IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 4, No. 3
August 2015

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

Int. J. Engg. Res. & Sci. & Tech. 2015

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 4, No. 3, August 2015

Research Paper

EMOTION SPEECH RECOGNITION USING MFCC AND RESIDUAL PHASE IN ARTIFICIAL NEURAL NETWORK

S Saranya Swaminathan[1]* and T Jayasankar[1]

Corresponding Author: **S Saranya Swaminathan ✉ Saranya20swaminathan @gmail.com*

The main objective of this paper is to develop a speech emotion recognition system using residual phase and MFCC features with neural network. The speech emotion recognition system classifies the speech emotion into predefined categories such as anger, fear, happy, neutral or sad. The proposed technique for speech emotion recognition has two phases: Feature extraction, and Classification. Initially, speech signal is given to feature extraction phase to extract residual phase and MFCC features. Based on the feature vectors extracted from the training data, neural network are trained to classify the emotions into anger, fear, happy, neutral or sad. Using residual phase and MFCC features the performance of the proposed technique is evaluated. The rate of recognizing the emotion from the speech signal is about 88.3%. Experimental results show neural network, Classifier is better which offers a new efficient way of solving problems.

Keywords: Speech emotion recognition, Residual phase, MFCC, Neural network classifier, Feature extraction

INTRODUCTION

Emotions play an important role in expressing our feeling. In addition, speakers generally have own speaker dependent style, i.e., a characteristic articulation rate, intonation habit and loudness characteristic. So, emotions that are expressed and inferred in speech depend upon the speaker's community culture and language, gender, age, education, social status, health, physical engagements, etc. In many human computer interactions where speech is used as an interface require the detection of emotion in the speech. But a very few human machine interfaces being implemented currently are able to achieve that. Hence, the requirement of building an interface for machine and human communication is highly anticipated that can detect emotions effectively and efficiently. Emotions expressed in a speech can be recognized using three factors – the speech contents, the speaker's face expressions and the features extracted from the emotional speech. This paper is an attempt to present the basic approach to recognize emotion from a given speech signal. We discuss about the database, feature extraction techniques and

[1] Department of ECE, University college of Engineering (BIT Campus), Trichy, India.

classifiers are discussed in this paper. It has been noticed over the years that the recognition of emotion is growing much better due to development in speech analysis, and machine learning techniques. However recognition of emotion is still an interesting area and a lot of development has to be done. Various approaches for emotion recognition have been reported in the literature. Mixed modeling of human's emotional states from speech and visual information has been used as well. A B Kandali and his team in their work on Assamese language found that the choice of delta-Log-energy and delta-delta Log-energy features do not improve the recognition score. So, the feature set they choose for further study consisted of 1 Log-energy, 14 MFCC, 14 delta-MFCC, and 14 delta-delta-MFCC. After exhaustive trials of experiment with GMM classifier for M[6, 12], the highest mean classification score rose to 74.4%. They also found that the surprise emotion is the most difficult one to disambiguate from other emotions, since surprise may be expressed along with any other emotion such as angry-surprise, fear-surprise, happy-surprise, etc. Perusing used eight features chosen by a feature selection algorithm developed a real time emotion recognizer for call center applications. He achieved 77% classification accuracy in two emotions ("agitation" and "calm") using neural networks. Joy Nicholson and his team gained an accuracy rate of approximately 50% while developing a speaker and context independent system for recognition of emotion in speech using neural networks. They have designed and implemented a one-class-in-one network for emotion recognition. Y Pan found that the system that uses both spectral and prosodic features for recognition of emotion achieves better recognition rate than that only uses spectrum or prosodic features. More distinctly they have showed that, the recognition rate of the system which uses the combination of features MFCC, LPCC, MEDC, energy and pitch is a bit less than the systems that only use energy, pitch MFCC and MEDC features. Thus there has been considerable past work in the area of characterizing and detecting emotion in speech. However the problem of recognition of emotion is not fully solved yet and a lot of future work is left in the area. The emotion recognition systems consist of the basic steps as follows: collection of a correct database, extracting features from the database and finally classify the different emotions using different classifiers available. There are a few speech databases available from the recent works. There are six available databases which we can refer: two of them are publicly available, namely the Danish Emotional Speech corpus (DES) and Berlin Emotional Database (EMO-DB), and the remaining four databases are from the Interface project with Spanish, Slovenian, French and English emotional speech. All of these databases contain acted emotional speech. The speech database is of two types. Acted and real life data. As the name suggests itself, in an acted emotional speech corpus, a professional actor is asked to speak in a certain emotion, whereas in real databases, speech databases for each emotion are obtained by recording conversations in real-life situations such as call centers and talk shows. But it has been observed that there is a difference in the features of acted and real emotional speeches. This is because acted emotions are not felt while speaking and thus come out more strongly. However collecting real life data for emotion is not easy and it's also not helpful in training and testing due to noise and other disturbances.

Figure 1: The Audio Signal Intended For The Analysis in This Case

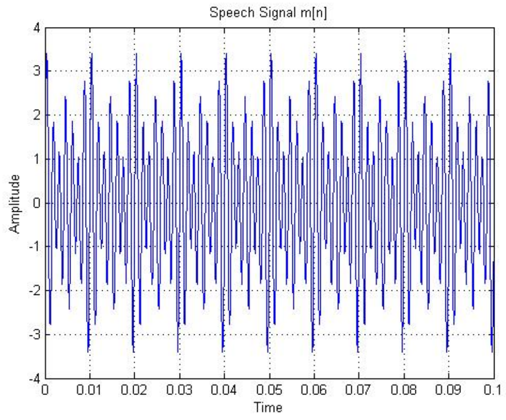
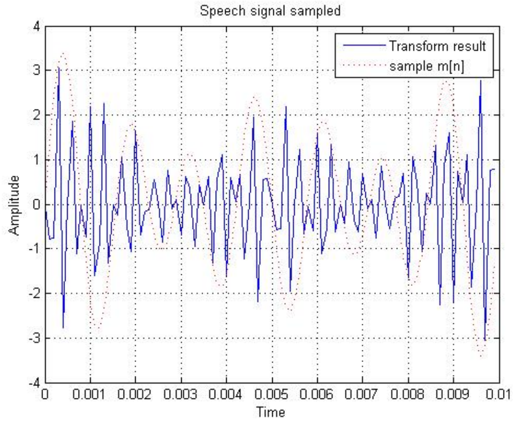


Figure 1.1: Sampled Speech Signal



There are thus a large number of speech processing techniques whose performance can be dramatically affected if polarity is inverted. In parallel, these same techniques are developed to automate systems which are nowadays expected to work properly on huge amounts of data. These data can be acquired through a variety of microphones, and consequently with different polarities. Detecting correctly the speech polarity is then a necessary step to ensure the good behavior of the aforementioned techniques. In this paper, a new simple algorithm for automatic speech polarity detection is proposed. This method, called Residual phase extraction, is based on the skewness of two excitation signals: the traditional residual signal, and a rough approximation of the glottal flow derivative. Contrarily to state-of-the-art approaches, it has the advantage of not requiring voicing decisions or F0 estimation. It therefore allows a fast computation, and is shown through our experiments to provide the best results in clean conditions as well as in noisy environments.

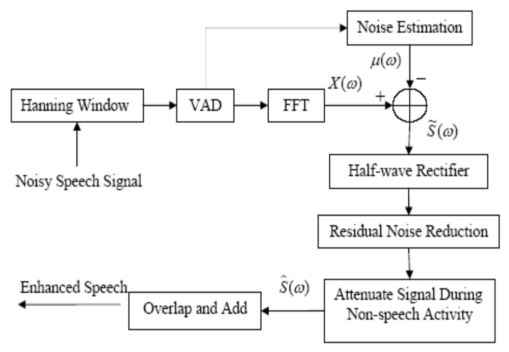
RELATED WORK

Speech signal contains large number of information from which we can detect the speaker, speech and emotion. There are several techniques available for extracting features from speech signal. Once the database is ready we can proceed towards extracting useful features from the recorded speech signals. A lot of commonly used features of speech signal have been extracted in the literature, i.e. the rate of speech, the energy features, pitch of the speech, etc. Spectrum features such as Linear Prediction Coefficient (LPC), Linear Prediction Cestrum Coefficient (LPCC), Mel Frequency Cepstrum Coefficients (MFCC) and the 1st order derivatives of them. Prosodic features: Speech power (P), Pitch (p) and phonetic features12 LPC parameters (c1, c2... c12) – Delta LPC parameter (d) are also used. Some of the common features are: Energy and related features, The pitch and other features, Qualitative Features. The Energy is one of the basic and most important features in speech signal. Statistics of energy in the whole speech sample can be obtained from computing energy, in the form of max and mean value, the variance and the range of variation and the energy contour. The frequency of the pitch can be computed in all of the speech frames individually and then the pitch statistics of the complete sample can be obtained. The values of pitch calculated give the global properties of the characteristic parameters. Emotional contents of an utterance highly relates to the voice quality of the utterance. The quality of the voice can be represented numerically by using the parameters estimated directly from a speech signal. The acoustic parameters which are related to the quality of speech are as given: (a) level of voice: amplitude of signal, the energy and the duration of the speech are shown as reliable measures; (b) pitch of the voice; (c) phrases, phonemes, words and feature boundaries; and (d) temporal structures.

RESIDUAL PHASE METHOD

The speech production and perception theory indicate that the source contains speaker information and also more robust due to its impulsive nature. Motivated us to explore methods for modeling the speaker-specific source information from the complete source. These attempts demonstrate that the source indeed significant speaker information. However, the recognition performance is far better than the vocal tract information. The reason may be that the methods employed in representing source information may not model all aspects of speaker information. By that we mean, LPCC or MFCC captures formants and their bandwidth information characterizing the vocal tract completely, pitch is only one aspect of speaker information due to source. Thus to improve the performance of source features, methods need to developed that tries to capture the complete source information. Source information contained in the LP residual of the speech signal. The LP residual can be processed in time, frequency, Cepstral or time – frequency domains to extract and model information. Processing of LP residual in time-domain has the advantage that the artifacts of digital signal processing like block processing or windowing (see Figure 2) effect that creeps in other domains of processing like will be negligible. Thus processing LP residual in time-domain is expected to model the speaker information in the best possible manner. Much work is not done on processing of LP residual at different levels. A unified frame work may be evolved where a given LP residual is processed at sub-segmental, segmental and supra-segmental levels.

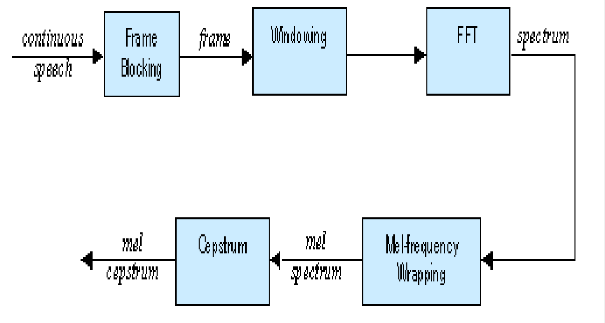
Figure 2: Block Diagram of Training Phase



MEL-FREQUENCY CEPSTRUM COEFFICIENTS PROCESSOR

A block diagram of the structure of an MFCC processor is given in Figure 3. The speech input is typically recorded at a sampling rate above 10,000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.

Figure 3: Block Diagram of the MFCC Processor



Frame Blocking

In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal

Typically the Hamming window is used, which has the form:

Fast Fourier Transform (FFT)

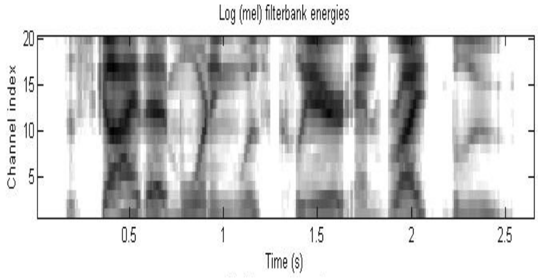
The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {xn}, as follow:

In general Xk's are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence {Xk} is interpreted as follow: positive frequencies $0 \le f < F_s / 2$ correspond to values $0 \le n \le N/2-1$, while negative frequencies $-F_s/2 < f < 0$ Here, Fs denote the sampling frequency. The result after this step is often referred to as spectrum or periodogram.

Mel-frequency Wrapping

As mentioned above, psychophysical studies have Figure shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

Figure 3.2: Log Energy Computation



One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel-scale (see Figure 3.1). That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, K, is typically chosen as 20. Note that this filter bank is applied in the frequency domain, thus it simply amounts to applying the triangle-shape windows as in the Figure 4 to the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those mel power spectrum coefficients that are the result of the last step are $\tilde{S}_0, k = 0,2,...,K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right],$$

$$n = 0, 1, \ldots, K\text{-}1$$

Note that we exclude the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal, which carried little speaker specific information. By applying the procedure described above, for each speech frame of around 30 ms with overlap, a set of mel-frequency cepstrum coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a mel-frequency scale. This set of coefficients is called an acoustic vector. Therefore each input utterance is transformed into a sequence of acoustic vectors.

NEURAL NETWORK MODEL (NN)

Neural Networks is one of the most advanced classifiers in the testing category. The neural has a feed forward method. The feed forward method takes one input as a training sample and another input as the target sample. The input sample is the data stored in the database on the basis of all the features which have been extracted at the time of training.

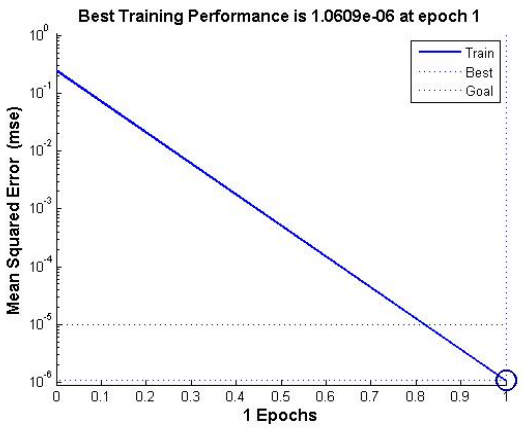
If P is the input sample then P would be defined as

P = sum (all. features (input));

Here the central block is termed as the Neural classifier. There are two input samples where the neural classifier generates the weight accordingly for the first input which has been taken from the database. The second input is the target set which is to be tested. The neural takes each sample as a neuron and explains to the architecture that how the input is going to react and how the result is going to be proceeded. Finally, it produces a binary result. If we do proceed more than one sample category, the neural will have be combined.

The below diagram shows the entire working of the testing of the speech files with the help of the MFCC and neural network Classifier. In this method, first of all, feature extraction has been done using MFCC algorithm and it is saved to database. Later on, at the time of testing a speech file is uploaded to test the scenario. The same process is applied to the uploaded file also and it has been passed to neural network. On the basis of the trained set data, the neural network predicts the output of the system (see Figure 4).

Figure 4: Neural Network Training Performance



During neural network training, the weights of the network are adjusted to minimize the mean square error obtained for each feature vector. If the adjustment of weights is done for all feature vectors once, then the network is said to be trained for one epoch. For successive epochs, the mean square error averaged over all feature vectors. During testing phase, the features extracted from the test data are given to the trained neural network model to find its match.

CONCLUSION

Although none of the approaches proved to be good enough for practical purposes with the present extent of development, they were good enough to prove that translating speech into trajectories in a feature space works for recognition purposes. Speech emotion could be useful in speech understanding, recommendation, retrieval and some other related applications. In this project, focus on challenging issue of recognizing speech emotions such as happy, sad, anger, fear, and neutral. Speech database collected from linguistics laboratory. Mel Frequency Cepstral Coefficient (MFCC) features are extracted from the speech signal. The neural network model classifier used to recognize the emotion from the features taken. The rate of recognizing the emotion from the speech signal is about 88.3%. Experimental results show neural network classifier is better which offers a new efficient way of solving problems. Finally, the new approach developed for training the neural network's architecture proved to be simple and very efficient. It reduced considerably the amount of calculations needed for finding the correct set of parameters. If the traditional approach had been used instead, the amount of calculations would have been higher.

FUTURE WORK

In future work, we will improve current model to increase the efficiency, besides this, the research on recognition of emotion intensity will be performed through the analysis of audio features, which is different from the approach in. Also, we notice that some new dimensional reduction and the pattern classification methods like tensor based analysis proposed recently; we will carry out study on its application in emotion recognition field.

REFERENCES

1. Furui S (1989), "Digital Speech Processing, Synthesis And Recognition," Marcel Dekker Inc.

2. Hasegawa H and Inazumi I (1993), "Speech Recognition By Dynamic Recurrent Neural Networks," Proceedings Of 1993 International Joint Conference On Neural Networks.

3. Jamshidi M (1983), "Large-Scale Systems: Modelling And Control," North-Holland.

4. Jankowski C R Jr., Vo H H and Lippmann R P (1995), "A Comparison Of Signal Processing Front Ends For Automatic Word Recognition," *IEEE Transactions On Speech And Audio Processing*, Vol. 3, No. 4.

5. Lee K-F, Hon H W and Reddy R (1990), "An Overview Of The Sphinx Speech Recognition System," *IEEE Transactions On Acoustic, Speech, And Signal Processing*, Vol. 38, No. 1, January.

6. Negroponte N (1995), "Being Digital," *Vintage Books.*

7. Pinker S (1995), "The Language Instinct," *Harperperennial.*

8. Prator C H and Robinett B W (1985), "Manual of American English Pronunciation," Harcourt, Brace and Co.

9. Rabiner L and Juang G (1993), "Fundamentals of Speech Recognition," Prentice-Hall.

10. Tebelskis J (1995), "Speech Recognition Using Neural Networks," Phd Dissertation, Carnegie ellon University.

11. Torkkola K And M. Kokkonen (1991), "Using The Topology-Preserving Properties Of Soms In Speech Recognition," Proceedings Of The Ieee Icassp.

12. Zegers P (1992), *"Reconocimiento De Voz Utilizando Redes Neuronales,"* Engineer Thesis, Pontificia Universidad Católica De Chile.

IJERST

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com

ISSN 2319-5991
03
9 772319 599001