Sindh Univ. Res. Jour. (Sci. Ser.) Vol.47 (1): 67-72 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Audio-Visual Emotion Classification using Filter and Wrapper Feature Selection Approaches

S. HAQ, M. ASIF, A. ALI*, T. JAN++, NaveedAHMAD**, Y. KHAN***

Department of Electronics, University of Peshawar, Pakistan

Received 8th June 2014 and Revised 3rd January 2015

Abstract: A comparative analysis of filter and wrapper approaches of feature selection has been presented for the audio, visual and audio-visual human emotion recognition. A large set of audio and visual features were extracted, followed by speaker normalization. In filter approach, feature selection was performed using the Plus l-Take Away r algorithm based on Bhattacharyya distance criterion. In wrapper method, features were selected based on their classification performance using support vector machine (SVM) classifier. Finally, an SVM classifier was used for human emotion recognition. The filter approach provided a slight better performance in comparison to wrapper approach for seven emotions on SAVEE database.

Keywords: Emotion Recognition, Feature Selection, Filter Approach, Wrapper Technique, Feature Normalization, Classification

1. INTRODUCTION

In recent years, humans have developed advanced technologies which have made the human-computer interaction more attractive for humans. Primarily, human beings use speech to communicate with each other, but the complete message cannot be conveyed only through verbal content. Some additional information are required to completely understand the message, these includes vocalized emotions, facial expressions, hand gestures and body language (Sebe *et al.,* 2005; Kessous *et al.,* 2010). From human's point of view, the human-machine interaction will be more interesting if machines can recognize human emotions and are able to respond accordingly (Picard, 1997). On the other hand, communication will be more reliable if machines are able to understand human emotions.

Automatic emotion recognition covers a wide range of applications including automobile systems, customer services systems, health care, and game and film industries (Zeng *et al.,* 2009). In recent years, researchers from different disciplines have taken interest in the field of human emotion recognition and significant progress has been made in several areas including recording of emotional databases, feature extraction, feature selection, and classification (Zeng *et al.,* 2009; Haq and Jackson, 2010).

Previous studies have mainly focused on unimodal approaches (e.g., speech, facial expressions) for emotion classification. The modalities have been analyzed independently and the interrelation between different modalities has not been explored. In actual fact, speech and facial expressions are highly correlated, and emotions and linguistic content influence the relationship between these two modalities (Busso and Narayanan, 2007). Humans utilize both speech and gesture to express their emotions, and for this reason an ideal emotion recognizer should be based on both verbal and non-verbal information (Sebe *et al.,* 2005).

One of the factors upon which the reliability of an emotion recognizer depends is the quality of emotional data used to build the emotion recognition system. Examples of audio databases are the AIBO corpus and Berlin database (Batliner *et al.,* 2004; Burkhardt *et al.,* 2005). Visual databases include the Cohn-Kanade and MMI databases (Kanade *et al.,* 2000; Pantic *et al.,* 2005). The audio-visual databases are Facial Motion Capture database, GEMEP, and Belfast Naturalistic database (Busso and Narayanan, 2007; Bänziger *et al.,* 2006; Douglas-Cowie *et al.,* 2003). The emotional databases are either acted or natural.

Audio and visual features of various types have been analyzed to achieve better emotion classification results. Important audio features are pitch, formants, duration, spectral energy, and Mel frequency cepstral coefficients (MFCCs). These features have been used both at utterance-level (Haq and Jackson, 2009) and frame-level (Lin and Wei, 2005). The vision-based techniques are mainly based on facial expressions, since face plays a vital role in conveying emotions. Facial features can be subdivided into two broad categories: geometric and appearance (Pantic and Bartlett, 2007). The technique (Pantic and Bartlett, 2007) is based on geometric features, while (Bartlett *et al.,* 2006) used appearance features.

++ Corresponding Author: Tariqullah Jan, tariqullahjan@uetpeshawar.edu.pk, +92-91-9216498,
* Department of Electrical Engineering, Sarhad University of Science and IT, Peshawar.
** Department of Computer Science, University of Peshawar.
*** Department of Electrical Engineering, IQRA National University, Peshawar.

In the field of pattern recognition, feature selection techniques are used to remove the uninformative, redundant and noisy information. These techniques improve the classification accuracy and computational efficiency. The feature search techniques used in the field of emotion recognition include sequential floating forward selection (Haq *et al.,* 2008), genetic algorithms (Casale *et al.,* 2007) and best-first (Gunes and Piccardi, 2005).

In any pattern recognition problem, the choice of classifier is very important. These classifiers include hidden Markov model (Lin and Wei, 2005), neural network (Petrushin, 1999), support vector machine (Ashraf *et al.,* 2007), and adaptive boosting (Bartlett *et al.,* 2005). Multimodal approaches have been adopted to improve the emotion classification by fusion of data at feature, decision, and model levels.

In the field of emotion recognition, different audio and visual features have been used for audio-visual emotion classification. We have focused on filter and wrapper feature selection techniques to achieve better classification performance for the audio, visual and audio-visual human emotion classification in the speaker-independent scenario. The following sections present the SAVEE database, method, experimental results and conclusion.

2. SAVEE DATABASE

We used Surrey Audio-Visual Expressed Emotion (SAVEE) database (Jackson and Haq, 2015) for our analysis. The database consists of data from four British male speakers in Ekman's six basic emotions (anger, fear, disgust, sadness, happiness and surprise) plus neutral (Ekman *et al.,* 1987). The text material consisted of 15 phonetically-diverse sentences for each of the six emotions and 30 sentences for the neutral. The distribution of sentences in this way resulted in 120 utterances per actor and 480 utterances in total.

The sampling rate for audio data was 44.1 kHz, while that for video was 60 fps. To extract facial features, each actor's frontal face was painted with 60 markers. The SAVEE database was evaluated at utterance level by 20 subjects (10 male, 10 female). Each actor's data were evaluated by 10 subjects. The classification accuracy for seven emotion classes averaged over four actor's data and 10 evaluators was 66.5 % for audio, 88.0% for visual, and 91.8 % for audio-visual data.

3. MATERIAL AND METHODS

Proposed Method

The speaker-independent emotion classification was performed using a four steps method: feature extraction, feature normalization, feature selection, and classification and fusion.

3.1 Feature Extraction

For emotion classification in the speaker-independent scenario, features were extracted at utterance-level consisting of 830 audio and 290 visual features. The details of the audio and visual features are given below.

Audio Features

A total of 830 audio features were extracted, consisting of 38 intensity, 38 loudness, 35 pitch, 38 probability of voicing, 304 line spectral frequencies, 35 zero-crossing rate, 10 duration, 60 energy and 272 spectral envelope features. The audio features were extracted at utterance-level using the open EAR toolkit (Eyben *et al.,* 2009), Speech Filing System (Huckvale, 2015), and HTK (Young *et al.,* 2015).

Audio features: The low-level descriptors (LLD) including pitch (f_0), intensity, loudness, probability of voicing, 8 line spectral frequencies and zero-crossing rate (ZCR) were extracted using the openEAR toolkit. The delta regression coefficients were computed these LLD, and the following functions were applied to the LLD and their delta coefficients: maximum/minimum value and respective relative position, range, arithmetic mean, 2 linear regression coefficients, linear and quadratic errors, standard deviation, skewness, kurtosis, 3 quartiles (25%, 50%, 75% percentile), and inter-quartile ranges.

Duration features: Phone labels were used to extract duration features. The phone labeling was performed in two steps: first HTK was used for automatic labeling of the audio, and second the Speech Filing System was used to correct the automatic phone labels. The extracted were voiced speech duration, unvoiced speech duration, sentence duration, average voiced phone duration, average unvoiced phone duration, voiced-to-unvoiced speech duration ratio, average voiced-to-unvoiced phone duration ratio, speech rate, voiced-speech-to-sentence duration ratio, and unvoiced-speech-to-sentence duration ratio.

Energy features: The extracted features were mean and standard deviation of total log energy; minimum, maximum, range, mean and standard deviation of normalized energies in the original speech signal and speech signal in the frequency bands 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz, 2-4 kHz and 4-8 kHz; minimum, maximum, range, mean and standard deviation of first order difference of normalized energies in the original speech signal and speech signal in the same frequency bands.

Spectral features: The HTK was used to extract the spectral envelope features. The feature vector consisted of 12 MFCCs, their delta, and delta delta coefficients. The following functions were applied to the feature vector: maximum/minimum value, range, arithmetic mean, standard deviation, skewness and kurtosis.

Visual Features

The visual features were extracted by painting markers on actors' foreheads, eyebrows, cheeks, lips and jaws. The markers were automatically tracked and normalized relative to a reference point at the bridge of the nose (Busso and Narayanan, 2007). For the speaker-independent emotion classification, we extracted the marker position (MP) and marker angle (MA) features. A total of 290 visual features were extracted from the adjusted marker coordinates after translation, rotation and mapping of markers.

Marker position (MP) features: A total of 236 visual features were obtained from the 2D marker coordinates as mean and standard deviation of the adjusted marker coordinates.

Marker angle (MA) features: A total of 54 marker angle features were extracted related to mouth height, width, and roundness; left and right cheek height, width, roundness, and angles; nose upper half, lower half, and total length, nose angle and area; forehead angle and areas; left eyebrow outer and inner angles, and right eyebrow outer and inner angles; chin upper and lower angles. The roundness feature was calculated as a ratio of height to width.

3.2 Feature Normalization

Feature normalization plays an important role in pattern recognition by restricting the variation of different types of features to the same range. Audio features of different types have different value ranges and normalization is required for the equal weighting of those features. In the case of visual features, due to head pose, difference in face structure of the speakers and variation in distance from the camera, some kind of translation, rotation and scaling is required to process the data.

Audio Feature Normalization

Audio emotion classification was performed using speaker normalization. Normalization of features corresponding to a specific speaker was performed using the mean and standard deviation of that speaker's data. The entire speaker's data were used for normalization. The actual emotional state of any observed utterance was unknown (Vlasenko *et al.*, 2007b). Speaker normalization is given by

$$x_{sn} = \frac{x - \mu_s}{\sigma_s} \qquad (1)$$

where x_{sn} is the speaker-normalised feature, x is the original feature, μ_s is the mean and σ_s is the standard deviation of feature calculated from speaker's data.

Visual Feature Normalization

Visual feature normalization was performed in two steps: firstly, the facial markers were translated and rotated to align frames (Busso and Narayanan, 2007), and in the next step, radial basis functions were used to map the marker data of each speaker onto a single speaker (Lorenzo *et al.*, 2003).

3.3 Feature Selection

We used two methods for feature selection: Filter approach and wrapper approach.

Filter Approach

In filter method the features selection is independent of classification technique used for recognition. Features are selected based on some criterion.

The Plus l-Take Away r algorithm is a feature search method based on some criterion function (Kittler, 1978). It combines the sequential forward selection (SFS) and sequential backward selection (SBS) algorithms to achieve better results. At each step, l numbers of features are included to the current feature set and r numbers of features are discarded. The process continues until the required feature set size is achieved. The feature search was performed with $l = 2$ and $r = 1$, i.e., one feature was added at each step. We used this algorithm for feature selection based on Bhattacharyya distance criterion (Campbell, 1997). Feature selection was performed over the full range of audio, visual and audio-visual feature sets.

The Bhattacharyya distance provides the separability between two classes. The distribution of classes was assumed to be normal. For m number of classes the separability measure is obtained by averaging it over all binary combinations of the classes.

Wrapper Approach

In wrapper method, features are selected based on their classification performance using a specific classifier. We used support vector machine (SVM) classifier for the wrapper approach.

3.4 Classification and Fusion of Modalities

The SVMs are one of the popular classifiers that have been used by many researchers for emotion recognition, and high performance is often achieved (Schuller *et al.*, 2007c; Vlasenko *et al.*, 2007a). The SVMs were originally developed for binary classification, and were later extended to multi-class problems. Multi-class approaches can be divided into two main categories: combining several binary

classifiers, and considering all classes in a single optimization function (Bishop, 2006). The approaches based on binary classification include one-versus-one, one-versus-rest, and directed acyclic graph (DAG). Weston and Watkins (1999) defined a single objective function which considers all the classes simultaneously. Considering input data that are two sets of vectors in n-dimensional space, the SVM will construct a separating hyperplane in that space. When data points are separated by a non-linear region, it is difficult to separate them by simply constructing an n-dimensional hyperplane. The SVM handles this problem by using a kernel function to map the data onto a high dimensional space where a hyperplane can easily separate the data. Common kernel functions of the SVM are linear, polynomial, radial basis function (RBF) and sigmoid. We used polynomial and radial basis function kernels for our analysis.

The audio-visual emotion classification was performed by fusion of modalities at decision-level. The probabilities obtained from the audio and visual modalities were multiplied with equal weighting to obtain the final result.

4. RESULTS AND DISCUSSION

For the wrapper approach features were selected based on the classification performance of an SVM classifier. Feature selection was performed using 1/3 of the data from each speaker with leave-one-speaker-out cross-validation. The SVM classifier was trained on the full data from three speakers and tested on the unseen 2/3 of the data from the fourth speaker. Features were selected for each of the audio and visual modalities, and the audio-visual emotion classification was based on decision-level fusion. In this way, four sets of experiments were performed and the results were averaged. The experiments were performed on SAVEE database for seven emotion classes. The audio feature set consisted of 250 features (speaker-normalized) selected based on the Bhattacharyya distance, and the visual features consisted of two sets: 236 MP features, and 290 MP + MA features.

The results for seven emotion classes are given in Table 1. The SVM classifier performed better for the visual features compared to the audio, and an average classification accuracy of above 85% was achieved for the bimodal scenario on the development set. For the evaluation set, average classification accuracy close to 80% was achieved for the audio-visual features. The classification performance for the visual modality improved when MA features were combined with MP features, and resulted in an increase in the overall performance for the bimodal scenario. The results indicate that MA features are helpful for emotion classification.

For comparison purposes, we repeated the experiments using the best set of features selected with the filter-based approach using Bhattacharyya distance. The results are given in Table 2. For the evaluation set, the average classification accuracies of 61% and 58% were achieved for the audio, 56% and 66% for the visual, and 73% and 83% for the audio-visual with the RBF and polynomial kernels, respectively. In comparison, the best classification accuracies using the wrapper-based approach for the audio were 54% and 56%, for the visual were 68% and 58%, and for the audio-visual were 79% and 81% with the RBF and polynomial kernels, respectively. The filter approach provided a better performance for the audio and audio-visual modalities. On the other hand, the wrapper approach performed better in the case of visual modality. To conclude, the filter-based approach provided a slight better performance in comparison to wrapper-based approach.

The humans' classification accuracy for seven emotions on the SAVEE database was 67% for audio, 88% for visual, and 92% for audio-visual data. Differences existed between the classification accuracies of machine and humans. The possible reasons are differences in the training data, i.e., the machine was trained/tested in a speaker-independent scenario but humans were adapted to a small amount of data, the task was discrete emotion classification, and the expressed emotions may have been lacking in naturalness.

Table 1: Average classification accuracy (%) for seven emotions on SAVEE database using the SVM classifier (radial basis function (RBF) and polynomial (Pol.) kernels) with wrapper-based feature selection. The audio was speaker-normalized and the visual features consisted of MP + MA features. The mean is averaged over 4 leave-one-speaker-out cross-validations.

Feature Set	Data Set	Audio		Visual		Audio-Visual	
		RBF	Pol.	RBF	Pol.	RBF	Pol.
Audio: 250; Visual: MP	Dev.	66.9 ± 5.1	86.3 ± 7.1	81.3 ± 12.4	90.0 ± 6.0	76.2 ± 13.3	86.9 ± 5.4
	Eval.	53.8 ± 15.9	55.6 ± 16.7	62.2 ± 18.0	66.3 ± 7.6	78.1 ± 7.8	77.8 ± 9.1
Audio: 250; Visual: MP+MA	Dev.	66.9 ± 5.1	86.3 ± 7.1	95.0 ± 5.3	94.4 ± 4.2	86.2 ± 10.9	89.4 ± 8.3
	Eval.	53.8 ± 15.9	55.6 ± 16.7	68.1 ± 16.1	57.5 ± 16.2	79.4 ± 5.1	80.9 ± 9.1

Table 2: Comparison between average classification accuracy (%) of filter-based and wrapper-based feature selection approaches for seven emotions on SAVEE database. The mean is averaged over 4 leave-one-speaker-out cross-validations.

Feature Selection	Audio		Visual		Audio-Visual	
	RBF	Pol.	RBF	Pol.	RBF	Pol.
Filter method	61	58	56	66	73	83
Wrapper method	54	56	68	58	79	81

5. CONCLUSION AND FUTURE WORK

For the speaker-independent task using the SAVEE database, the filter and wrapper approaches of feature selection were compared. A large set of audio and visual features were extracted, followed by feature normalization. In the filter approach, features were selected using the Plus l-Take Away r algorithm based on Bhattacharyya distance criterion. In the wrapper approach, features were selected based on their classification performance using an SVM classifier. Finally, an SVM classifier was used to classify between seven emotions on SAVEE database using the audio, visual and audio-visual modalities.

For the filter approach, the best classification accuracies of 61%, 66% and 83% were achieved for the audio, visual and audio-visual modalities, respectively. In the case of wrapper approach, the best classification results of 56%, 68% and 81% were achieved for the audio, visual and audio-visual modalities, respectively. The filter approach provided a better performance for the audio and audio-visual modalities, whereas the wrapper approach performed better in the case of visual modality. The overall classification performance of the filter approach was slight better than the wrapper approach.

The humans' classification accuracy for seven emotions on the SAVEE database was 67% for audio, 88% for visual, and 92% for audio-visual data. The machine performance is lower than the humans' performance on SAVEE database. In future, we will further improve the classification performance by investigating the extraction of audio and visual features, feature selection, features reduction, and classification and fusion techniques.

REFERENCES:

Ashraf, A. B., S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon and B. J. Theobald (2007) The Painful Face – Pain Expression Recognition Using Active Appearance Models. Proc. ACM Int'l Conf. on Multimodal Interfaces, Japan, 9–14.

Bänziger, T., H. Pirker and K. R. Scherer (2006) GEMEP-GEneva Multimodal Emotion Portryals: A Corpus for the Study of a Multimodal Emotional Expressions. Proc. LREC Workshop on Corpora for Research on Emotion and Affect, Italy, 15–19.

Bartlett, M. S., G. Littlewort, M. Frank, C. Lainscsek, I. Fasel and J. Movellan (2005) Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, 568–573.

Bartlett, M. S., G. Littlewort, M. Frank, C. Lainscsek, I. Fasel and J. Movellan (2006) Fully Automatic Facial Action Recognition in Spontaneous Behavior. Proc. Int'l Conf. Automatic Face and Gesture Recognition, UK, 223–230.

Batliner, A., C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russel and M. Wong (2004) "You Stupid Tin Box" - Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech Corpus. Proc. Int'l Conf. on Language Resources and Evaluation, Portugal, 171–174.

Bishop, C. M. (2006) Sparse Kernel Machines, Pattern Recognition and Machine Learning, Springer, Cambridge, UK.

Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss (2005) A Database of German Emotional Speech. Proc. Interspeech, Portugal, 1517–1520.

Busso, C. and S.S. Narayanan (2007) Interrelation Between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study, IEEE Transactions on Audio, Speech and Language Processing 15(8):2331–2347.

Campbell, J.P. (1997) Speaker Recognition: A Tutorial. Proceedings of the IEEE 85(9):1437–1462.

Casale, S., A. Russo and S. Serranoa (2007) Multistyle Classification of Speech Under Stress using Feature Subset Selection Based on Genetic Algorithms. Speech Communication 49(10):801–810.

Douglas-Cowie, E., N. Campbell, R. Cowie and P. Roach (2003) Emotional Speech: Towards a New Generation of Databases. Speech Communication 40(1):33–60.

Ekman, P., W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, K. Scherer and M. Tomita (1987) Universals and Cultural Differences

in the Judgments of Facial Expressions of Emotion. Journal of Personality and Social Psychology 53(4):712–717.

Eyben, F., M. Wöllmer and B. Schuller (2009) openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. Proc. IEEE 4th Int'l HUMAINE Association Conference on Affective Computing and Intelligent Interaction, The Netherlands, 576-581.

Gunes, H. and M. Piccardi (2005) Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, USA, 3437–3443.

Haq, S. and P.J.B. Jackson (2009) Speaker-Dependent Audio-Visual Emotion Recognition. Proc. Int'l Conf. on Auditory-Visual Speech Processing, UK, pp. 53–58.

Haq, S. and P.J.B. Jackson (2010) Multimodal Emotion Recognition, Machine Audition: Principles, Algorithms and Systems, IGI Global Publisher, USA.

Haq, S., P.J.B. Jackson and J.D. Edge (2008) Audio-Visual Feature Selection and Reduction for Emotion Classification. Proc. Int'l Conf. on Auditory-Visual Speech Processing, Australia, 185–190.

Huckvale. M. (2015) Speech Filing System, http://www.phon.ucl.ac.uk/resource/sfs/. Last accessed on 1 January 2015.

Jackson, P. and S. Haq (2015) Surrey Audio-Visual Expressed Emotion (SAVEE) Database, http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/References.html. Last accessed on 1 January 2015.

Kanade, T., J. Cohn and Y. Tian (2000) Comprehensive Database for Facial Expression Analysis. Proc. IEEE Int'l Conf. Face Gesture Recognition, France, 46–53.

Kessous, L., G. Castellano and G. Caridakis (2010) Multimodal Emotion Recognition in Speech-Based Interaction using Facial Expression, Body Gesture and Acoustic Analysis, Journal on Multimodal User Interfaces 3(1): 33–48.

Kittler, J. (1978) Feature Set Search Algorithms, Pattern Recognition and Signal Processing, Sijthoff & Noordoff Int'l Publishers, The Netherlands, 41–60.

Lin, Y. and G. Wei (2005) Speech Emotion Recognition Based on HMM and SVM. Proc. Int'l Conf. Machine Learning and Cybernetics, China, 4898–4901.

Lorenzo, M., J. Edge, S. King and S. Maddock (2003) Use and Re-use of Facial Motion Capture Data. Proc. Vision, Video, and Graphics, UK, 1-8.

Pantic, M. and M.S. Bartlett (2007) Machine Analysis of Facial Expressions, Face Recognition, I-Tech Education and Publishing, Austria, 377–416.

Pantic, M, M. Valstar, R. Rademaker and L. Maat (2005) Web-Based Database for Facial Expression Analysis. Proc. ACM Int'l Conf. Multimedia and Expo, The Netherlands, 317–321.

Petrushin, V.A. (1999) Emotion in Speech: Recognition and Application to Call Centers. Proc. Artificial Neural Networks in Engineering, Missouri, USA, 7–10.

Picard, R. W. (1997) Affective Computing, MIT Press, Cambridge.

Schuller, B., B. Vlasenko, R. Minguez, G. Rigoll and A. Wendemuth (2007) Comparing One and Two-Stage Acoustic Modeling in the Recognition of Emotion in Speech. Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Japan, 596-600.

Sebe, N., I. Cohen and T. S. Huang (2005) Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision, World Scientific, Singapore.

Vlasenko, B., B. Schuller, A. Wendemuth and G. Rigoll (2007a) Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. Proc. Interspeech, Belgium, 2249-2252.

Vlasenko, B., B. Schuller, A. Wendemuth and G. Rigoll (2007b) Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. Proc. Int'l Conf. on Affective Computing and Intelligent Interaction, Portugal, 139-147.

Weston, J. and C. Watkins (1999) Multi-class Support Vector Machines. Proc. European Symposium on Artificial Neural Networks, Belgium, 219-224.

Young, S., G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland (2015) Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/.Last accessed on 1 Jan. 2015.

Zeng, Z., M. Pantic, G. I. Roisman and T.S. Huang (2009) A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1):39–58.