Speech Emotion Recognition

Group member:
Tianheng Wang (N17192477)
Yuehan Cai (N17290298)

1. Introduction

Speech emotion recognition (SER) is an active topic in the human–computer interaction field in recent years. After years of research and application of speech recognition, machines can easily and effectively recognize the speech content and the speaker, but there are still difficulties for machine to detect speech emotions. Solving the SER difficulties will be very helpful for current human–computer interaction.

Since the emotion recognition is nothing but a pattern recognition system, two major problems are: what these patterns are and how these patterns are like. In this study, speech databases where emotions are well annotated were used for the first problem, while machine learning methods for classification are adopted for the second problem.

Previous works[1–5] have achieved SER accuracy variously from 60% to 90%, depending on different selections of speech emotion databases, audio features and machine learning methods.

The experiments were conducted on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[6] and Toronto Emotional Speech Set (TESS)[7]. Recordings in 2 sets are portraying each of 7 emotions: neutral, happy, sad, angry, fear, surprised and disgust. Both sets are in English. RAVDESS contains audio and video recordings of 24 actors (male and female) reading 2 sentences in 7 emotions and in two level of tension, normal and strong. TESS contains recordings of 2 female actors reading different words in 7 emotions. In this work, 1248 audio recordings were selected from RAVDESS and 1370 recordings were selected from TESS.

2. Speech Emotion Recognition System

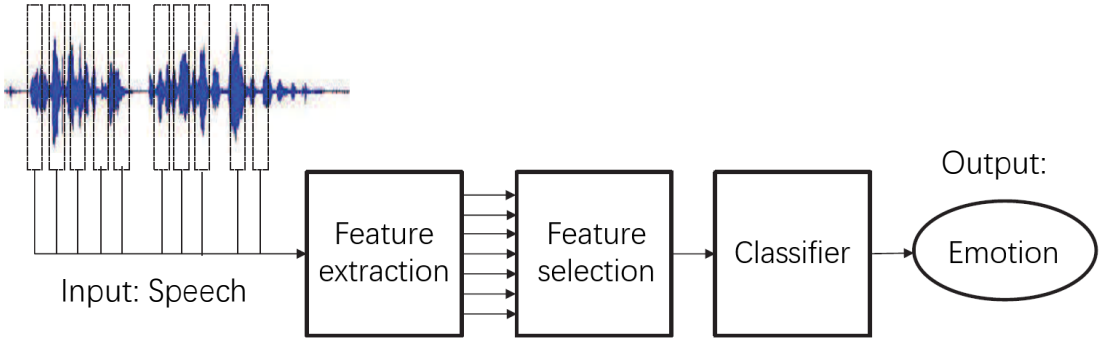

Figure 1. Framework of the SER System.

The framework of the SER system in this work is depicted in Figure 1. Like a typical pattern recognition system, it contains 5 modules: emotional speech input, feature extraction, feature selection, classification, and recognized emotion output.

2.1 Feature Extraction

A large number of parameters reflect the emotional characteristics. Research agrees that Mel-frequency cepstrum coefficients (MFCCs)[4] is a highly significant feature, but it seems that there are no more final conclusions other than this.

In this work, 5 short-term features were extracted from audio: pitch, energy, spectral centroid, MFCCs and zero-crossing rate (ZCR)[5]. Pitch was computed using unbiased autocorrelation function (ACF). Root-mean-square (RMS) of energy was finally used. 50 MFCCs were computed in total.

Other features, including formant frequencies, speaking rate, linear predictive coding (LPC) and linear prediction-based cepstral coefficients (LPCCs), are also suggested to be close related to speech emotions.

2.2 Feature Selection

Support vector machine (SVM) model was used as selection model. First, all features were inputted and the overall accuracy the model can achieve was observed. Then, features were removed one by one (except MFCCs, which were considered as the most important feature) to discover the effect it has on overall accuracy.

2.3 Classifiers

In this study, three main classifiers, SVM (Support Vector Machine), CNN (Convolutional Neutral Network) and LSTM-CNN (Long Short Term Memory-Recurrent Neutral Network), are implemented. Each of them had advantages and limitations.

SVM is a technique for data classification and regression by Vapnik and Chervonenkis derived from statistical learning theory in 1990s. Its main idea is to transform the original input set to a high-dimensional feature space kernel mapping, then achieve optimum classification in the new feature space. The original and classical use of SVM is in two-class classification, but Libsvm provide SVC model suitable for multi-class classification, which uses one versus one method traversing all classes.

Deep learning is an emerging field in machine learning in recent years, they have the characteristic to learn high level invariant features from raw data, which is useful in emotion recognition area. To prevent the mixture of unrelated features, 1-dimensional convolutional operator are used, which can only calculate along time. But this method will lose lot of information at each time point, we implement parallel structure with two operator calculates on 2 dimensions first and combine results together.

LSTM-RNN was originally introduced to solve backpropagated error in RNN when face to long-time sequence. LSTM make use of recurrently connected memory blocks, each of them connects to several memory cells along with three multiplicative ‘gate’ units: the input, output, and forget gates (Figure 2). This structure allows the network to store and access information over long period of time. In this work, emotion is a continuous event, we would like to know the change or continuity along time. LSTM layer in Keras gives an excellent model.

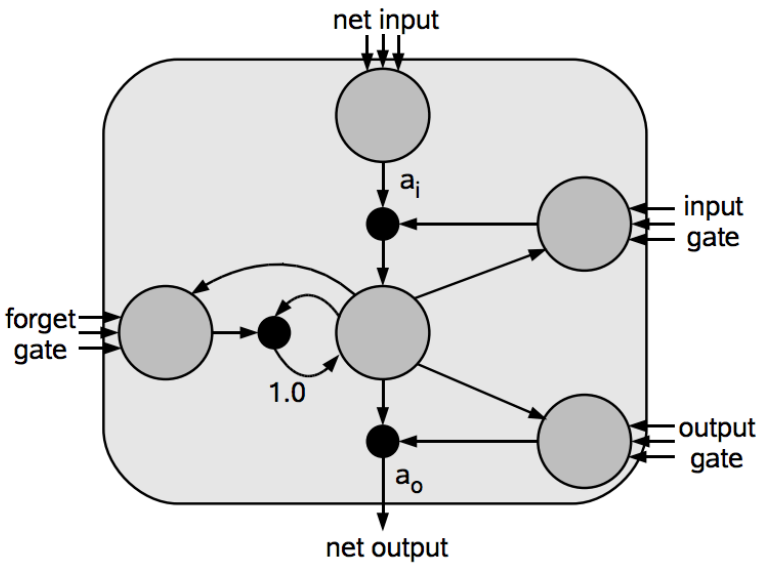

Figure 2. Structure of LSTM

3. Implementation of Feature Extraction

A series of pre-processing on two sets was conducted first, including unifying the sample rate, removing audio tags and converting few stereo recordings into mono in order that they can be easily and steadily read in the experiment. Recordings in RAVDESS were down-sampled at 24414 Hz to be consistent with recordings in TESS.

Feature extraction was implemented with both MATLAB API for Python and LibROSA[8]. LibROSA is a python package for audio analysis using together with NumPy and SciPy, two packages for scientific computing. In practice, MATLAB API was astonishingly nearly 100 times slower than LibROSA for extraction, due to huge delay and inefficiency of the communication between MATLAB engine instance and Python. Therefore, only LibROSA was used in following experiments.

The feature extraction part read all audio files under 7 emotions of both sets, extracted features and yielded a three-dimensional feature matrix. The feature matrix was then passed to classifiers together with a list of emotion label for each recording.

A Hann window was used to segment signal into frames, and features were computed for each window, resulting in a two-dimensional feature vector for each audio. Stacking of all feature vectors formed a three-dimensional matrix, i.e. the feature matrix of all recordings.

However, recordings have different lengths, so their feature vectors have different lengths and become unable to stack. In order to keep the feature matrix size fixed for all recordings, signal of every recording was trimmed or zero-padded to 43195 samples (1.8 s, which is around the median and mode length of all recordings) before feature extraction.

Since the timbre and patterns of emotions may be different for male and female, gender was also one-hot encoded as a “feature” row for each recording in the feature matrix.

4. Results

4.1 Feature Extraction

Combination of window length of 1024 samples (42 ms) and hop length of 768 samples (31 ms) was found to be little more effective for the SER system accuracy. In fact, window and hop lengths have little effect on the result, as long as the hop length does not exceed 2048 samples (84 ms).

4.2 Feature Selection

Removing ZCR, spectral centroid and RMS showed no obvious changes in overall accuracy. Therefore, features of MFCCs, pitch, energy and gender are selected to obtain best performance.

Validation on gender features was also done. The accuracy for female emotion recognition is around 63.2%, for male is around 64.15%, meanwhile with gender features accuracy is 59.26%, without gender features is 57.03%. This shows gender is also important and determine feature in emotion.

4.3 Model Performance Comparison

In our work, all the features were randomly split into train and test set, and normalization was applied on each audio.

In SVM model, we used loops to iterate and found the best model parameters. The overall accuracy was around 88% when using 100 MFCCs.

In CNN, parallel CNN and LSTM model, the overall accuracy was around 84%, 85% and 74% respectively.

For each model, a confusion matrix was used to evaluate the accuracy on each class.

acc	0	1	2	3	4	5	6
0	0.94	0	0	0	0.01	0.02	0.01
1	0.01	0.88	0.01	0	0.05	0	0
2	0.01	0	0.84	0.05	0.03	0.02	0.05
3	0.03	0.02	0.04	0.81	0.03	0.03	0.07
4	0.01	0.06	0.03	0.1	0.82	0	0
5	0.06	0	0.03	0.01	0	0.92	0
6	0.04	0.02	0.05	0.02	0.06	0.03	0.76

Table 1. Confusion Matrix of SVM

acc	0	1	2	3	4	5	6
0	0.93	0	0	0.01	0	0.01	0.04
1	0.02	0.94	0	0	0.03	0	0
2	0.02	0	0.81	0.06	0	0.05	0.07
3	0	0.08	0.03	0.81	0.01	0.03	0.08
4	0.04	0.12	0.04	0.05	0.76	0.02	0.01
5	0.03	0	0.01	0	0	0.97	0
6	0.07	0.39	0.03	0.01	0.05	0.05	0.76

Table 2. Confusion Matrix of CNN

acc	0	1	2	3	4	5	6
0	0.9	0	0.03	0.01	0.04	0	0.01
1	0.03	0.88	0.01	0	0.04	0	0
2	0.01	0	0.85	0.05	0	0.05	0.04
3	0.03	0.06	0.05	0.81	0.03	0.02	0.04
4	0.01	0.04	0.03	0.05	0.82	0.02	0.04
5	0.04	0	0.03	0.01	0	0.92	0.01
6	0	0.02	0.08	0	0.08	0.06	0.76

Table 3. Confusion Matrix of Parallel CNN

acc	0	1	2	3	4	5	6
0	0.77	0.06	0.01	0.02	0.05	0.05	0.03
1	0.04	0.85	0	0.04	0.03	0	0
2	0.06	0.02	0.81	0.02	0.03	0.02	0.05
3	0.01	0.09	0.08	0.76	0.04	0.05	0.01
4	0.14	0.17	0	0.08	0.61	0.01	0.05
5	0.03	0.02	0.06	0.1	0	0.78	0.04
6	0.04	0.1	0.06	0.07	0.05	0.03	0.65

Table 4. Confusion Matrix of LSTM

5. Conclusion

Feature selection and classifier implementation are most important part in audio emotion recognition system. In this work, we mixed two English database TESS and RAVDESS to analyze, train and test. MFCC, energy, pitch and gender are used as features to provide the same level prediction as well as achieve least calculation time, and SVM, Neutral Network achieves accuracy over 80%, and LSTM achieves over 70% accuracy, which is far away from the requirements of automatic emotion recognition in human-computer interaction.

There are three main problem for our work. Firstly, the recognition on separate or mixed gender still need more work, features like frequency, formant or spectral may help in recognize mixed gender speech. Second, dataset is limited, more various data is need for this topic, especially variety in gender, age and language. Third, feature selection part is not satisfying, features contains noise information that need to be purified or extracted further and deeper neutral network are need for analyzing these features.

References

[1] A. B. Ingale and D. S. Chaudhari, “Speech Emotion Recognition,” in *International Journal of Soft Computing and Engineering*, vol. 2, no. 1, Mar. 2012, pp. 235–238.

[2] K. Han and D. Yu, “Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine,” in *INTERSPEECH 2014*, Sep. 2014, pp. 223–227.

[3] D. Le and E. M. Provost, “Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks,” in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2013, pp. 216–221.

[4] Y.-L. Lin and G. Wei, “Speech emotion recognition based on HMM and SVM,” in *2005 International Conference on Machine Learning and Cybernetics*, Aug. 2005, pp. 4898–4901.

[5] T. Seehapoch, S. Wongthanavasu, “Speech emotion recognition using Support Vector Machines,” in *5th International Conference on Knowledge and Smart Technology (KST)*, Feb. 2013, pp. 86–91.

[6] S. R. Livingstone and F. A. Russo, “The Ryerson audio-visual database of emotional speech and song (RAVDESS),” in *Zenodo*, 2018. [Online]. Available: https://zenodo.org/record/1188976. [Accessed May 8, 2018].

[7] K. Dupuis and M. K. Pichora-Fuller, “Toronto emotional speech set (TESS),” in *TSpace Repository*, 2010. [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487. [Accessed May 8, 2018].

[8] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, O. Nietok, “librosa: audio and music signal analysis in Python,” in *Proceedings of the 14th Python in Science Conferences (SciPy)*, Jul. 2015, pp. 18–25.