Speech Emotion Recognition Based on Gender Influence in Different Languages Using Various Classifiers

Abstract. Speech emotion recognition (SER) is an attractive and challenging task in human-computer interaction and artificial intelligence technologies. SER is the process of recognizing emotions from speech utterances. This work is based on emotion recognition from speech, and our purpose is to study the effect of gender classes on the SER. Three emotional databases with different languages: ADED, EMO-DB and ShEMO in Algerian dialect, German and Persian languages respectively, are used to evaluate the performance of the SER system. Extracting features from speech is a required step in the creation of the SER system. Combinations of prosodic and MFCC features are exploited as speech features in the system of recognition. This last is based on Linear Discriminate Analysis (LDA), Deep Neural Networks (DNN) and Support Vectors Machine (SVM) as methods of classification. The results obtained show us that the system of SER is influenced by gender classes. The recognition rates of SER systems with gender distinction are higher than the recognition rates of SER systems without gender distinction.

Keywords: DNN, intensity, LDA, MFCC, pitch, Speech emotion recognition, SVM.

1 Introduction

The human-machine interface becomes more significant if the machines can recognize emotions. Speech signal processing to recognize emotions has become one of the important areas of speech research. Speech emotion recognition (SER) has wide applications in human-machine interaction and artificial intelligence technologies. SER is applied in intelligent tutoring systems, speech translation systems, telephone banking, call center conversation, robots, lie detection, and medical field [1].

Speech emotion recognition is the technique used to recognize the emotional state of a speech signal. SER is defined as the task of automatically classifying emotions from an unknown utterance of speech from a list of emotions. The purpose of this work is to study the effect of gender classes on the SER system. We based on anger and neutral emotions on the system of recognition. Each SER system needs a database for its operation. Three emotional speech databases with different languages are exploited in this work. These databases are the Algerian Dialect Emotional Database (ADED) [2], Berlin Database of Emotional Speech (EMO-DB) [3] and the Sharif Emotional Speech Database (ShEMO) [4]. To design a system for emotion recognition in speech, choosing suitable features is a crucial step. A combination of speech features, including prosodic (pitch and intensity) and MFCC features are used in this

work. The SER systems are based on classification methods. The system of SER in this work is based on three classifiers: Linear Discriminate Analysis (LDA), Deep Neural Networks (DNN) and Support Vectors Machine (SVM).

This document is structured as follows: Some of the previous works are discussed in the second section. In the third section, the methodology of this work is described, including emotional databases, speech features, and classifiers. Section four presents the experiments and results. Finally, we finish with a conclusion.

2 Previous works

Speech signal contains various important information such as language, emotion, gender, and phonetic information. Recognition of emotion from speech is a very important field for researchers. The technique for automatically identifying the emotion of a particular speech is called Speech Emotion Recognition (SER). Some emotional databases, speech features, and classification methods used in SER are presented briefly in this section.

Selection of databases for training and testing of speech emotion recognition systems is one of the major issues of debate in research. A suitable database is crucial for SER to improve its performance. Among the most used databases are the Berlin database (EMO-DB) [3], the Danish (DES) database [5] and the Polish emotional database [6]. EMO-DB has been widely used to improve and develop several emotional recognition systems, and it has been exploited to make comparisons with other databases. This database contains seven emotions. The Danish (DES) database contains five emotional states. The database of Polish emotional speech contains 288 segments pronounced in six emotions. There are many other speech emotional databases in different languages, such as the SAVEE database in English [7], KISMET [8] in American English [9], IITKGP-SEHSC in Indian [10], Sharif Emotional Speech Database (ShEMO) in Persian [4], and CEMO in French [11]. There are a few databases on Arabic speech and emotion. Among these databases, we cite the Tunisian dialect database [12], the Moroccan emotional database (MEDB) [13], the Emirati speech database (ESD) [14] and the Egyptian Arabic speech emotion (EYASE) [15].

Extraction of features from speech is a very important step in analyzing the signal of speech. To recognize information such as emotion, gender, and language from speech, speech features related to the information involved were needed. Thus, the performance of the SER system depends on speech features. In literature, for the SER, various speech features have been used to develop the SER systems. Prosodic features, including pitch, intensity, and duration, have been widely exploited for recognizing speech emotions. Statistical values of pitch and energy features have been used to classify the emotional states in the SER systems [16]. Pitch and intensity have been employed as speech features to recognize emotions in Telugu speech [17]. Spectral features such as Mel Frequency Cepstrum Coefficients (MFCC), Linear Prediction (LP), Linear Predictive Cepstral Coefficients (LPCC) and perpetual linear prediction (PLP) have been used in many works of SER [16]. The MFCC and DWT (Discrete Wavelet Transform) have been investigated in the classification system of different emotions from speech [18]. Voice quality features like jitter, shimmer and Harmonic to Noise Ratio (HNR) have been also used in the system of SER in numerous works

[16]. Jitter and shimmer have been utilized for recognizing different emotions by using multiple classifiers [19]. Different combinations of various speech features have been exploited to improve the performance of recognition in the SER systems. In [20], a combination of various features such as fundamental frequency (F0), energy, formants, and energy in Mel and MFCCs has been explored to classify speech emotions. A combination of spectral, vocal quality and prosodic features has been exploited in emotion recognition using different databases [21]. Speech features such as duration, energy, pitch, MFCC, PLP, HNR, shimmer, jitter, and teager operator have been used in systems of automatic speech recognition [22].

Classifiers are essential and important parts of emotion recognition systems. To develop SER systems, several classification methods have been applied. Support Vectors Machine (SVM), Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Deep Neural Networks (DNN) were the most commonly used classifiers in the systems of SER [16][23].

3 Methodology

The objective of this work is to study the effect of gender classes on the system of speech emotion recognition (SER). The process for building an SER system requires emotional emotion databases, features, and classification techniques. Scheme of the SER systems is illustrated in figure 1. To achieve our aim, three databases with different languages are used: the Algerian Dialect Emotional Database (ADED), the Berlin Database of Emotional Speech (EMO-DB) and Sharif Emotional Speech Database (ShEMO). The influence of gender on SER is evaluated in two emotional states: anger and neutral. The system of SER is divided into a gender distinction part and without gender distinction part. In the gender distinction part, the speech utterances of each database are divided into male and female utterances. For the feature extraction step, the prosodic features (pitch and intensity) and the MFCC features are exploited. In the classification step, three classifiers: LDA, SVM and DNN are applied to evaluate the performance of SER. The response in the form of a recognition rate is obtained and studied for its accuracy.

3.1 Emotional speech databases

A suitable database is required for the speech emotion recognition (SER). As mentioned above, three databases: ADED, EMO-DB and ShEMO are used to achieve our objective. The Algerian Dialect Emotional Database (ADED) is an Algerian database. ADED database composed of 200 audio files, contains four emotional states [2]. The Berlin database of emotional speech (EMO-DB) is a German database; it contains 535 sound files recorded into 7 emotion states [3]. The Sharif Emotional Speech Database (ShEMO) is a Persian database; it consists of 3000 speech samples in six different emotions [4].

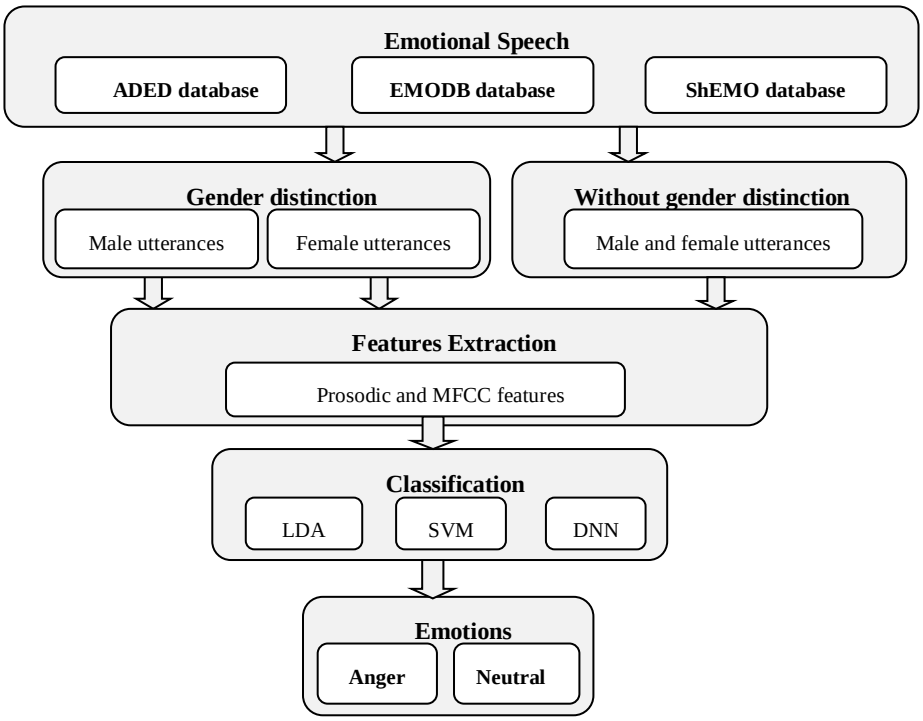

Fig. 1. Scheme of emotions recognition systems based on gender classes.

3.2 Features extraction

Extraction of features is a crucial step in recognition of emotion from speech. The statistical values of pitch and intensity, and MFCCs parameters are exploited in this paper. Table 1 show the features selected in this work. PRAAT software [24] is exploited to directly extract the statistical values of pitch and intensity features. The MFCCs parameters are extracted using MATLAB.

Table 1. Speech features extracted in this work.

Features extracted	
Prosodic features	Mean of pitch Maximum of pitch Minimum of pitch Standard deviation of pitch Mean of intensity Maximum of intensity Minimum of intensity
MFCC	13 MFCCs

Speech signal consists of unvoiced and voiced parts. This last is produced by the periodic vibration of the vocal cords. Its periodicity is known as pitch [25]. The pitch contour of speech is shown in Figure 2 by a blue line. The pitch feature has been used to develop the SER systems in many works [13],[16–17].

Intensity provides a measure of the voice loudness and represents the force with which the sound is emitted [26]. The contour of intensity is shown by a green line in figure 3. The intensity feature has been investigated in the SER systems [13],[17].

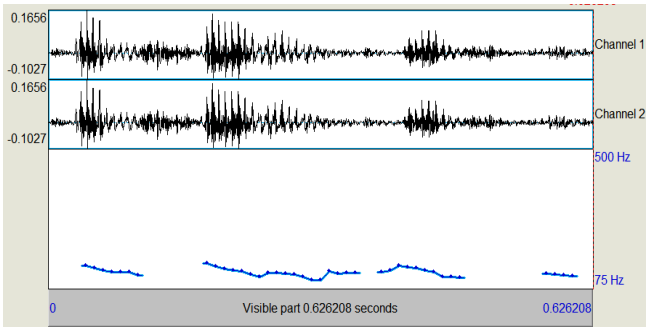

Fig. 2. Speech signal and its corresponding pitch contour.

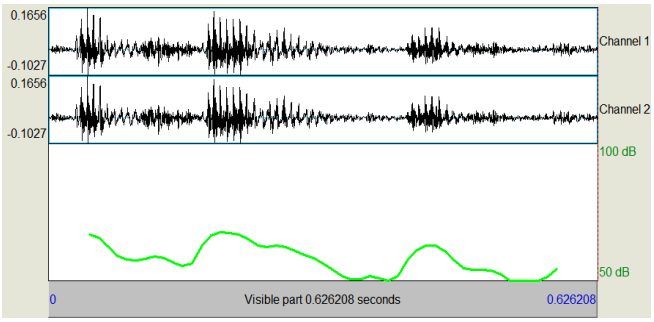

Fig. 3. Speech signal and its corresponding intensity contour.

Mel Frequency Cepstrum Coefficients (MFCC) features are the most spectral features used in the field of speech emotion recognition [13][18][21-22]. MFCC based on how we actually percept the speech sound. MFCC represents the power spectrum of a speech signal based on a non-linear scale of frequency (Mel-scale).

The formula to convert the frequency to Mel scale frequency:

$$m(f) = 2595 \log(1 + f/700) \quad (1)$$

Figure 4 shows the procedure to calculate the MFCC parameters. The procedure is composed of the following blocks: pre-emphasize, hamming window, FFT, triangular band-pass filter, logarithm, and discrete cosine transformation (DCT) [27].

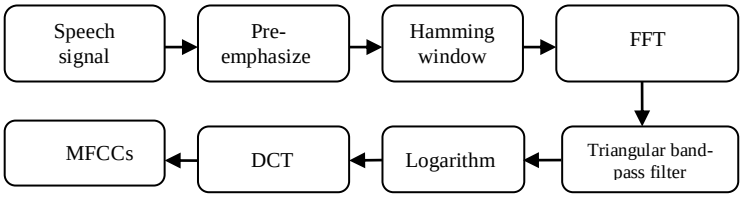

Fig. 4. Scheme of the procedure to calculate the MFCC parameters.

3.3 Classifiers

Classifier is perquisite step in SER systems. Three methods of classification are exploited for SER in this work. These classifiers are: Linear Discriminate Analysis (LDA), Deep Neural Networks (DNN) and Support Vectors Machine (SVM). LDA is a robust and simple classification technique with a linear decision boundary. The method of linear separation finds the linear transformation and maximizes the distance between the two classes in the new space [28]. Deep Neural Networks (DNN) has achieved great success in classification tasks. DNN is a feed-forward neural network with fully connected layers and hidden layers. Between the input and output layer more hidden layers are present [29]. Support Vectors Machine (SVM) is a simple machine learning that is applied in classification fields. The process in SVM is to transform the input features into a high dimensional feature space by using the kernel function [30].

4 Experiments and results

As we mentioned above, the aim of this work is to study the effect of gender classes on the speech emotion recognition (SER) in three different languages. To achieve our goal, three databases in three different languages are used, ADED, EMO-DB and ShEMO databases in Algerian, German and Persian, respectively. Several experiments are performed to study the influence of gender on SER. The systems of SER in the experiments are divided into system with gender distinction (male or female) and system without gender distinction (male and female). A combination of prosodic and MFCC features are used as feature vectors in the system of recognition. The experiments are based on LDA, SVM and DNN classifiers. The results are obtained using MATLAB software. The experiment results are presented in table 2. According to the results obtained, it is observed that the recognition rates in system with gender distinction are higher than the recognition rates in system without gender distinction. It is concluded that the recognition accuracy of the SER is influenced by the gender classes. In previous works, the effect of gender classes on SER was studied. The recognition accuracies of system with gender distinction are higher than the system without

gender distinction in [31], [32] and [33] which used German, Indian, and English emotional databases, respectively.

Table 4. Experiments results of gender effect on SER.

Databases	Gender class	Classifiers		
		LDA	***SVM***	***DNN***
ADED Database	Without gender distinction (M and F)	*79.17%*	*83.33%*	*85.41%*
	Male utterances (M)	*88.63%*	*93.18%*	*100%*
	Female utterances (F)	*93.18%*	*97.72%*	*95.45%*
EMO-DB Database	Without gender distinction (M and F)	*94.44%*	*93.05%*	*94.44%*
	Male utterances (M)	*100%*	*100%*	*100%*
	Female utterances (F)	*98.61%*	*100%*	*100%*
ShEMO Database	Without gender distinction (M and F)	*84.02%*	*68.75%*	*84.72%*
	Male utterances (M)	*93.05%*	*94.44%*	*100%*
	Female utterances (F)	*88.89%*	*93.05%*	*91.66%*

5 Conclusion

Speech emotion recognition (SER) plays an important role in human-machine interaction and artificial intelligence fields. This work was based on the recognition of speech emotions. And the purpose was to study the effect of gender classes on SER in different languages. Three databases with different languages, ADED, EMO-DB, and ShEMO, were used. Combinations of prosodic and MFCC features were exploited as features vectors in the SER system. This last was based on three classification techniques: Linear Discriminate Analysis (LDA), Deep Neural Networks (DNN) and Support Vectors Machine (SVM). Several experiments were performed to study the influence of gender on SER. It is concluded that the performance of SER is influenced by the gender classes. It was observed that the recognition accuracy in the system with gender distinction is higher than the recognition accuracy in the system without gender distinction. In the future, the present work can be extended to other emotions, speech features, and emotional databases in other languages.

References

1. Ramakrishnan, S., El Emary, I. M. M.: Speech emotion recognition approaches in human computer interaction. Telecommunication Systems 52 (3), 1467–1478(2011).

2. Horkous, H., Guerti, M. Recognition of Emotions in the Algerian Dialect Speech. International Journal of Computing and Digital Systems , 10(2): 245- 254, 2021.
3. Burkhardt, F., Paeschke, A., Rolfe, M., Sendlmeier, W., Weiss, B. A database of german emotional speech. Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal. pp.1-4, 2005.
4. Nezami, O. M., Jamshid, P., Karami, M. ShEMO: a large-scale validated database for Persian speech emotion detection. Language Resources and Evaluation, 53: 1–16, 2018.
5. Engberg, I., Hansen, A. (1996). Documentation of the Danish emotional speech database DES.Center for Person Communication, Institute of Electronic Systems, Alborg University, Aalborg, Denmark.
6. Staroniewicz, P., Majewski, W. Polish Emotional Speech Database – Recording and Preliminary Validation. Lecture Notes in Computer Science, 42–49, 2009.
7. Jackson, P., Haq, S,. Edge. J. D Audio-visual feature selection and reduction for emotion classification. In: Proc. Int'l Conf. on Auditory-Visual Speech Processing. pp. 185–90, 2008.
8. Breazeal. C., Aryananda, L. Recognition of affective communicative intent in robot-directed speech. Autonomous Robots, 12: 83–104, 2002.
9. Ambrus. D. C. Collecting and recording of an emotional speech database. Advances in speech technology, 239-244, 2000.
10. Rao, K. S., Koolagudi, S. G. Identification of Hindi dialects and emotions using spectral and prosodic features of speech. Systemics, Cybernetics, and Informatics, 9 (4): 24–33, 2011.
11. Vidrascu, L., Devillers, L. Real-life emotions in naturalistic data recorded in a medical call center," In 1st International Workshop on Emotion: Corpora for Research on Emotion and Affect, Genoa, Italy, pp. 20–24, 2006.
12. Meddeb, M Hichem, K., Alimi, A. Speech Emotion Recognition Based on Arabic Features. 15th international conference on Intelligent Systems design and Applications (ISDA15), Marrakesh, Morocco, 2015.
13. Agrima, A., Farchi, A., Elmazouzi, L. Mounir, I., Mounir, B. Emotion recognition from Moroccan dialect speech and energy band distribution. International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019.
14. Shahin, I., Nassif, A. B., Hamsa, S. Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network. IEEE Access, 7: 26777–26787, 2019.
15. Abdel-Hamid, L. Egyptian Arabic Speech Emotion Recognition using Prosodic, Spectral and Wavelet Features. Speech communication, 2020.
16. De Lope, J., Graña, M. An ongoing review of speech emotion recognition. Neurocomputing, 528: 1-11.
17. Mannepalli, K., Sastry, P. N., Suman, M. Analysis of Emotion Recognition System for Telugu Using Prosodic and Formant Features. Speech and Language Processing for Human-Machine Communications, India, pp. 137-144. 2017.
18. Saste, S. T., Jagdale, S. M. Emotion recognition from speech using MFCC and DWT for security system. International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2017.
19. Jacob, A. Speech emotion recognition based on minimal voice quality features. International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 2016.
20. Kwon, O., Chan, K., Hao, J., Lee, T. Emotion recognition by speech signals. In Eurospeech, Geneva, pp. 125–128, 2003.

21. Li, Y., Chao, L., Liu, Y., Bao, W., Tao, J. From simulated speech to natural speech, what are the robust features for emotion recognition? International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 2015.
22. Batline, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V., Amir, N. The automatic recognition of emotions in speech. Emotion-Oriented Systems, Springer, Berlin, Heidelberg , pp. 71–94, 2010.
23. Youddha Beer Singh , Shivani Goel. A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492: 245-263, 2022.
24. Boersma, P., Weenink, D. Praat, a system for doing phonetics by computer. Glot Int, 5(9), 341–345, 2002.
25. Ververidis, D., Kotropoulos, C. Emotional speech recognition: resources, features, and methods. Speech Communication, 48: 1162–1181, 2006.
26. Batliner, A., Schuller, B. The automatic recognition of emotions in speech. Emotion-Oriented Systems, Springer, Berlin, Heidelberg, pp. 71–94, 2010.
27. Zhang, G., Yin, J., Liu, Q., Yang, C. The Fixed-Point Optimization of Mel Frequency Cepstrum Coefficients for Speech Recognition. Proceedings of 2011 6th International Forum on Strategic Technology, Heilongjiang, Harbin, pp. 1172-1175, 2011.
28. Liu, Z, T., Xie, Q., Wu, M., Cao, W. H., Mei, Y, Mao, J. W. Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing, 309: 145-156, 2018.
29. Goodfellow, I.., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, 2016.
30. Swain, M., Sahoo, S., Routray, V. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. International Journal of Speech Technology, 18: 387–393, 2015.
31. Vogt, T., Andre, E. Improving automatic emotion recognition from speech via gender differentiation. Proceedings of Language Resources and Evaluation Conference, Genoa, Italy, pp. 1123-1126, 2006.
32. Verma D, Mukhopadhyay D, Mark E. Role of Gender Influence in Vocal Hindi Conversations: A Study on Speech Emotion Recognition. 2016 International Conference on Computing Communication Control and automation(ICCUBEA); Pune, India, pp.1-6, 2016
33. Singh V, Prasad S. Speech emotion recognition system using gender dependent convolution neural network. Procedia Computer Science, 218: 2533-2540, 2023.