PROPOSING A HYBRID APPROACH FOR EMOTION CLASSIFICATION USING AUDIO AND VIDEO DATA

Reza Rafeh[1], Rezvan Azimi Khojasteh[2], Naji Alobaidi[3]

[1] Centre for Information Technology, Waikato Institute of Technology, Hamilton, New Zealand
[2]Department of Computer Engineering, Malayer Branch, Islamic Azad University, Hamedan, Iran
[3]Department of Computer Engineering, Unitec Institute of Technology, Auckland, New Zealand

ABSTRACT

Emotion recognition has been a research topic in the field of Human Computer Interaction (HCI) during recent years. Computers have become an inseparable part of human life. Users need human-like interaction to better communicate with computers. Many researchers have become interested in emotion recognition and classification using different sources. A hybrid approach of audio and text has been recently introduced. All such approaches have been done to raise the accuracy and appropriateness of emotion classification. In this study, a hybrid approach of audio and video has been applied for emotion recognition. The innovation of this approach is selecting the characteristics of audio and video and their features as a unique specification for classification. In this research, the SVM method has been used for classifying the data in the SAVEE database. The experimental results show the maximum classification accuracy for audio data is 91.63% while by applying the hybrid approach the accuracy achieved is 99.26%.

KEYWORDS

Emotion Classification, Emotions Analysis, Emotion Detection, SVM, Speech Emotion Recognition;

1. INTRODUCTION

Emotion is one of the humans' characteristics by using which they can express their wishes, interests, needs and goals as well as communicate with each other. Most emotional states involve physiological responses which affect the people's voice when they speak. For example, high-energy emotions such as anger often increases the vibration of the vocal cords and changes the shape and rhythm of the breathing muscles. Therefore, researchers have used data like human's voice in emotion recognition [1]. Recently, many research projects have been conducted in order to use this kind of data to achieve better results. Several hybrid approaches have been proposed for emotion classification based on text, speech, and image [2]. Changes in the voice are independent of the speaker and the language.

Usually, when classifying emotion, researchers consider only acoustic sound features [2]. This is while features like pitch, energy and changing the speed of emotional speech can also differ in

Natarajan Meghanathan et al. (Eds) : CSTY, AI, MaVaS, SIGI, FUZZY - 2019
pp. 31-40, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.91403

various emotions. Therefore, using acoustic features alone cannot accurately classify emotions especially strong emotions.

In addition to the use of spectral and prosodic features of voice in emotion classification, video features can be used as a complementary factor. In this way, the accuracy of classifying emotions can be improved.

In this research, a hybrid emotion classification is proposed which uses audio-video features. Emotions fall into seven classes: happiness, anger, fear, sadness, disgust, surprise and neutral. To evaluate the impact of using video features on the accuracy of classification, before applying the proposed hybrid method, classification is performed based on audio only. Then, the results are compared with classification when using both audio and video features.

The rest of the paper is organized as follows. Existing studies on emotion recognition and classification are reviewed in Section 2. The data collection process and the database are detailed in Section 3. Feature extraction is explained in Section 4. The hybrid method is proposed in Section 5. The experimental results are shown in Section 6. Finally, Section 7 concludes the paper and proposes directions for future research.

2. RELATED WORKS

Recognition of human emotions is a sense analysis and emotion classification topic by using a variety of physiological and audio and video signals or a combination of them. The aim is to maximize the accuracy of the classification and to propose a better solution for emotion recognition. In [3] negative emotions such as sadness, fear, stress and surprise were studied. Electro Dermal Activity (EDA), Electrocardiogram (ECG), Skin Temperature (SKT), Photo Plethysmography (PPG) were registered and analysed as physiological signals. Researchers have done studies about emotion recognition from human expression. They found that emotion changing will have a significant impact on individual voice features [4]. Therefore, emotional states and human sentimentality can be explored by investigating these features. The most important feature to use in extracting emotion from human voice is the basic frequency called f0. f0 is one of the sound pitch characteristics of speech. The voice of men is thicker than the voice of women. This is because basic frequency for men is 80-160 Hz, for women it is 150-250 Hz and for children the basic frequency is 200-400. Apart from basic frequency, speaking speed, speech quality, and energy are among the features that can be used in detecting the emotion. For some of these features, there have been several studies on speaker-dependent models [5] and several studies on speaker's dialect-dependent models [6]. There are also some models which are completely independent of the speaker [7].

Emotion recognition is considered in emotion classification as well. Speech emotion recognition has a wide range of applications including vehicle management system in which mental information state of drivers is recorded and is used for maintaining the safety of the driver [8]. Detecting the system from a low level of consciousness of the driver due to sickness or fatigue can reduce the driver's ability to control the vehicle. The system can manage these situations effectively in order to protect the lives of the drivers and passengers by identifying critical conditions. In [9] emotion is recognized by using audio signals for classification. Four classes have been used in the study: happiness, anger, sadness and neutral. In another research, Hidden Markov model (HMM) has been used for emotion recognition [10].

3. DATABASE

Using appropriate data is crucial when creating and evaluating new models. A variety of databases are available for emotion recognition some of which are publicly available for free.

eNTERFACEA'05 EMOTION is an audio-visual database in emotional contexts [11]. The AVICAR database [8] that was collected by researchers at the University of Illinois in 2003-2004 is a speech corpus in the vehicle and contains multi-channel audio-visual records. The IEMOCAP1 database is a multimodal information capture. It has been recently collected by the SAIL2 lab at the University of Southern California (USC) [12], [13].

SAVEE audio-visual database has been used in this study which includes data from native male English speakers between 27-31 years who are referred by these labels: KL, DC, DJ and JK [14]. Emotion has been described psychologically in discrete categories and seven classes: happiness, anger, fear, sadness, surprise, disgust and neutral. The text material consisted of 15 TIMIT sentences per emotion and 120 utterances for each speaker. This resulted in a total size of 480 utterances. Data was recorded dynamically and there are 60 painted markers on the actors' frontal face as facial markers. Figure 1 shows the distribution of database emotion classes. As we can see, the columns represent the number of emotional video files, and all emotions are equal except the neutral state.

Figure 1.SAVEE Database emotion classes distribution [15]

The database uses only male speakers which is a disadvantage for the database. This study classifies emotion independent from speakers' gender so it will not affect the operation of classification.

4. FEATURE EXTRACTION

A. Audio Features

Energy and related features: energy is a crucial feature of speech signals. To obtain the statistics of energy feature, the value of energy per frame of speech should be extracted. Thus, statistics of energy in the whole speech sample are obtained by calculating the energy, such as maximum value, minimum value, average and standard deviation [16].

Pitch and related features: pitch is another important feature in speech emotion recognition. The shape of vocal cords and how they vibrate are affected in different emotional states. Since pitch depends on vocal cords tension and pressure under larynx, and it also contains information about emotion. Pitch signal is also called glottal wave-form. Maximum value, minimum value, average and variation range are different in various emotions.

Mel-Frequency Cepstrum coefficients (MFCC): MEL frequency scale is a feature widely used in speech with a simple calculation. MFC has a good resolution in low frequency region and its strength is its excellent resistance against noise. However, the accuracy of emotion recognition is not satisfactory [6].

Formant, bandwidth for the first four formants: formants determination is based on vocal cords that are affected differently in different emotional states. For instance, the highest peak spectral peaks in the spectrum of sound is the first formant frequency. In other words, formant is the concentration of energy around a certain frequency. Linear predictive coding method is used for formant calculation [17].

B. Video Features

In image processing algorithms and functions, including "DetectFeatures", the features of the image can be extracted. The main issue is that the output is high dimension. Faces are marked by small blue signs in the SAVEE database. The marks are used to identify the essential and effective points in determining a facial expression and to reduce the dimensions of the features extracted. Colour tracking algorithms can be utilized to find these points. A sample of a database image is shown in Figure 2.

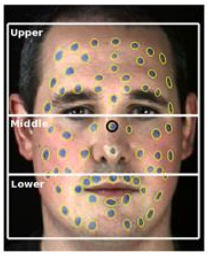

Figure 2. A sample of SAVEE database image and colour marker of face

As shown in Figure 2, marker on the edge of the nose (encircled in black) is taken as a reference. It is considered as the centre of coordinate, and the remaining coordinates are obtained based on it. With the extraction of these features, i.e., by using the same coloured points marked on faces a 130-dimensional set is obtained. This dimensional set includes only three speakers (JK, DC, DJ). It is due to the fact that 65 coloured points are detected on their faces. However, 60 points can be detected and visualized on the fourth speaker face (KL); therefore, extracted features from his related files include 120 dimensions.

To assimilate dimensional features, identify points of difference and their coordinate were removed from other speakers feature files. In Figure 3, the points of difference between the fourth speaker and others (here is JE, the second for example) can be seen. In addition, the second speaker face is marked with yellow circles to compare with the fourth speaker face.

Figure 3. Compare markers of two speakers in SAVEE database

5. EMOTION CLASSIFICATION

In recent years, researchers have suggested many classification algorithms for emotion recognition using audio. These algorithms include Neural Networks (NN) [18], Gaussian Mixture

Model (GMM) [19], Hidden Markov model (HMM) [10], [13], Maximum Likelihood Bayesian classifier (MLC), Kernel Regression, K-nearest Neighbours (KNN) [20], and Support Vector Machines (SVM) [21], [22]. SVM creates high-dimension vectors in the vector space which is actually the hyperplane max distance. Since SVM is a simple and efficient algorithm in machine learning, it is widely used for pattern recognition and classification problems. It could have a very good classification performance compared with other classifiers when the training data is limited. Thus, SVM is selected for emotion classification in this study as the classifier.

Hard margin SVM is non-linear. There is another type of extended SVM which is called soft margin where for the data items with class violation, a penalty coefficient C is defined. They are in the range of other classes. The amount of violation is shown byξ_i.

If (1) $\min \frac{1}{2}\|w\|^2 = \frac{1}{2}w^T w$

Then (2) $\min_{w.b.\xi} \left\{\frac{1}{2}\|w\|^2 + C\sum_i \xi_i\right\}$

$C > 0$ $l_i(w.x_i + b) \geq 1 - \xi_i$ $1 \leq i \leq n\,.\xi_i \geq 0$

The kernel function is used for non-linear SVM. In this way, it makes space of main entrance to high-dimensional feature space and keeps SVM non-linear and separate different classes. Calculating the feature space can be expensive in terms of size and, in general. This space has unlimited dimensions. Thus, the kernel is used to overcome this problem. RBF kernel function:

(3) σ>0 $K(x_i.x_j) = \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$

When using RBF the choice of two parameters is very important. Penalty C in case of conflict and constant σ in (2) and (3). If the values of C and σ are identified, then the classifier can predict emotions more accurately. To generalize SVM to more than two classes, OAA algorithm is used [16].

A: Emotion Classification Based on Audio Features

In this study, seven main emotions have been used for identifying emotions: happiness, sadness, anger, fear, disgust, surprise and neutral. Different variations are created in human voice features such as pitch, energy and spectrum in various emotional states. Initially, audio features are extracted using one of the feature selection methods. Thus, more effective features can be selected. In this study, z-score method of normalized has been used. Generally, the classification uses only audio features with 121 audio features including energy, pitch, formant, coefficient Mel and speed of speech and values associated with them. These features have been selected as they are common in most of the works done in this area and the experiments are repeated on multiple choices of audio features.

A classification task usually involves separating data into training and testing sets. The classifications were done by 7-fold Cross-Validation and they were examined by different values of sigma (σ) too. In section 6, the results of the experiments are investigated.

B: Emotion Classification on Audio-Visual Features (Hybrid Approach)

In the second phase, classification is done by combination both audio and video features. In other words, part of the work is common with classification based on only audio features. In this phase, extraction of features must be done first, and the process of dimension reduction must be accomplished. Finally, normalization should take place.

The vectors have been created from extracted features which have been used to train SVM in the hybrid approach should use audio features vectors and video features vectors simultaneously. A model is created based on SVM classifier and multi-classification done by OAA algorithm. 7-fold cross validation has been used for determining training and testing data sets. The classification accuracy can be improved by appropriate solutions including changes in selected features and checking other ways for feature extraction.

6. EXPERIMENTS

In this study, MATLAB 2016a has been used for implementation. The educational algorithms of the University of Rochester have been used [17] to extract audio features. They have been used to improve the performance of SVMs such as One-Against-All too. Using different kernels had excellent results to solve SVM problems including [15] who have used the polynomial kernel in their research. They could improve classification accuracy by increasing four percent. The support vector machine is used for classification. Compared to other classification methods, this method has proved to be effective and popular. In this research, the visual features presented in the SAVEE database have been used. There are several toolboxes to easily work for audio and video features extraction such as Open SMILE and Praat for audio features and Open CV for video features that can be used instead of writing an algorithm.

This study uses a limited number of features and has achieved good results compared with other emotion classifiers which are based on audio or audio-visual.

For the first stage, classification was done only with audio features in different conditions. Different and remarkable results were obtained. Table 1 shows the result of classification based on just audio features. In this study, we considered σ=9 and three different states of audio features. According to Table 1, if the selected audio features are taken energy and pitch, the classification accuracy is 75.62%. While adding formants to audio features set and retesting, classification accuracy is increased to 82.68%. Next, classification is done with new features set, this time taking into account Mel-Frequency Cepstrum coefficients and speed of speech besides energy, pitch and formants features, it can be seen that classification accuracy increases to 91% based on the result shown.

Table 1.Audio classification of 7 emotion with sigma (σ)=9

Features / Class	Energy, Pitch	Energy, Pitch, Formant	Energy, Pitch, Formant, MFCCs
Happiness	74.18	77.37	89.23
Anger	89.43	87.53	92.38
Sadness	65.97	83.94	94.07
Fear	75.25	84.11	88.80
Disgust	56.17	69.89	90.05
Natural	83.11	87.31	89.69
Surprise	85.19	88.63	92.79
Accuracy	**75.62**	**82.68**	**91.00**

Emotion classification was done again with the same audio features, but different values of σ. Values considered for σ were 5, 6.5, 8.5, 9, 10, of which 8.5 achieved the best result. Table 2 shows the overall result of models with three values of sigma.

Table 2. Obtained models accuracy with only audio features and different values for sigma (σ)

(σ)Sigma	Audio Features		
	Energy, Pitch	**Energy, Pitch, Formant**	**Energy, Pitch, Formant, MFCCs**
5	81.28	88.77	89.69
6.5	77.49	86.5	91.03
8.5	75.49	83.44	91.63

Maximum and minimum classification accuracy of seven emotions with only audio features are 91.63% for the model with audio features energy, pitch, formants, Mel-Frequency Cepstrum coefficients, speed and sigma as 8.5 and 75.49 for the model with audio features energy, pitch and sigma as 8.5, respectively.

Finally, we wanted to test the impact of adding video features on accuracy. Thus, we repeated experiments and changed the number of features and sigma with audio and visual features together. The results are summarized in Table 3.

Table 3. Models accuracy by audio-visual features with different values for sigma (σ)

Sigma	Audio-Visual Features		
	Energy, Pitch, VF	**Energy, Pitch, Formant, VF**	**Energy, Pitch, Formant, MFCCs, VF**
5	98.95	96.65	90.06
6.5	99.26	97.86	95.04
8.5	99.16	98.76	97.58

The maximum classification accuracy of seven emotions by the hybrid approach is 99.26% achieved from the model with energy, pitch as audio features and video features and sigma as 6.5.Classification accuracy with the same conditions but using only audio features was 77.49. The minimum accuracy was 90.06% obtained from the model with energy, pitch, formants, Mel-Frequency Cepstrum coefficients (MFCC), speed as audio features and video features and sigma as 5.Using the same conditions classification accuracy with only audio features was 89.69%. The comparison of classifications based on only audio features to the hybrid approach (classification on audio-visual features) determines that the hybrid approach increases classification accuracy in all three models. Therefore, the proposed hybrid approach produced more promising results.

[23] used audio signals for emotion recognition and used SAVEE database. Selected features in the project are mainly related to energy, pitch, and statistics and spectral features MFCC as well. They recognized emotions by using linear kernel with binary tree classification strategy OAO and OVsR. The best results of that work and this research with the same number of common data and the same audio features are shown in Table 4. By comparing the results in Table 4, good performance of classification by a hybrid approach with the audio-visual features can be seen.

Table 1. Comparing Sinith's project with the proposed hybrid approach

Class	Sinith's work	Hybrid Approach
Anger	65	96.55
Happiness	45	98.78
Natural	70	98.78
Sadness	65	97.5
Accuracy	**61.25**	**97.9**

The best accuracy in [23] is 61.25% using linear kernel and binary tree, while by using the proposed method and the same database, the hybrid approach accuracy equals to 97.90%. In [10]Hidden Markov model is used and they used SAVEE database with four classes: surprise, sadness, fear and disgust. The work used only one audio feature which is Mel-Frequency Cepstrum coefficients (MFCCs) to recognize emotions. The accuracy rate of emotion recognition is 94.17% while using the same database and the same feature, the proposed hybrid approach can improve the accuracy to 97.82%. Table 5 shows the results and their comparison.

Table 5.Comparing Chandni's project with the proposed hybrid approach

Class	Chandni's work	Hybrid Approach
Anger	90	98.24
Happiness	100	99.09
Natural	97	94.78
Sadness	90	99.18
Accuracy	**94.17**	**97.82**

7. CONCLUSION

In this paper, a hybrid approach was introduced for classifying emotions. The proposed approach uses both audio and video data and SVM as the classifier. The proposed approach achieved 99.16% accuracy which is higher than similar previous researches.

This study was aimed at introducing an emotion recognition system independent of language and speaker. Existing researches in classification and emotion recognition use different audio features. In this study, two audio features have been used: prosodic features and spectral features. We intend to investigate the effect of dialect and language on emotion recognition in a close future. It is expected that recognizing emotions in different languages and analysing their audio features will improve the accuracy of the classifier. A feature may be a prominent transmitter of emotion in one language but not in another language or the effect of audio features on one language may be different from other languages. Another interesting topic would be the impact of the accent on expressing emotions when analysing audio features. We hope our future research can address these concerns. We may need to create especial databases for different languages and investigating the effective features in every language based on the database. Also, multi-faceted analysis and using hybrid approaches can provide a rich set of information about the hidden aspects of the text, audio, visual, image and any other transmitter. In other words, the information taken from each side can complete the other aspects.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Hosein Darabi for his help and support. This research was funded by Wintec Research Office and Wintec Data Science Group.

REFERENCES

[1] Ververidis,Dimitrios,Kotropoulos, Constantine, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol. 48, no. 9, pp. 1162-1181, 2006.

[2] Bhaskar, Jasmine, Sruthi, K. Nedungadi and Prema, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining," Procedia Computer Science, vol. 46, pp. 635-643, 2015.

[3] E. H. Jang, B. J. Park, S. H. Kim and J. H. Sohn, “Emotion classification based on physiological signals induced by negative emotions: Discriminantion of negative emotions by machine learning,” in Networking, Sensing and Control (ICNSC), 2012 9th IEEE International Conference on Beijing, 2012.

[4] C. Parlak. and B. Diri, “Emotion recognition from the human voice,” in Signal Processing and Communications Applications Conference (SIU), 2013 21st, 2013.

[5] E. Ayadi, M. Kamel, M. S. and K. Fakhri, “Survey on speech emotion recognition: Features, classification schemes, and databases,” vol. 44, no. 3, pp. 572-587, 2011.

[6] Y. Pan, P. Shen and L. Shen, “Speech Emotion Recognition Using Support Vector Machine,” International Journal of Smart Home, vol. 6, no. 2, pp. 101-108, 2012.

[7] C. Lijiang, M. Xia, X. Yuli and C. L. Lung, “Speech emotion recognition: Features and classification models,” Digital Signal Processing, vol. 22, no. 6, pp. 1154-1160, 2012.

[8] N. Rajitha, D. David, L. B, P. J., Sridharan, S. Fookes and C. B., “Recognising audio-visual speech in vehicles using the AVICAR database,” in Proceedings of the 13th Australasian International Conference on Speech Science and Technology Melbourne, Vic, 2010.

[9] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema and S. Rajan, “Emotion recognition from audio signals using Support Vector Machine,” in IEEE Recent Advances in Intelligent Computational Systems (RAICS) Trivandrum, 2015.

[10] G. Chandni, M. Vyas, K. Dutta, K. Riha and J. Prinosil, “An automatic emotion recognizer using MFCCs and Hidden Markov Models,” in Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2015 7th International Congress on Brno, 2015.

[11] “eNTERFACE'05 EMOTION Database,” [Online]. Available: http:// www.enterface.net/enterface05/..

[12] C. Busso, M. Bulut, C CLee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee and S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” vol. 42, pp. 335-359, 2008.

[13] A. Metallinou, C. Busso, S. Lee and S. Narayanan, “Visual emotion recognition using compact facial representations and viseme information,” in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing ,Dallas, TX, 2010.

[14] “SAVEE Database,” [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/Database.html.

[15] M. Sidorov, E. Sopov, I. Ivanov and W. Minker, “Feature and decision level audio-visual data fusion in emotion recognition problem,” in Informatics in Control, Automation and Robotics (ICINCO), 2015 12th International Conference on Colmar, 2015.

[16] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman and M. Sturge-Apple, “Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion,” in Spoken Language Technology Workshop (SLT), 2012 IEEE Miami, FL, 2012.

[17] “Bridge Project,” 2013. [Online]. Available: http://www.ece.rochester.edu/projects/wcng/project_bridge.html.

[18] E. Sopov and I. Ivanov, “elf-Configuring Ensemble of Neural Network Classifiers for Emotion Recognition in the Intelligent Human-Machine Interaction,” in Computational Intelligence, 2015 IEEE Symposium Series on Cape Town, 2015.

[19] S. Agrawal and S. Dongaonkar, “Emotion recognition from speech using Gaussian Mixture Model and vector quantization,” in Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on Noida, 2015.

[20] M. R. Mehmood and H. J. Lee, “Emotion classification of EEG brain signal using SVM and KNN,” in Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on Turin, Italy, 2015.

[21] N. R. Kanth and S. Saraswathi, “Efficient speech emotion recognition using binary support vector machines & multiclass SVM,” in IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) Madurai, 2015.

[22] Y. Chavhan, M. L. Dhore and P. Yesaware, “Article: Speech Emotion Recognition Using Support Vector Machine,” vol. 1, pp. 6-9, 2010.

[23] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema and S. Rajan, “Emotion recognition from audio signals using Support Vector Machine,” in IEEE Recent Advances in Intelligent Computational Systems (RAICS) Trivandrum, 2015.

[24] A. Metallinou, A. Katsamanis, W. M, F. Eyben, B. Schuller and S. Narayanan, “Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract),” in Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on Xi'an, 2015.

AUTHORS

Reza Rafeh is a senior lecturer at Waikato Institute of Technology. He received his PhD in computer science from Monash University. His research areas cover data mining, big data and analytics, recommender systems, software engineering and modelling, constraint programming, and health informatics.

Rezvan Azimi Khojasteh received her MSc in Software Engineering from Islamic Azad University, Malayer Branch. Her research area includes emotion mining and data analytics.

Naji Alobaidi received his MSc in Computer Science from Unitec Institute of Technology. His research areas cover data analytics, emotion mining, and vehicular ad-hoc networks.