Comparison between Different Feature Extraction Techniques to Identify the Emotion 'Anger' in Speech

Bageshree V. Pathak[1] and Ashish R. Panat[2]

[1] Cummins College of Engg for Women, Pune, India
bvpathak100@yahoo.com
[2] Priyadarshani Indira College of Engg., Nagpur, India
ashishpanat@gmail.com

Abstract. In this paper, three different techniques of feature extraction for identification of emotion in speech have been compared. Traditional feature like LPCC (Linear Predictive Cepstral Coefficient) and MFCC (Mel Frequency Cepstral Coefficient) have been described. Linear features like LFPC which is FFT based have been explained. Finally TEO (Teager Energy Operator) based nonlinear LFPC features in both time and freqnency domain have been proposed and the performance of the proposed system is compared with the traditional features. The comparison of each approach is performed using SUSAS (Speech Under Simulated and Acid Stress) and ESMBS (Emotional Speech of Mandarin and Burmese Speakers) databases. It is observed that proposed system outperforms the traditional systems. Analysis will be carried for identification mainly of the emotion 'Anger' in this paper.

Keywords: feature vector, VQ, MFCC, cepstrum, HMM, TEO, LPCC, LFPC.

1 Introduction

Speech is the vocalized form of human communication. Each spoken word is created out of the phonetic combination of a vowel and consonant speech units [9]. In order to obtain good representation of speaker characteristics, speech data needs to be analysed using suitable analysis technique. In the analysis technique proper frame size is selected and extracting the relevant features is carried out in the feature extraction stage.

A number of studies have been conducted to investigate acoustic indicators to detect emotion in speech. The characteristics most often considered include Fundamental frequency F0, duration [2], [17],intensity [l], spectral variation [2], [5] and wavelet based subband features [6]. In these researches, features used are mostly derived from linear speech models. However, in recent years, non-linear features derived from Teager Energy Operators (TEO) [7], [8] are explored.

Human auditory system is assumed to have a filtering system in which the entire audible frequency range is partitioned into frequency bands [9]. According to Fletcher [10], speech sounds are pre-processed by the peripheral auditory system through a bank of bandpass filters. These auditory filters perform the process of frequency weighing for frequency selectivity of ear.

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 637–643, 2012.

2 Pre Processing

Initially, the acoustic wave is transformed into a digital signal, which is suitable for voice processing. A microphone or telephone handset converts the acoustic wave into an analog signal [12]. This analog signal is conditioned with antialiasing filtering to compensate for any channel impairments. Before sampling, the antialiasing filter limits the bandwidth of the signal to approximately the Nyquist rate. [12]. The conditioned analog signal is then sampled to form a digital signal by an analog-to-digital (A/D) converter [17]. Today's A/D converters for speech applications typically sample with 12–16 bits of resolution at 8000–20000 samples per second. The speech is further subjected to windowing by passing it through a Hamming window and a frame size of 10 to 30 m sec is chosen for analysis .

3 Feature Extraction

The speech signal can be represented by a sequence of feature vectors. In this section, the selection of appropriate features is discussed. This is known as feature selection. There are a number of feature extraction techniques based on speaker dependent parameters like Pitch, Formants, Energy, Intensity, LPC etc.

3.1 Traditional Features

MFCC [15] and LPCC are the most widely used feature extraction techniques. In sound processing, the Mel-Frequency Cepstrum (MFC) in sound signal processing is basically a representation of the short-term power spectrum of the sound signal. By taking linear cosine transform of a log power spectrum on a nonlinear scale i.e. mel scale frequency we get MFCC [3]. Mel-frequency cepstral coefficients (MFCCs) are derived from a type of cepstral representation of the audio clip. This frequency warping can allow for better representation of sound and act a distinctive feature for every speaker. The following equation is used to compute the Mel for given frequency f in HZ:

$$F\,(\text{Mel}) = [2595 * \log 10[1+ f /700]$$

LPCC yield better results than MFCC while discriminating different languages. It also shows that language identification performance may be increased by encompassing temporal information by including acceleration features. Gaussian Mixture Model (GMM) is normally used along with these techniques for modeling of the speech signal [15].

3.2 Linear Features

The Linear parameters that are generally considered in evaluating changes in speech signal characteristics are intensity, duration, pitch, spectrum of vocal tract, glottal source effect and vocal tract articulator profiles. The last two parameters cannot be derived directly from the speech signal but require measurements directly related to the speaker which restricts the flexibility [8].

Intensity: In general the average intensity observed increases with emotions like anger or some type of high workload. It was also found that mainly vowels and semivowels show a significant increase in intensity while consonants do not.

Pitch: Pitch is the most widely considered parameter of emotion evaluation. Pitch contours, variance and distributions show variations when speech is subjected to emotions as shown in figure 1.

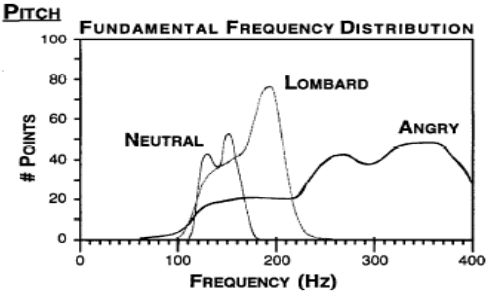

Fig. 1. Variation of Pitch, for Neutral and Angry emotions

FFT based Log-Frequency Power Coefficients (LFPC) are designed to simulate logarithmic filtering characteristics of human auditory system by measuring spectral energies [13]. First, the signal is segmented into short-time windows which are 16ms for emotion speech samples. The reason for using short frame length for emotion database is that it includes female speech utterances which have shorter pitch period than male speech and frame size needs to cover two pitch period of fundamental frequency. The window is moved with the frame rate 9ms for emotion speech samples. The frequency content is calculated in each frame using Fast Fourier Transform (FFT) method. This power spectrum is accumulated into a bank of log-frequency filters. The filter bank splits input speech signal into multiple outputs by passing through the parallel set of band pass filters. Energy in the filter bank output is calculated. For each speech frame 12 LFPCs are obtained.

3.3 Non-linear Features

An alternative way to characterize speech production is to model the airflow pattern in the vocal tract. The underlying concept here is that while the vocal tract articulators do move to configure the vocal tract shape, it is the resulting airflow properties which serve to excite those models which a listener will perceive as a particular phoneme. Studies by Teager emphasized this approach with further investigations by Kaiser to support those concepts [7]. In an effort to reflect the instantaneous energy of nonlinear vortex-flow interactions, Teager developed an energy operator. TEO (Teager Energy Operator) is a non-linear differential operator which detects

modulations in the speech signal and further decomposes the signal into AM and FM components. Using the shape of a pitch normalized TEO profile, good performance can be obtained for speech produced under angry, loud, clear, and Lombard effect speaking conditions. The features relating to spectral shape should be incorporated into Teager Energy Operation as well. For this reason, TEO based nonlinear properties in combination with the LFPC are investigated, which is commonly applied in the time domain [7],[8]. In this paper, TEO in both time and frequency domain are considered [13]. In Time Domain LFPC (NTDLFPC) the speech signal is windowed and passed through TEO and then FFT and LFPC are applied to the signal. In and Nonlinear Frequency Domain LFPC (NFDLFPC), the speech signal is windowed and converted to frequency domain by using FFT and then applied to TEO and finally LFPC are extracted as shown in figure2 below.[13]

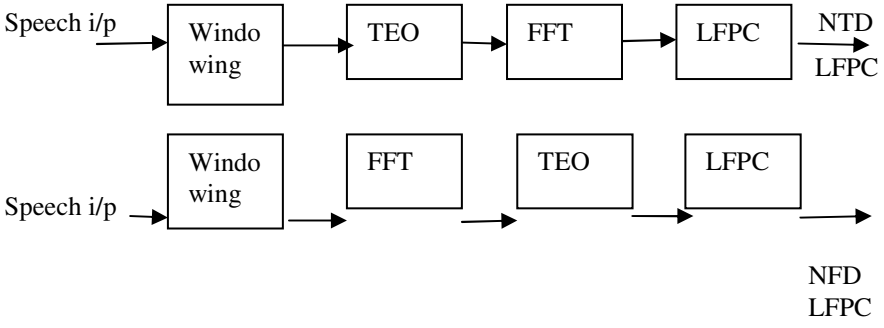

Fig. 2. TEO in time and frequency domain

4 Emotion Database

For emotion classification, a database ESMBS which is specifically designed and set up for text-independent emotion classification is used [14]. The database includes short utterances covering the six emotions, namely Anger, Disgust, Fear, Joy, Sadness and Surprise. A total of six native Burmese language speakers (3 males and 3 females), six native Mandarin speakers (3males and 3females) are employed to generate 720 utterances. Sixty different utterances, ten each for each emotional mode, are recorded for each speaker. The recording is done in a quiet environment using a mouthpiece microphone.

SUSAS (Speech Under Simulated & Actual Stress Database) [13], [14] was established in order to conduct research into the analysis and recognition of speech produced in noise and under stress. SUSAS consists of a wide variety of stresses and emotions of 32 speakers (13 female, 19 male) to generate in excess of 16,000 isolated-word utterances. The stress domains included were: i) talking styles (slow, fast, soft, loud, angry, clear, question), ii) single tracking computer response task or speech produced in noise (Lombard effect), iii) dual tracking computer response task, iv) subject motion-fear tasks (G-force, Lombard effect, noise, fear), and v) psychiatric analysis data (speech under depression, fear, anxiety). A common highly confusable vocabulary set of 35 aircraft communication words also make up the database.

5 Result

It is seen that different types of stress and emotion may affect different frequency bands differently and an improved stress classification features should be obtained by analyzing energy in different frequency bands [13].

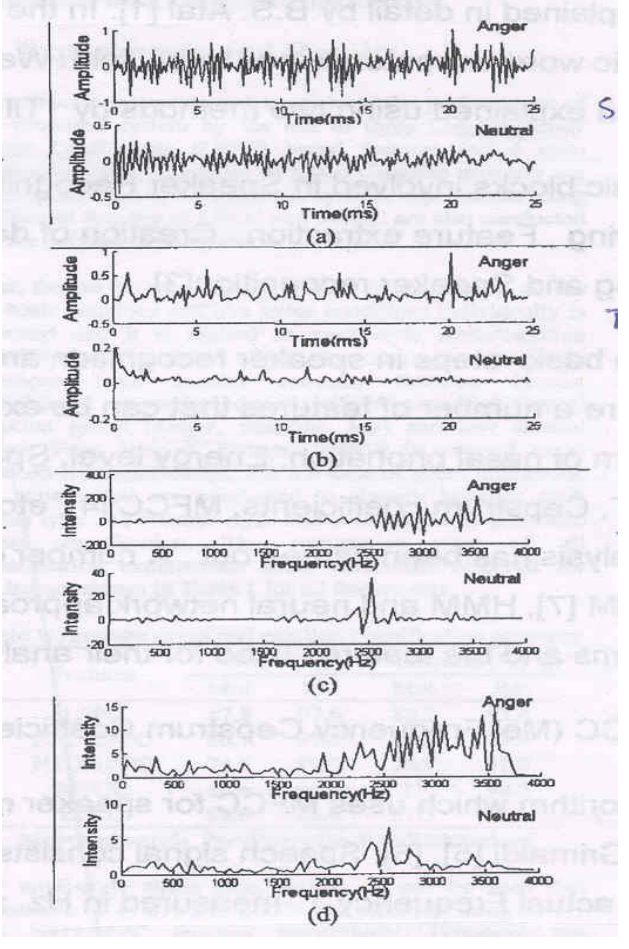

Fig. 3. a) Time domain representation of Anger and Neutral condition of the word 'destination'. b) TEO in Time Domain. c) TEO in Frequency Domain d) Frequency domain version of the word 'destination' using SUSAS database.

The time domain and frequency domain representations together with the results after the TEO operation for the emotion Anger and Neutral speaking styles are shown in Figure3. Anger has the higher frequency content than Neutral. Furthermore, as can be seen from Figure 3(c) for Anger, TEO operation suppresses certain intensity values in the frequency range 3 to 3.2k Hz down to near zero because of nonlinear property

analysis. This results in loss of important information for high frequency range, which is an essential feature of Anger [6]. Between NFD-LFPC (Figure 3(c)) and NTD-LFPC (Figure 3(b)), it can also be observed that nonlinear energy variations in frequency domain, present more significant discrimination among different speaking conditions. Anger has high intensity in higher frequency regions [14]. Neutral has higher intensity values in lower frequency scales. This emphasizes that Teager Energy Operation in frequency domain is more capable than in time domain to detect stress. The same trend has been observed between Anger (high arousal) and Sadness (low arousal) emotions. However, the graphical representations using emotion samples are not included and can be found in [14].

6 Conclusion

In this paper, comparison between novel systems for emotion classification is carried out. Traditional features like MFCC and LPCC are compared along with linear acoustic feature LFPC and nonlinear acoustic features NTD-LFPC which is in time domain and NFD-LFPC which is in frequency domain [14]. It is found that linear LFPC and nonlinear acoustic feature in frequency domain are important in representing speaking styles. After comparing the two approaches for TEO operation, it is observed that nonlinear variation of energy distribution using frequency domain analysis gives a better representation than in the time domain analysis. When comparing LFPC based features and two traditional features MFCC and LPCC, it is found that features LFPC, NFD-LFPC and NTD-LFPC perform well over the two traditional features. This has already been shown in fig3 above.

Refrences

1. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals. Pearson Education Publication
2. Bou-Gbde, S., Hansen, J.H.L.: A novel training approach for improving speech recognition under adverse stressful environments. In: EUROSPEECH 1997, Rhodes, Greece, vol. 5, pp. 2387–2390 (September 1997)
3. Emerich, S., Lupu, E., Apatean, A.: Emotion Recognition from speech and Facial Expressions Analysis. In: 17th European Signal Processing Conference, EUSIPCO 2009 (2009)
4. Tarng, W., Chen, Y.-Y., Li, C.-L., Hsie, K.-R., Chen, M.: Applications of Support Vector Machines on Smart Phone Systems for Emotional Speech Recognition. World Academy of Science, Engineering and Technology (2010)
5. Bou-Ghazale, S.E., Hame, J.H.L.: Stress perturbation of neutral speech for synthesis based on hidden Markov models. IEEE Transactions on Speech & Audio Processing 6(3), 201–216 (1998)
6. Sarikaya, R., Gowdy, J.N.: Subband based classification of speech under stress. Proceedings of the IEEE on Acoustics, Speech, and Signal Processing 1, 569–572 (1998)
7. Non, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech & Audio Processing 9(3), 201–216 (2001)

8. Cairns, D., Hansen, J.H.L.: Nonlinear analysis and detection of speech under stressed conditions. J. Acoust. Soc. Am. 96(6), 3392–3400 (1994)
9. Rabiner, L.R., Jug, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs (2010)
10. Fletcher, H.: Auditory Patterns. Review of Modern Physics 12, 4745 (1940)
11. Nadeu, D.M., Hemando, J.: Time and frequency filtering of filter-bank energies for robust HMM speech recognition. Speech Communication 34(1-2), 93–114 (2001)
12. Campbell Jr., J.P.: Department of Defense Fort Meade, MD. Speaker Recognition
13. Nwe, T.L., Foo, S.W.: Detection of Stress and Emotion in Speech Using Traditional and FFT based Log Energy Features. In: ICICS-PCM 2003, Singapore (2003)
14. Nwe, T.L.: Analysis and Detection of Human Stress and Emotion from Speech Signals. Ph. D. Thesis, National University of Singapore (2003)
15. Kandali, A., Routray, A., Basu, T.K.: Emotion recognition from Assamese speeches using MFCC features and GMM Classifier. In: Proceedings of IEEE Region 10 Conference 2008, Hyderabad, India (2008)
16. Recognition of emotions in speech by a hierarchical approach. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, Amsterdam, September 10-12, pp. 1–8 (2009)
17. Kotti, M., Patterno, F., Kotropoulos, C.: Speaker-independent negative emotion recognition. In: Proc. 2nd Int. Workshop Cognitive Information Processing, Elba Island, Italy, pp. 417–422 (June 2010)
18. Mower, E.: A Framework for Automatic Human Emotion Classification Using Emotion Profiles. IEEE Transactions on Audio, Speech and Language Processing 19(5) (July 2011)