ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 101-106
doi:10.18178/wcse.2019.03.018

Dataset for Depression Detection from Speech Emotion Recognition

Lwin Lwin Mar [1+], Win Pa Pa[2+] and Tin Lay Nwe[3 +]

[1, 2, 3] UCSY, Myanmar

Abstract. The recognition of internal emotional state of a person plays an important role in several human related fields. Emotions constitute an essential part of our existence as it exerts great influence on the physical and mental health of people. Depression is a common mental disorder. Developments in affective sensing technology with focus on acoustic features will potentially bring a change due to depressed patients' slow, hesitating, monotonous voice as remarkable characteristics. The system is intended for classification of emotions and depression by using speech signals. Both time and frequency domain features will be used in feature vector extraction. In feature extraction, the system will use wavelet transform and MFCC. DenseNet will be used to detect the emotion, classify the type of emotion and then depression. This paper will present about the datasets collected for the system and the experimental results on the dataset using Support Vector Machine.

Keywords: internal emotional state, feature vector extraction, wavelet transform, MFCC, Densenet, depression

1. Introduction

Emotion plays important role in human's daily life. It indicates the mental state of a person. Depression may also be detected from emotions. Speech emotion Recognition (SER) can be defined as the extraction of the emotional state of the speaker from his or her speech signal. The motivation of the system is that features play an important role in speech emotion recognition. As features, wavelet features and MFCC (Mel Frequency Cepstral Coefficient) features are used for the system. Different wavelet decomposition structures are used for feature vector extraction. Most of the signals in practice are time- domain signals in their raw format. The most distinguished information is hidden in the frequency content. Wavelet transform decomposes a signal into wavelets. In MFCC feature extraction, these cepstral vectors are given to pattern classifier for speech emotion recognition purpose. Convolutional networks can be substantially deeper, more accurate, and efficient to train if they contain shorter connections between layers close to the input and those close to the output. Dense Convolutional Network (DenseNet) connects each layer to every other layer in a feed-forward fashion [4]. The dataset that is used in the system is also important. Firstly, the dataset that will be used in the proposed system is experimented with different classifiers. In this paper, experimental results classified with Support Vector machine and MFCC feature extraction will be presented.

In the next sections, feature extraction used in the system, Densenet classifier, data preparation, and experimental results on the dataset with Support Vector Machine will be presented.

Author [1] proposed to perform classification of speech emotions in step-by-step manner using different feature subsets for every step. They applied the maximal efficiency feature selection criterion for composition of feature subsets in different classification levels. The multi-level organization of classification and features was tested experimentally in two emotions, three emotions, and four emotions recognition tasks and was compared with conventional feature combination techniques. Author [2] introduces a first approach to emotion recognition using RAMSES, the UPC's speech recognition system. The approach is based on standard speech recognition technology using hidden semi- continuous Markov models. Both the selection of

[+] Corresponding author.
E-mail address: [1]lwinlwinmar@ucsy.edu.mm [2]winpapa@ucsy.edu.mm [3]tlnma@i2r.a-star.edu.sg

low level features and the design of the recognition system are addressed. Results are given on speaker dependent emotion recognition using the Spanish corpus of INTERFACE Emotional Speech Synthesis Database. Author [3] proposed to recognize the human emotion through speech using Hidden Markov Model and Support Vector Machine. To recognize emotion through speech, various speech features were extracted. Based on these speech features, classification of the emotions has been done and the classification performance of Hidden Markov Model and Support Vector Machine is discussed.

2. Feature Extraction

A proper choice of feature vectors is one of the most important tasks. The feature type used in different approaches may be acoustic: duration, energy, pitch, spectrum, cepstrum (MFCC features), voice quality, wavelets or linguistic: bag of words (BOW), part of speech (POS), higher semantics (SEM) and varia (disuencies/non-verbals such as breathing or laughter).

2.1 Wavelet Transform

The commonly used tool for signal analysis is Fourier Transform, which breaks down a signal into constituent sinusoids of different frequencies [6]. Wavelet transform decomposes a signal into a set of basis functions (wavelets). Wavelets are obtained from a single prototype wavelet Ψ (t) called mother wavelet by dilations and shifting:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$$

where a is the scaling parameter and b is the shifting parameter. The Discrete Wavelet Transform (DWT) is computed by successive low-pass and high-pass filtering of the discrete time-domain signal. At each level, the high pass filter produces detail information (Di) while the low pass filter produces coarse approximations (Ai). The output of the filters is decimated in order to maintain orthogonality, halving the number of coefficients at each iteration. The approximations are filtered again at each decomposition step.

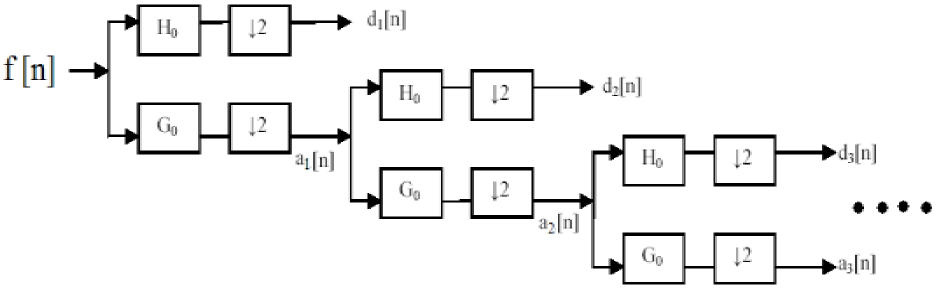

Fig. 1: Discrete Wavelet Transform

2.2. MFCC

MFCCs are the most widely used acoustic feature for speech recognition, speaker recognition, and audio classification. MFCCs take into account certain properties of the Human auditory system:

–Critical band frequency resolution approximately

– Log-power (dB magnitudes)

Speech is analysed over short analysis window. For each short analysis window, a spectrum is obtained using FFT. Spectrum is passed through Mel-Filters to obtain Mel- Spectrum. Cepstral analysis is performed on Mel-Spectrum to obtain Mel-Frequency Cepstral Coefficients. Thus speech is represented as a sequence of Cepstral vectors.

MFCC is most widely used spectral representation in ASR [10]. Pre-emphasis is boosting the energy in the high frequencies. The spectrum for voiced segments has more energy at lower frequencies than higher frequencies. This is called **spectral tilt**. Spectral tilt is caused by the nature of the glottal pulse. Boosting high- frequency energy gives more info to Acoustic Model. It improves phone recognition performance [5]. Speech is not a stationary signal; we want information about a small enough region that the spectral information is a useful cue.

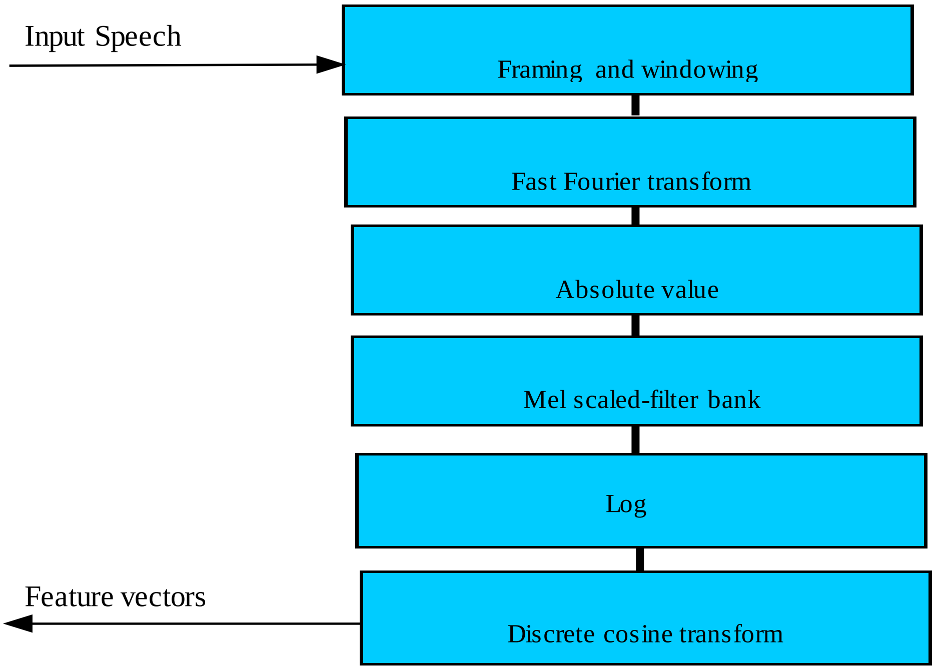

Fig. 2: Feature extraction using MFCC

3. DenseNet

As CNNs become increasingly deep, a new research problem emerges; it can vanish and “wash out” by the time it reaches the end of the network [8]. Dense Convolutional Network (DenseNet), which connects each layer to every other layer in a feed-forward fashion.Wheras traditional convolutional network with L layers have L connections---one between each layer and its subsequent layer---the network has L (L+1)/2 direct connections. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameter [6].

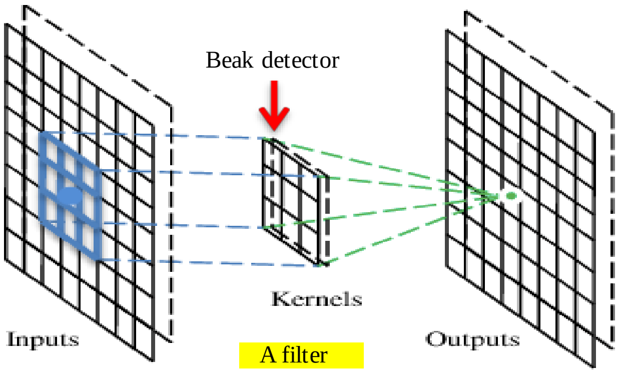

Fig. 3: Convolutional Layer

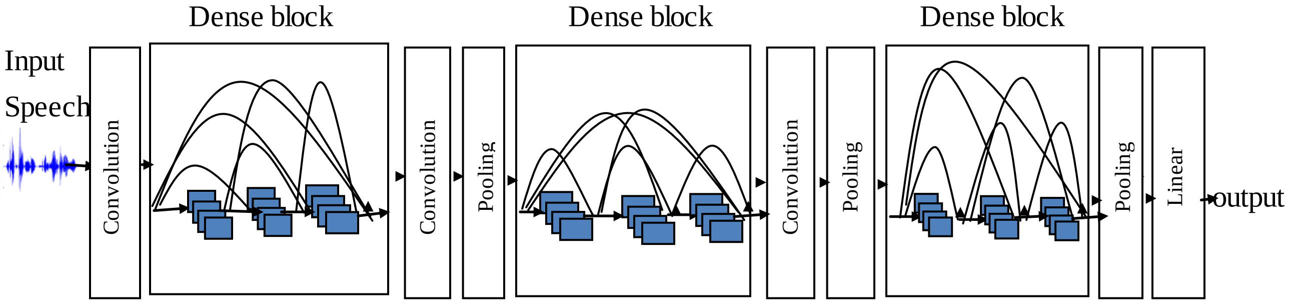

Fig. 4: Densenet classifier with 3 dense blocks

Figure 4 shows Densenet classifier. It connects all layers directly (with matching feature-map sizes).To preserve the feed-forward nature, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers. In this network, features are combined by concatenating them.L^{th} layer has l inputs consisting of the feature-maps of all preceding convolutional blocks. Its own feature-maps are passed on to all L−l subsequent layers. This introduces L (L+1)/2 connections in an L-layer network, instead of just L, as in traditional architectures. Because of its dense connectivity pattern, the approach is referred to as Dense Convolutional Network (DenseNet) [4].To classify the emotions from speech signal, the system intends to use Densenet.

4. Data Preparation

There will be about 10000 utterances for 10 hours movies. There are seven types of emotions. They are anger, disgust, fear, happy, sadness, and surprise, neutral and depression utterances. From these emotions, depression can be classified.

Table 1.Data Separation

Training	11000 utterances
Testing1	5000 utterances
Testing2(opened test)	2400 utterances

Angry, Disgust, Happy, Sad, Surprise, Fear, Neutral emotion and depression each contains about 2000 utterances.

The utterances are available from www.youtube.com/maharmovies.The data are collected from Myanmar movies of Mahar.They are speech utterances of different actors. Actors include both male and female. The speech utterances really describe the types of emotions.

5. Experimental Results

This is the result of experiment of dataset for depression detection from speech emotion recognition. The classifier used is Support Vector Machine and MFCC feature extraction is used for feature extraction. The training accuracy is 0.

Table 2.Testing results

	Accuracy	Precision
Closed test data	0.87	0.80
Opened test data	0.70	0.67
10 minutes Test data	0.86	0.87

Table3.Confusion matrix of 10 minutes test data

	Anger	Happy	Sad	Fear	Surprise	Neutral	Disgust	Depression
Anger	13	0	0	0	0	0	0	0
Happy	3	8	0	0	0	0	0	1
Sad	0	0	12	0	0	0	0	0
Fear	2	1	0	4	0	1	0	1
Surprise	0	0	0	0	12	0	0	0
Neutral	0	1	0	1	0	10	0	0
Disgust	0	0	0	0	0	0	14	0
Depression	0	0	0	0	0	2	0	10

The training accuracy with train-test split is 0.85 and testing accuracy is 0.84.

Table4.Confusion matrix using train-test split

	Anger	Happy	Sad	Fear	Surprise	Neutral	Disgust	Depression
Anger	44	0	0	0	1	0	0	0
Happy	3	21	0	0	0	0	0	0
Sad	0	0	24	0	0	2	0	0
Fear	0	0	0	49	0	1	0	1
Surprise	1	0	1	0	0	0	2	1
Neutral	2	0	2	2	2	4	1	0
Disgust	1	0	0	2	0	0	5	1
Depression	1	0	0	0	0	2	3	15

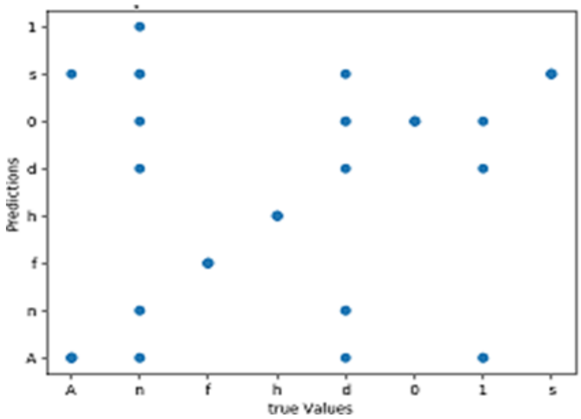

Fig .5: Using StratifiedShuffleSplit10-fold cross validation

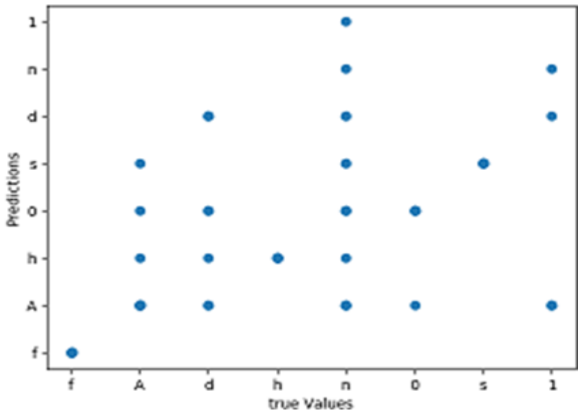

Fig.6: Using train_test split

6. Conclusion

Speech emotion recognition is quite new but a quickly growing field in the vast area of digital signal processing. Depression is a severe mental health disorder with high societal costs [9]. The proposed system will use wavelet features and MFCC features and densenet to classify the emotion and then depression is classified. The dataset that will be used in the proposed system is experimented with SVM classifier and it shows that the dataset produces good result with SVM.In the two methods of testing, Stratified ShuffleSplit gets better results than train_test split.

7. References

[1] Gintautus , TAMULEVIČIUS, Tatjana LIOGIENĖ, “Low-Order Multi-Level Features for Speech Emotion Recognition”, *Baltic J.Modern Computing*, Vol. 3(2015), No.4, 234-247

[2] Albino Nogueiras,Asuncion Moreno, Antonio Bonafonte,Jose B.Marino, “Speech emotion recognition using hidden Markov models”, *EUROSPEECH 2001 Scandinavia,7th European Conference on Speech Communication and Technology,2nd INTERSPEECH Event*, Aalborg, Denmark September 3-7, 2001

[3] Ashish B.Ingale, Dr.DS.Chaudhari, “Speech Emotion Recognition Using Hidden Markov Model And Support Vector Machine”, *International Journal of Advanced Engineering Research and Studies,IJAERS*/Vol.I/Issue III/April-June, 2012/316- 318

[4] Gao Huang, Zhuang Liu, Laurens van der Maaten, “Densely Connected Convolutional Networks”,*arXiv:1608.06993v5[cs.CV]* 28,Jan,2018

[5] Dan Jurafsky, “LSA 352 Speech Recognition and Synthesis”, *https:// nlp.stanford.edu / courses/lsa352 /* lsa352.lec6.6up.pdf

[6] “Multi-Scale/Multi-Resolution: Wavelet Transform”, *http://pami.uwaterloo.ca/~basir/SD575/*lectwk9-1.pdf

[7] Manish Chablani, “DenseNet : Towards Data Science”, *https://towardsdatascience.com/*densenet-2810936aeebb

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, “Densely Connected Convolutional Networks”,*Computer Vision Foundation (CVPR)* 2017

[9] Hailiang_Long, Xia Wu, Zhenghao_Guo, Jianhong_Liu and Bin Hu, “Detecting Depression in Speech: A Multi-

classifier System with Ensemble Pruning on Kappa-Error Diagram*",* *J Health Med Inform* 2017, Vol 8(5): 293

[10] "Mel Frequency Cepstral Coefficient (MFCC) tutorial", *http://practicalcryptography.com/miscellaneous /machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/*

[11] P.Vijayalakshmi, A.Anny Leema, "Real-time Speech Emotion Recognition Using Support Vector Machine", *International Journal ofSystem and Software Engineering,* Volume 2 Issue 1 June 2014

[12] Shanthi Therese S., Chelpa Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition", *International Journal ofScientific Engineering and Technology* Volume No.2, Issue No.6, pp: 479-484

[13] Shreya Narang, Ms.Divya Gupta, "Speech Feature Extraction Techniques: A Review", *International Journal ofComputer Science and Mobile Computing*, Vol.4, Issue.3, March 2015, Pg.107-114

[14] Panagiotis Tzirakis,Georage Trigeorgis,Mihalis A.Nicolaou, Member, IEEE,Bjorn W.Schuller,and Stefanos Zafeiriou, Member, IEEE, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks", *IEEE Journal of Selected Topics in Signal Processing*, Vol.11, No.8, December 2017