Biologically Inspired Cognitive Architectures xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

journal homepage: www.elsevier.com/locate/bica

Research article

Emotion recognition from Marathi speech database using adaptive artificial neural network

Raviraj Vishwambhar Darekar[a,*], Ashwinikumar Panjabrao Dhande[b]

[a] Research Scholar, Dr. D.Y. Patil Institute of Technology, Pimpri, SPPU, Pune, India.
[b] Pune Institute of Computer Technology, Pune, India

ARTICLE INFO

Keywords:
Emotions
Recognition
Multimodal fusion
Hybrid PSO-FF
Classifier

ABSTRACT

Nowadays, recognition of emotion from the speech signal is the wide spreading research topic since the speech signal is the quickest and natural approach to communicate with humans. A number of investigations has been progressed related to this topic. With the knowledge of many investigated model, this paper intends to recognize the emotions from the speech signal in a precise manner. To accomplish this, we intend to propose an adaptive learning architecture for the artificial neural network to learn the multimodal fusion of speech features. It results in a hybrid PSO-FF algorithm, which combines the features of both the PSO and FF towards training the network. The performance of the proposed recognition model is analyzed by comparing it with the conventional methods in correspondence with varied performance measures like Accuracy, Sensitivity, Specificity, Precision, FPR, FNR, NPV, FDR, F_1Score and MCC. Finally, the experimental analysis revealed that the proposed modal is 10.85% better than the conventional modals with respect to the accuracy for both the Marathi database and Benchmark database.

Introduction

Since the speech signal is one of the natural approaches to communicating one another, researchers are motivated to develop interaction approaches among humans as well as machines. However, the vital fact is that machine must have necessary intelligence for the recognition of human's voice (Grimm, Kroschel, Mower, & Narayanan, 2007; Rong, Li, & Chen, 2009; Shami & Verhelst, 2007). From the past fifty years, a number of research on speech recognition has been developed for the process of converting human speech into the series of words. Although great progress has been made in recognition of speech, improvement is required to progress the recognition in a precise manner since the machine does not understand the emotional state of the humans (Murray & Arnott, 2008; Ververidis & Kotropoulos, 2008).

One of the robust recognition systems named SER has the great objective of automatic identification of human emotional states from his/her voice (Morrison, Wang, & De Silva, 2007; O'Shaughnessy, 2008). This system focuses on speech signals. Basically, speech signal has some major noticeable characteristics like carrying data about the human's or speaker's age, religion, gender, origin and emotion stage (Garvin & Ladefoged, 1963). Moreover, emotions of humans have a prominent impact on the extracted features from their speech signal (Cowie & Cornelius, 2003; Luengo, Navas, & Hernaez, 2010). Emotions of human can be categorized into six classes including happy, sad, anger, disgust, anxiety and surprise (Ekman, 1992). It also plays a vital role in decision making, problem-solving, and intelligence, because the vision of future computation relies on artificial intelligence. Thus, the performance of the emotion recognition model must be robust, and therefore, it is highly useful in HCI (Calvo & D'Mello, 2010; El Ayadi, Kamel, & Karray, 2011). HCI is normally an important aspect of emerging an automatic SER. In addition, the rising of HCI real-world applications along SER has a large influence on real-world environments. More especially in diagnosis fields, SER is the major tool for the treatment of mental disorder via speaker interaction as well as games (Kostoulas et al., 2012). SER is also one of the important measures in the detection of diseases like Parkinson and Alzheimer (Lopez-de-Ipiña

Abbreviations: PSO, Particle Swarm Optimization; FF, Firefly; FPR, False Positive Rate; FNR, False Negative Rate; NPV, Negative Predictive Value; FDR, False Discovery Rate; MCC, Mathews correlation coefficient; SER, Speech Emotion/Stress Recognition; HCI, Human-Computer Interface; MFCC, Mel-frequency Cepstrum Coefficients; LPCC, Linear Predictive Cepstrum Coefficients; SVM, Support Vector Machines; HMM, Hidden Markov Model; KNN, k-Nearest Neighbors; MLC, Maximum Likelihood Bayesian Classifier; ANN, Artificial Neural Network; NMF, Non-negative Matrix Factorization; FSS, Feature Subset Selection; SAVEE, Surrey Audio-Visual Expressed Emotion Database; SUSAS, Speech under Simulated and Actual Stress; BES, Berlin Emotional Speech Database; FDA, Functional Data Analysis; PCA, Principal Component Analysis; MLLR, Maximum Likelihood Linear Regression; DSL, Double Sparse Learning; DaLSR, Domain-Adaptive Least-Squares Regression; LSR, Least-Squares Regression; STFT, Short-time Fourier Transform; NN, Neural Network; LM, Levenberg Marquardt

* Corresponding author.
E-mail address: ravirajvishwambhardarekar@gmail.com (R.V. Darekar).

https://doi.org/10.1016/j.bica.2018.01.002
Received 7 September 2017; Received in revised form 27 November 2017; Accepted 9 January 2018

et al., 2013). Furthermore, SER is the primary useful tool for the people who have autism and can also be applicable in environments such as learning environment, entertainment, games, educational software and lie detection system (Cowie et al., 2001).

In general, two types of features like prosodic and spectral features are used for recognition of emotions, since both the features comprise emotional information or data. Some of the foremost spectral features are MFCC and LPCC. In order to design various emotions, some major prosodic features are used including loudness, pitch, frequency, glottal parameters as well as speech intensity (Zhou, Sun, Zhang, & Yan, 2009). Varied classifiers are also available for SER such as SVM (Li, Tian, Li, Zhou, & Yang, 2017), HMM (Gómez-Lopera, Martínez-Aroza, Román-Roldán, Román-Gálvez, & Blanco-Navarro, 2017), Kernel Regression and KNN (Maillo, Ramírez, Triguero, & Herrera, 2017), MLC (Stella & Amer, 2012), and ANN (Sitton, Zeinali, & Story, 2017).

Generally, the speech signal processing is extensively exploited in many applications such as remote observation of patients, non-contact medical diagnosis, human-computer interaction (HCI), etc. Moreover, the speech signals can be analyzed to detect heart rate that can naturally enable a remote diagnosis of a patient by exploiting audio data. The main aim of this paper is to recognize the emotions from the speech signals in an accurate manner. In order to accomplish this, this paper intends to propose an adaptive architecture for artificial neural network to learn the multimodal fusion of speech features. The speech features include NMF, cepstral and prosodic features. Multimodal features can represent each speech signal at unique margin, which is highly recommended by the researchers. The organization of the paper is represented as follows. Section ‘Literature review’ illustrates the summarized literature review, and Section ‘Analysis of emotional speech’ depicts the analysis of emotional speech. Section ‘Hybrid classifier for emotion recognition’ presents the hybrid classifier for emotion recognition. Section ‘Results and discussion’ presents the details associated with the simulation results. Section ‘Conclusion’ states the conclusion of the paper.

Literature review

In Yogesh et al. (2017a) has offered a study to choose the spectral features for SER systems. They has extracted some features like bicoherence and bispectral and its corresponding glottal waveform. The extracted features were combined with Inter-Speech 2010 for the improvement of recognition rates. In the proposed model, they have also implemented the FSS. FSS normally comprises two major stages. In the initial stage, they have adopted the selection of Multi-cluster Features for the diminishing of feature space and also to identify its relevant feature subset from Interspeech 2010 features. In the second stage, they have adopted Biogeography and Particle swarm BBO_PSO Hybrid optimization for the reduction of dimensions of feature space and also to find the exact feature subset. The proposed model was compared to the conventional methods and from the results it was found that the introduced model was more efficient than the existing models.

In Yogesh et al. (2017b) have proposed a new features selection algorithm based on Biogeography. They have simulated the proposed model with the aid of database named SAVEE, SUSAS and another database namely BES. In addition, they have evaluated the model with the help of eight benchmark datasets and those datasets had varied classes and dimensions. They have conducted a total of eight varied experiments and analyzed the performance of the model by comparing it with conventional methods. The results of the investigation have proved the effectiveness of the introduced feature selection algorithm.

In Arias, Busso, and Yoma (2014) have developed the usage of neutral reference methodology for the detection of local emotional prominence. They have presented a new approach on the basis of FDA that has the aim of capturing the intrinsic variability of F0 contours. They have also determined the neutral model, which was based on functions and testing of F0 contour. They also evaluated the PCA for the detection of emotions. The proposed model was assessed with models like lexicon-independent and dependent model. The investigational results have shown that the proposed system could lead to more accuracy of 75.8%. Moreover, the developed approach could be implemented at the sub-sentence level, which facilitates the detection of emotional data. The method was validated with the SEMAINE database, which was a spontaneous corpus. The investigational results have designated that the proposed scheme could be effectively employed in real-world applications in terms of recognizing emotions.

In Kim and Park (2016) have dealt the efficient SER model which have utilized the data accumulated on devices. With the advantage of the personal devices, they have developed an SER model that employed MLLR. This was a speaker adaption technique, which has chosen the useful information that conveys the acoustic emotion characteristics. With the use of log-likelihood distance-based measure as well as universal background approach, they have conducted multistage collection. The proposed model was compared to the existing methods and from the results, it has proved the superiority of the proposed approach.

In Deng, Xu, Zhang, Frühholz, and Schuller (2017) have suggested a new unsupervised domain adaptation methodology named Universum auto encoders for the improvement of system performance in various testing as well as training conditions. In order to address the mismatch, the developed model does not only learned the data from the labeled information, but also learned for the incorporation of the prior knowledge from unlabelled data. The experimentation of the proposed model was conducted on the labeled Geneva Whispered Emotion Corpus database and with other three unlabeled databases. The results have shown the effectiveness of the proposed model while comparing with the existing models.

In Deng, Xu, Zhang, Frühholz, and Schuller (2016) have investigated the efficiency of phase info for whispered recognition of speech emotion. They have nominated two kinds of phase-based features and both the features have shown the wide applicability to the entire sorts of varied speech analysis. With the exploitation of those features, they have introduced a novel speech recognition of emotion and have employed the outer product with the combination of L2 normalization as well as power. In accordance with the developed technique, they have encoded the variable length series of the phase based features. The resultant demonstration was fed to train a classification approach with a linear kernel classifier. The proposed model was analyzed and compared with the conventional methods. Experimental results have demonstrated the efficiency of the proposed method when compared with other recent systems.

In Zong, Zheng, Cui, and Li (2016) have handled the problems associated with SER using DSL model. They have an integrated pyramid structure-based feature extraction approach to this model. The main advantageous part of the proposed model was the selection of features by concerning two scales of the pyramid structure-based features, which have considered as a great contribution to SER. The respective proposed model was carried out in both AFEW and eNTERFACE emotion databases to validate the effectiveness of the proposed model. Later the experimental results have confirmed the ability of DSL with the pyramid structure-based feature extraction approach in SER process.

In Zong, Zheng, Zhang, and Huang (2016) have introduced the DaLSR model for developing the cross-corpus SER method. They have linked the unlabeled data set from target speech corpus as an auxiliary data with the labeled training data set to train the DaLSR approach. One of the beneficial parts of DaLSR model was its ability of handling mismatch problem existed between the source and target speech corpora, as compared to the LSR method. In fact, the experiment was carried out in three emotional speech corpora and have compared with several traditional algorithms associated with solving SER. Thus the performance comparison has proven the better performance of the DaLSR method as it attains high accuracy.