell, Cowie, & Roach, 2003; Ververidis & Kotropoulos, 2006). The emotional speech databases include Berlin (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), Danish (Engberg & Hansen, 1996), and TESS (Dupuis & Pichora-Fuller, 2011) emotional speech databases. The Cohn-Kanade (Kanade, Cohn, & Tian, 2000) and MMI (Pantic, Valstar, Rademaker, & Maat, 2005) are the visual databases. SAVEE (Haq & Jackson, 2011), RAVDESS (Livingstone & Russo, 2018) and AVIC (Hassan & Damper, 2012) are examples of audio-visual emotional databases.

The speech signal consists of two types of information: linguistic and paralinguistic. Paralinguistic features are also known as acoustic features which include prosodic, spectral and voice quality features. The prosodic features have been widely used for emotion recognition. These features consist of rhythm, loudness, pitch, duration, pause, intonation of speech, and speaking rate (Ververidis & Kotropoulos, 2006; Schuller, Vlasenko, Eyben, Rigoll, & Wendemuth, 2009). Frequently used pitch features include fundamental frequency (f_0) and glottal air velocity. Different statistical methods like median, mean, maximum, minimum, standard deviation and 3rd or 5th moment are applied to the contour for analysis. Loudness is the energy of sound produced and is directly related to intensity of emotions. The famous feature is found to be speaking rate. The intensity, duration, and pitch of basic emotions were examined by (Ververidis & Kotropoulos, 2006). Anger was found to be the one with highest intensity, high energy, and pitch level, when the relation between emotions and features were investigated.

The fundamental frequency in the form of harmonics produces the spectral features. The air flow in the vocal track is nonlinear which produces harmonics of various amplitudes and frequencies. The Mel

Frequency Cepstral Coefficients (MFCC) between 20 Hz to 300 Hz band is used to model pitch, which has performed better as compared to pitch features. Alternatives to MFCC features are Mel Filter Bank (MFB), Linear Predictive Cepstral Coefficients (LPCC), and Relative Spectral Transform-Perceptual Linear Perdition (RAST-PLP) features (Douglas-Cowie, et al., 2007; Neiberg, Elenius, & Laskowski, 2006; Steidl, Batliner, Noth, & Hornegger, 2008).

The voice quality features include jitter, shimmer, and harmonics to noise ratio (HNR). Several researchers improved their results by merging acoustic features with linguistic features (Ververidis & Kotropoulos, 2006; Schuller, Villar, Rigoll, & Lang , 2005; Scherer, 2000). The brute force approach is the advanced method of extracting a large set of features from speech. Different sets of features were used by researchers to develop a reliable speech emotion recognition system. Batliner, et al. (2004) improved their results by introducing more than 6000 features from individual sets. An open-source toolkit known as openSMILE was introduced by Eyben et al. (2009) for the extraction of audio features. Similarly, a large set of acoustic features was extracted by Hassan and Damper (2012) for emotion recognition.

To improve the emotion recognition system performance, unnecessary and redundant features should be removed from the extracted features. The filter and wrapper feature selection methods are used for this purpose. In filter method features are ranked according to their ability of separation between classes based on some criterion (Aharon, Elad, & Bruckstein, 2006). The examples of filter methods are chi-square, gain ratio and info gain ratio. The wrapper method evaluates features based on a prediction model for a specific classification task. The wrapper method is more effective and predominately provides better results. The wrapper method is computationally costlier as compared to filter method, but it provides better performance as compared to filter method (Kuhn & Johnson, 2013; Morgan, 2014; Kohavi & John, 1997). Hassan and Damper (2012) used k-nearest neighbor classifier and achieved 67% accuracy for the Danish database and 90% accuracy for the Berlin database by selecting 1052 features using Gaussian mixture model (GMM) classifier. Clavel et al. (2008) obtained 71% accuracy using the best 40 features selected from 1052 audio features.

The final step is the classification of different emotion classes. The frequently used classifiers include Bayesian network (BN), support vector machine (SVM), hidden Markov model (HMM), GMM, neural network (NN) and adaptive boosting (AdaBoost). Berlett et al. (2005)

classified seven emotions using SVM and obtained an average accuracy of 89%. Araño et al. (2021) utilized a hybrid set of features for classifying emotions from speech consisting of MFCCs and image features extracted from spectrograms. The MFCCs features along with long short-term memory (LSTM) network performed better as compared to SVM classifier. Mannepalli et al. (2022) introduced multiples support vector neural network classifier for speech emotion recognition. The proposed model performed better as compared to AFDBN, FDBN, and DBN models. Alluhaidan et al. (2023) combined the MFCCs and time-domain features (MFCCT) to achieve better classification accuracy. The convolutional neural network was used for classification, which performed better as compared to other machine learning classifiers.

Mohan et al. (Mohan, Dhanalakshmi, & Kumar , 2023) proposed 2D Convolutional Neural Network (2D-CNN) with eXtreme Grading Boosting (XG-Boost) for audio emotion classification. A classification accuracy of 96.5% was obtained for 16 emotions of RAVDESS dataset using MFCC features. The proposed ensemble model outperformed the Random Forest and CNN-LSTM. Bhanusree et al. (Bhanusree, Kumar, & Rao, 2023) used a time-distributed attention-layered CNN for extraction of features and Random Forest for classification. The proposed model obtained classification accuracies of 92.2% and 90.3% on the RAVDESS and IEMOCAP datasets, respectively. Novais et al. (Novais, Cardoso, & Rodrigues, 2022) used Random Forest, AdaBoost, Neural Network and their ensemble using majority vote for audio emotion classification. The classification accuracy of 75.6% was achieved on the RAVDESS dataset using Random Forest and 86.4% on a group of datasets consisting of RAVDESS, SAVEE, and TESS using Neural Network. The individual classifiers outperformed the ensemble learning with majority vote. Chalapathi et al. (Chalapathi, Kumar, Sharma, & Shitharth, 2022) proposed AdaBoost classifier with high-dimensional acoustic features for speech emotion recognition. A classification accuracy of 94.8% was achieved for 7 emotion classes of the RAVDESS dataset.

Methodology

The audio emotion recognition is performed in the following steps: emotional speech data, feature extraction, feature normalization, feature selection and classifications.

Emotional Speech Database

The effectiveness of an emotion recognition system mainly depends upon the data used for modeling the system. In this research, two famous emotional speech databases, i.e., SAVEE and EMO-DB, have been used for the analysis.

Surrey Audio-Visual Expressed Emotion (SAVEE) Database

This database is free of charge and publicly available for research (Haq & Jackson, 2011). It contains data of four British male actors. The recordings are in six distinct emotions, i.e., anger, happiness, disgust, fear, sadness, and surprise, plus neutral. The database is composed of 120 sentences per actor, which were selected from the TIMIT corpus (Fisher, 1986). These 120 sentences consist of 15 sentences for each emotion ($15 \times 6 = 90$ sentences) with an addition of 30 neutral sentences. The dataset consists of 480 sentences in total. The recorded data were evaluated by 20 subjects after being processed and labeled at CVSSP, University of Surrey, UK. The 20 evaluators included 10 male and 10 female speakers. The human accuracy for seven emotions is 66.5% for the audio data.

Berlin Emotional Speech Database (EMO-DB)

The EMO-DB was recorded in the German language at the Technical University Berlin (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). The database contains data of five male and five female German actors in six emotions, i.e., anger, boredom, happiness, fear, disgust, and sadness, plus neutral. The data were evaluated by 20 listeners, and those sentences which were more than 60% natural and having a recognition rate of 80% or above were selected. The database has 535 utterances. The database contains 10 sentences. Average human accuracy for the database is 86.0% for seven emotions.

Feature Extraction

The openSMILE toolkit (Eyben, Wollmer, & Schuller, 2009) was used to extract 6669 audio features. These features included signal energy, Mel spectrum, cepstral, pitch, spectral, raw signal, and voice quality related low-level descriptors (LLDs), their delta (Δ) and delta-delta (ΔΔ) coefficients. The details of these features are given in Table 1. A total of 39 statistical functions were applied to these features to obtain 6669 features for each speech signal.

$$(57\ LLDs + 57\ \Delta + 57\ \Delta\Delta) \times 39\ functions = 6669\ features$$

Feature Normalization

Feature normalization, also known as data scaling, is important for any classification problem. The ranges of values are different for various types of features, and therefore it is necessary to scale them to the same range. We used two types of normalization: z-score (Jain, Nandakumar, & Ross, 2005) and min-max (Jain & Bhandare, 2011). The z-score normalizes features to a zero mean and unity variance, while min-max method normalizes data to range [min max]. In our case we normalize features to the range [0 1].

Table 1
Details of low-level descriptors (LLDs)

Features group	**Details of features**	**No. of features**	**Δ coefficients**	**Δ Δ coefficients**
Signal energy	Log energy per frame	1	1	1
Mel- spectrum	Energy in Mel-frequency bands (0-25)	26	26	26
Cepstral	Mel-frequency cepstral coefficients (0-12)	13	13	13
Pitch	Pitch (f_0) in Hz and its contour	2	2	2
Spectral	Energy in frequency bands 0-250 Hz, 0-650 Hz, 250-650 Hz, 1-4 kHz, and 3-9 kHz	5	5	5
	Flux, centroid, position of spectral max. and min. peaks	4	4	4
	Spectral roll off points 25%, 50%, 75% and 90%	4	4	4
Raw signal	Zeros crossing rate (ZCR)	1	1	1
Voice quality	Probability of voicing	1	1	1

Feature Selection

Features were selected using correlation-based feature selection (CFS), chi-square, gain ratio, and info gain ratio techniques. The Weka software was used for this purpose. The selected attributes are given in Table 2. The correlation-based feature selection method selects the most appropriate feature set for classification. On the other hand, chi-square, gain ratio, and info gain ratio are ranked methods which rank individual features based on their suitability for classification. The CFS method selected a set of 103 features for the SAVEE database, and a set of 232 features for the EMO-DB. The chi-square, gain ratio, and info gain ratio ranked the individual features for both the databases.

Classification

To achieve better classification performance, we used different types of classifiers: SVM, Bayes net, meta and trees. The Weka toolkit was used for classification.

Table 2
Selected attributes for the SAVEE-DB and EMO-DB

Attributes Evaluator	No. of selected attributes for SAVEE-DB	No. of selected attributes for EMO-DB
CFS	103	232
Chi-square	6604	6365
Gain Ratio	6604	6365
Info Gain Ratio	6604	6365

Results and Discussion

The experiments were performed using both z-score and min-max normalization. In the case of CFS, the subset of selected features was used. In the case of rank methods, i.e., chi-square, gain ratio, and info gain ratio, a step size of 50 was used up to 500 features and a step size of 100 was used from that point onwards. Different types of classifiers with 10-fold cross validation were used for classification.

The classification results for seven emotions of the SAVEE database is given in Table 3. The z-score normalized features provided slightly better results in comparison to min-max normalized features. The rank methods performed better as compared to CFS. The SVM classifier performed better in comparison to other classifiers. The best classification score of 78.54% was obtained using SVM classifier with 3500 features. The features were normalized by z-score and selected with chi-square and info gain ratio. The set of the best 3500 features is given in Table 4. The Mel-spectrum, cepstral and spectral features contributed mainly to classification. In addition, features from other groups also contributed.

The average classification accuracy for seven emotions of the EMO-DB is given in Table 5. The results were comparable for the two normalization techniques. The classification accuracies for all feature selection methods were quite close to each other. The performance of SVM classifier was better in comparison to other classifiers. The best performance of 87.1% was obtained using SVM classifier with 4000 features. These features were normalized with min-max and selected with gain ratio. The set of best 4000 features is given in Table 6. The Mel-spectrum, cepstral and spectral features were observed to be the most important, while features from other groups also contributed.

Table 3
Classification results for seven emotions on SAVEE-DB

Feature normalization	Attribute selector	Classifier	No. of features	Classification accuracy (%)
Z-score	Chi-squared	SVM	3500	**78.54**
	Gain Ratio	SVM	4000	77.3
	Info Gain Ratio	SVM	3500	**78.54**
	CFS	BayesNet	103	70.8
Min-Max	Chi-square	SVM	4100	78.33
	Gain Ratio	SVM	5200	78.33
	Info Gain Ratio	SVM	6000	77.96
	CFS	SVM	103	70.4

Table 4
Best set of selected features for the SAVEE-DB

Features group	Features in the group	Number of features
Mel-spectrum	Energy in Mel-frequency bands (0-25)	2177
Signal Energy	Logarithmic	75
Cepstral	Mel-frequency cepstral coefficients (0-12)	293
Pitch	Pitch (f_0) in Hz and its contour	108
Spectral	Energy in frequency bands: 0-250 Hz, 0-650 Hz, 250-650 Hz, 1-4 kHz, and 3-9 kHz	397
	Flux, centroid, position of spectral max. and min. peaks	212
	Spectral roll off points 25, 50, 75 and 90%	126
Raw signal	Zero crossing rate (ZCR)	84
Voice quality	Probability of voicing	28
Total		**3500**

Table 5
Classification results for seven emotions on EMO-DB

Feature normalization	Attribute selector	Classifier	Number of features	Classification accuracy (%)
Z-score	Chi-squared	SVM	6000	86.95
	Gain Ratio	SVM	6000	86.92
	Info Gain Ratio	SVM	6000	86.92
	CFS	SVM	232	86.16
Min-Max	Chi-square	SVM	5000	86.72
	Gain Ratio	SVM	4000	87.1
	Info Gain Ratio	SVM	6000	86.92
	CFS	Bayes Net	232	85.98

For both the SAVEE and EMO-DB databases, it is observed that the Mel-spectrum, cepstral and spectral features are crucial for emotion classification. Furthermore, the signal energy, pitch, raw signal, and voice quality features also contributed to classification.

Table 6
Best set of selected features for the EMO-DB

Features group	**Features in the group**	**Number of features**
Mel-spectrum	Energy in Mel-frequency bands (0-25)	2021
Signal Energy	Logarithmic	80
Cepstral	Mel-frequency cepstral coefficients (0-12)	887
Pitch	Pitch (f_0) in Hz and its contour	118
Spectral	Energy in frequency bands: 0-250 Hz, 0-650 Hz, 250-650 Hz, 1-4 kHz, and 3-9 kHz	338
	Flux, centroid, position of spectral max. and min. peaks	237
	Spectral roll off points 25%, 50%, 75% and 90%	192
Raw signal	Zero crossing rate (ZCR)	50
Voice quality	Probability of voicing	77
Total		**4000**

Conclusion

Audio feature analysis is performed on English (SAVEE) and German (EMO-DB) emotional speech databases. A large set of audio features was extracted. The extracted features were normalized using z-score and min-max techniques. Feature selection was performed using CFS, chi-square, gain ratio, and info gain ratio methods, followed by classification using various techniques.

In the case of SAVEE database, the best result of 78.5% was obtained using 3500 features with SVM classifier as compared to 66.5% by humans for seven emotions. For the EMO-DB, the best accuracy of 87.1% was achieved using 4000 features with SVM classifier as compared to 86.0% by humans for seven emotion classes.

For the SAVEE database, the ranker methods of feature selection, i.e., chi-square, gain ratio, and info gain ratio performed better as compared to CFS, while in the case of EMO-DB the results of ranker methods and CFS were comparable. The SVM classifier performed better in comparison to other classification techniques. For both databases, the Mel-spectrum, cepstral and spectral features were found to be most discriminative for emotion classification. In addition, other audio features including signal energy, pitch, raw signal, and voice quality also contributed.

In future, various combining classification techniques will be investigated to achieve better classification results. In these techniques, different sets of features will be used along with different classifiers and their results will be combined. The other direction of research is the hierarchical approach of emotion classification. In this approach,

different sets of features will be used at different levels to achieve better performance.

References

Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing, 54*(11), 4311-4322.

Arias, J. P., Busso, C., & Yoma, N. B. (2014). Shape- Based Modeling of the Fundamental Frequency Contour for Emotion Detection in Speech. *Computer Speech & Language, 28*(1), 278-294.

Bhanusree, Y., Kumar, S. S., & Rao, A. K. (2023). Time-Distributed Attention-Layered Convolution Neural Network with Ensemble Learning using Random Forest Classifier for Speech Emotion Recognition. *Journal of Information and Communication Technology, 22*(1), 49-76.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A Database of German Emotional Speech. *Interspeech*, (pp. 1517-1520). Lisbon.

Burleson, W., & Picard, R. W. (2004). Affective Agents: Sustaining Motivation to Learn through Failure and a State of Stuck. *Workshop on Social and Emotional Intelligence in Learning Environments.* Maceio.

Chalapathi, M. V., Kumar, M. R., Sharma, N., & Shitharth, S. (2022). Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal. *Security and Communication Networks, 2022*, 1-10.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40*(1-2), 33-60.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., et al. (2007). The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. *Affective Computing and Intelligent Interaction* (pp. 488-500). Lisbon: Springer.

Dupuis, K., & Pichora-Fuller, M. K. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics, 39*(3), 182-183.

Engberg, I. S., & Hansen, A. V. (1996). *Documentation of the Danish emotional speech database DES.* Denmark: Internal AAU report, Center for Person Kommunikation.

Eyben, F., Wollmer, M., & Schuller, B. (2009). OpenEAR-Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1-6). IEEE.

Fisher, W. (1986). The DARPA Speech Recognition Research Database: Specifications and Status. *DARPA Workshop on Speech Recognition,* (pp. 93-99). Palo Alto.

Haq, S., & Jackson, P. J. (2011). Multimodal Emotion Recognition. In *Machine Audition: Principles, Algorithms and Systems* (pp. 398-423). IGI Global.

Hassan, A., & Damper, R. I. (2012). Classification of emotional speech using 3DEC hierarchical classifier. *Speech Communication, 54*(7), 903-916.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition, 38*(12), 2270-2285.

Jain, Y. K., & Bhandare, S. K. (2011). Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology, 2*(8), 45-50.

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive Database for Facial Expression Analysis. *International Conference on Face and Gesture Recognition* (pp. 46-53). Grenoble: IEEE.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1), 273–324.

Kuhn, M., & Johnson, K. (2013). Filter methods. In *Applied Predictive Modeling* (pp. 499–500). New York: Springer.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one, 13*(5), e0196391.

Mohan, M., Dhanalakshmi, P., & Kumar , R. S. (2023). Speech Emotion Classification Using Ensemble Models with MFCC. *Procedia Computer Science, 218*, 1857-1868.

Morgan, J. (2014). Classification and Regression Tree Analysis. *Boston: Boston University, 298*.

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion Recognition in Spontaneous Speech Using GMMs. *International Conerence on Spoken Language Processing*, (pp. 809-812).

Novais, R. M., Cardoso, P. J., & Rodrigues, J. M. (2022). Emotion Classification from Speech by an Ensemble Strategy. *International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, (pp. 85-90). Lisbon.

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-Based Database for Facial Expression Analysis. *International Conference on Multimedia and Expo* (pp. 317-321). Amsterdam: ACM.

Picard, R. W. (2000). *Affective Computing.* Cambridge: The MIT Press.

Scherer, K. R. (2000). A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. *Interspeech*, (pp. 379-389).

Schuller, B. S. (2010). The INTERSPEECH 2010 Paralinguistic Challenge. *Interspeech.* Makuhari.

Schuller, B. W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., et al. (2020). The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. *Interspeech.*

Schuller, B., Villar, R. J., Rigoll, G., & Lang , M. (2005). Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition. *International Conference on Acoustics, Speech, and Signal Processing* (pp. 325-328). IEEE.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009). Acoustic emotion recognition: A benchmark comparison of performances. *Workshop on Automatic Speech Recognition and Understanding* (pp. 552–557). Merano: IEEE.

Steidl, S., Batliner, A., Noth, E., & Hornegger, J. (2008). Quantification of Segmentation and F0 Errors and Their Effect on Emotion Recognition. *International Conference on Text, Speech and Dialogue*, (pp. 525-534). Berlin.

Ververidis, D., & Kotropoulos, C. (2006). Emotional Speech Recognition: Resources, Features, and Methods. *Speech Communication, 48*(9), 1162-1181.