Depression Detection from Speech Emotion Recognition

Lwin Lwin Mar, Win Pa Pa
University of Computer Studies, Yangon
lwinlwinmar@ucsy.edu.mm, winpapa@ucsy.edu.mm

Abstract

The recognition ofthe 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. This paper will present classification of emotions and from it, depression is detected by using speech signals. Both time and frequency domain features will be used in feature vector extraction. In feature extraction, the paper will use wavelet transform and MFCC. DenseNet will be used to detect the emotion, classify the type of emotion and then depression.

***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. Depression is a mental health disorder with societal costs. Despite its high prevalence, its diagnostic rate is very low. To assist clinicians to better diagnose depression, researchers in recent years have been looking at the problem of automatic detection of depression from speech signals. Depression voice can change in the pitch, loudness and speaking rate. 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. Wavelet is best for non-stationary speech signal analysis. In MFCC feature extraction, these cepstral vectors are given to pattern classifiers 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.

In the next sections, feature extraction, Densenet, data preparation and experimental results are presented.

2. Related Work

Author **[2]** proposed to perform classification of speech emotions in step-by-step manner using different feature subsets for every step. Recognition of all emotions in one step is still a complicated process because of overlapping acoustic, prosodic and other features of the emotions. Each emotion is characterized by its own acoustic and prosodic features. Emotions, depending on the selected feature or features set, can be classified into various classes. 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 **[3]** 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. Short time pitch and energy, the contours of pitch and energy, the spectral shape, and duration and silence related measures are used as features for the system. Both the selection of 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 [4] 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. The energy related features, Mel-frequency cepstral coefficients (MFCC), fundamental frequency are some of the features which were used for the speech emotion recognition system. 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.

3. 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).

3.1. Wavelet Transform

Wavelet analysis is an exciting new method for solving difficult problems in mathematics, physics, and engineering with modern applications as diverse as wave propagation, data compression, signal processing, image processing, pattern recognition, computer graphics, the detection of aircraft and submarines and other medical image technology. Wavelets allow complex information such as music, speech, images, and patterns to be decomposed into elementary form at different positions and scales and subsequently reconstructed with high precision. The commonly used tool for signal analysis is Fourier Transform, which breaks down a signal into constituent sinusoids of different frequencies. 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:

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

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.

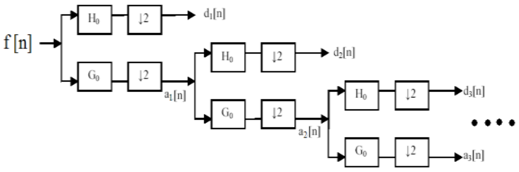

Figure 1.Discrete Wavelet Transform

3.2. 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 analyzed 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. 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.

Speech is not a stationary signal; we want information about a small enough region that the spectral information is a useful cue.

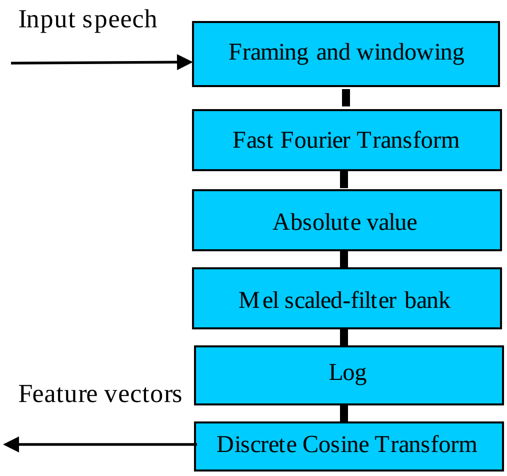

Figure 2.Feature extraction using MFCC

By combining the two feature extraction methods, the system is expected to be more effective and get the expected accuracy. The recognition rate of the speech recognition as well as speech emotion recognition system by using Mel-frequency cepstral coefficients (MFCC) is very good.

4. 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. Dense Convolutional Network (DenseNet), which connects each layer to every other layer in a feed-forward fashion. Whereas traditional convolutional networks with L layers have L connections—one between each layer and its subsequent layer—our 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 parameters.

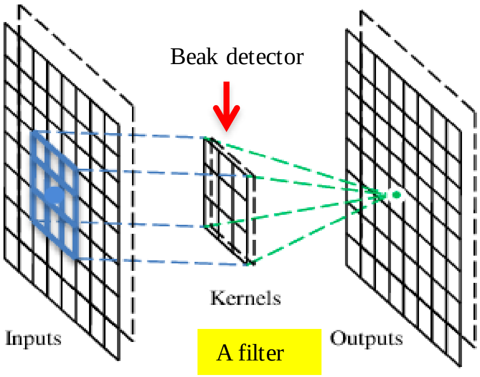

Figure 3.Convolutional Layer

Figure 3. shows convolutional layer in CNN.A convolutional layer has a number of filters that does convolutional operation.

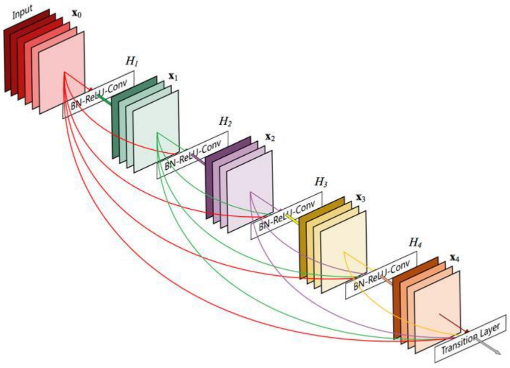

Figure 4. Five layers of a DenseNet block with a growth rate of 4 feature-maps per layer

In figure 4, 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).

5. Data Preparation

There are over 10000 utterances for about 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
Testing	5000 utterances

Angry emotion: 2000 utterances
Disgust emotion: 2000 utterances
Happy emotion: 2000 utterances
Sad emotion: 2000 utterances
Surprise emotion: 2000 utterances
Fear emotion: 2000 utterances

Neutral emotion: 2000 utterances

Depression: 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.

They are example sentences of data corpus.

Angry001: မင်းပါးစပ်ပိတ်စမ်း။

Angry002: မပြောနဲ့အမေရာ။

Angry003: ငါ့ကိုကောင်မလို့ မခေါ်နဲ့နော် အယုတ်တမာအရိုင်းအစိုင်းကောင်။

Disgust001: ဒုက္ခတွေတင်ပြမနေနဲ့ကွာ မင်းကလည်း အခိကက လ္ဘက်ရည်သောက်ရဖို့ အရေးကြီးတယ်။

Disgust002: ကဲကဲတော်ကြပါတော့ကွာ ကိုယ်နဲ့ ဘာဆိုင်လို့လဲ။

Disgust003: သြော် နေစမ်းပါဦး ချင်ချင်ရယ် လူကြီးတွေ စကားပြောနေတာ။

Happy001: နင် ရော ငါရော အောင်တယ်။

Happy002: ငါ့သားစာမေးပွဲအောင်တဲ့ အထိန်းအမှတ်နဲ့ ဒီနေ့ ဒေါ်ရွှေစင်ရဲ့ အကြော် ဆိုင် ပိတ်မည်။

Happy003: ဝမ်းသာလိုက်တာ ငါတို့အောင်တယ်။

Sadness001: ညီမလေး အိမ်ပေါ်ကဆင်းမယ်။

Sadness002: အဲဒီအတွက် ကျမ ရှင့််ကိုမုန်းတယ်။

Sadness003: သမီးတို့ ချစ်ချင်းကို မခွဲပါနဲ့ အဖေရယ်။

Surprise001: ရှင် ကိုကျော်အောင် မရှိဘူးရှင့်် မဖြစ်နိုင်တာ။

Surprise002: ဘာ သိန်း၂၀ ဟုတ်လား။

Surprise003: သြော် အန်တီတို့ အဖြစ်က ဒီလိုကိုး။

Fear001:ငါတောင်းပန်ပါတယ်ကွာ ငါတောင်းပန်ပါတယ် ကွာ။

Fear002: မလုပ်ပါနဲ့ ရှင် မလုပ်ပါနဲ့။

Fear003: ဟို ရေချမ်းအိုးဆိုတာ ကျမပါရှင်။

Neutral001: သိပါတယ် မေကြီးရ မေကြီးကို ချစ်လို့စတာ။

Neutral002: ဘာမှမဟုတ်ပါဘူး အမေရာ။

Neutral003: ကဲ အားလုံးမှာ ကြိုက်သလောက်စား။

Depression001: လူတွေကခက်သားလား အမှန်ကို အမှား ထင်နေကြတာ။

Depression002: ဘာမှမဖြစ်ပါဘူး သူတို့ခေါ်လာလို့ပေါ့ ဆရာ။

Depression003:အမှန်တော့လူတွေအားလုံး ယုတ်မာ နေကြတာ။

6. Experimental Results

There is result in that accuracy is 0.64%.The classifier used is Support Vector Machine (svm) classifier and MFCC feature extraction. The test results of 10 minutes and 80 minutes utterances using 10-folds cross-validation are shown in table 2.

Table 2. Testing Results

Test data	Opened test	Closed test
80minutes utterances	0.70	0.87
10minutes utterances	0.47	0.64

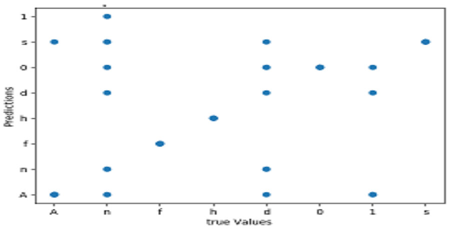

Figure 5: Results of 10 minutes utterances

7. 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. The proposed system will use wavelet features and MFCC features and densenet to classify the emotion and then depression is classified. Experimental results of SVM show the good result using the data corpora prepared.

References

[1]Tin Lay Nwe,Say Wei Foo,Liyanage C.De Silva, “Speech emotion recognition using hidden Markov models”, *Elsevier Speech Communications Journal* Vol.41,Issue 4,pp.603-623,November 2003

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

[3]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

[4]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

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

[6]Kun Han, Dong Yu, Ivan Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine: INTERSPEECH 2014

[7]Kishori R.Ghule, R.R. Deshmukh, "Feature Extraction Techniques for Speech Recognition: A Review", *International Journal of Scientific & Engineering Research*, Volume 6, Issue 5, May-2015

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

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

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

[11]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

[12]Aharon Satt, Shai Rozenberg,Ron Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms", INTERSPEECH 2017,August 20-24,2017,Stockholm,Sweden

[13]Dipti D.Joshi Prof.M.B.Zalte, "Speech Emotion Recognition: A Review", *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)* ISSN: 2278-2834, ISBN: 2278-8735.Volume 4, Issue 4(Jan.-Feb.2013), PP 34-37

[14]Aastha Joshi,Rajneet Kaur, "A Study of Speech Emotion Recognition Methods", *International Journal of Computer Science and Mobile Computing IJCSMC*,Vol.2,Issue.4,April2013 ,pg.28-31

[15]Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, "Speech Emotion Recognition: Methods and Cases Study", 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)-Volume2, pages175-182

[16]Rani P.Gadhe, Shalkh Niofer R.A.V.B. Waghmare, P.P.Shrishimal, R.R.Deshmukh, "Emotion Recognition from Speech: A Survey", *International Journal of Scientific & Engineering Research*, Volume 6, Issue 4, April-2015