Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 2, Issue 19; October-December, 2015, pp. 1649-1651

http://www.krishisanskriti.org/Publication.html

Comparative Analysis on the Performance of an Automated Assamese Vowel Recognition System with MFCC, Combination of MFCC and Energy, Combination of MFCC and Pitch Feature Vectors in Clean and Noisy Condition

Swapnanil Gogoi

Computer Science Gauhati University Institute of Distance and Open Learning
E-mail: swapnanil22@gmail.com

Abstract—*Mel-frequency cepstral coefficient (MFCC) is a very popular speech feature vector used in Automated Speech Recognition (ASR) Systems. In this paper, MFCC feature vectors are combined with two other speech feature vectors that are energy and pitch to construct two new set of speech vectors. Here the combination of MFCC and energy feature vectors is termed as MFCC-ENERGY and the combination of MFCC and pitch feature vectors is termed as MFCC-PITCH. Now a comparative analysis on the performance of an automated Assamese vowel recognition system with MFCC, MFCC-ENERGY and MFCC-PITCH feature vectors is performed in both clean and noisy condition. Here noisy speech database is constructed by adding five types of noises (babble, pink, white, volvo and factory) to the speech signals of Assamese vowels that are recorded in clean condition. Here for vowel recognition process, Hidden Markov Model (HMM) is used in training and testing phase. In training phase, a clean set of speech files are used and in testing phase, clean and noisy speech files are used. Experimented results shows that the performance of MFCC-PITCH is little better than MFCC and MFCC-ENERGY.*

Keywords*: MFCC, Energy, Pitch, Noise, ASR*

1. INTRODUCTION

In case of ASR systems, the performance is mainly dependent upon the selection of speech feature vectors. There are different types of feature vectors available in speech signals like energy, pitch, zero crossing rate, formants, MFCC etc. Now MFCC is very popular feature vector in ASR systems due to its improved robustness. It is observed that the performance of ASR systems with MFCC in case of recognition rate is better than the other mentioned feature vectors. In this paper, MFCC is combined with energy and pitch to form two new feature vectors and ASR experiments are performed.

2. ENERGY

The short time energy of a speech signal is defined as:

$$E_n = \sum_{m=n-N+1}^{n} x^2(m)$$

That is, the short-time energy at sample n is simply the sum of squares of the N samples n-N+1 through n [1].

Energy is an important feature vector of speech signals which can be used in speech recognition purpose [2]. But using only energy as a feature vector, the performance of a speech recognition system is not good and in noisy environment it will be degraded (see Table 1).

Short time energy of a speech signal is calculated by dividing the signal into frames so that the feature vectors can be calculated with useful information for ASR purpose.

3. PITCH

Pitch is a fundamental frequency produced due to the vibration of vocal folds and sub glottal air pressure to generate voiced signal. The voiced speech segments are near periodic in the time domain representation and the periodicity associated with such segments is defined as pitch period in the time domain and Pitch frequency or Fundamental Frequency in the frequency domain.

Pitch accents and phrase boundaries in speech has a close relationship with the lexico-syntactic structure of the utterance. The pitch accents are strongly correlated with syllable tokens that occur mostly in content words. These

dependencies can be used to augment the standard ASR model to improve recognition performance.[3]

There are different methods available in speech processing to estimate pitch. In our experiment we have used the following algorithm for estimation of pitch [4].

1. The analog signal is converted to digital by sampling with a suitable rate and quantized.
2. The digital signal is then hamming windowed to convert it into a suitable frame size. The signal is converted into frequency domain by using Fast Fourier Transform.
3. The absolute values of the signal are considered and then the logarithm of the signal is obtained.
4. The signal is then transformed into Cepstral domain by taking its Inverse Fast Fourier Transform. The very first signal peak represents the pitch frequency.

The performance of ASR system using pitch as a feature vector is not found to be good (see Table 1).

4. MFCC

In case of speech signal, the Mel Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [5]. In MFC, the frequency bands are equally spaced on the Mel scale, which is approximately similar to the human auditory system.

In general, MFCC of speech signals are estimated as follows:

1. Take the Fourier transform of a speech signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by (1) the number of filters, (2) the shape of filters, (3) the way that filters are spaced, and (4) the way that the power spectrum is warped [5].

In case of noisy speech signals, MFCC features contain some values that are not useful in ASR purpose and due to which the performance of the ASR is degraded (see Table 2).

5. MFCC-ENERGY

In this paper, a new speech feature vector is estimated termed as MFCC-ENERGY. At first MFCC feature vectors are estimated from each speech signal by considering the frame size 25ms and frame shift 10 ms. Then energy feature vectors are estimated by considering frame size 25ms and frame rate 10. Now MFCC-ENERGY is estimated by adding energy feature vectors to the MFCC feature vectors.

6. MFCC-PITCH

Here one more new speech feature vector is estimated termed as MFCC-PITCH. In this case also, MFCC feature vectors are estimated from each speech signal by considering the frame size 25ms and frame shift 10 ms .Then pitch feature vectors are estimated by considering frame size 25ms and frame rate 10. Now MFCC-PITCH is estimated by adding pitch feature vectors to the MFCC feature vectors.

7. SPEECH SIGNAL DATABASE PREPARATION FOR ASR EXPERIMENTS

The Assamese (অসমীয়া) is a main language in the state of Assam, India. In Assamese language , thirty two essential phonemes are available where total number of vowel phonemes is 8 that are ই (/i /) , এ (/ɛ/) , এ' (/e /) , আ (/a/) , অ' (/ɔ/) , অ (/ɒ/) , ঔ(/o/) and উ (/u/) [6].

For ASR experiments, here 10 men voices and 10 women voices are used to record speech signals for each of 8 Assamese vowels in a noise free environment. So, total 160 speech signals are recorded for the Assamese vowels with sampling rate 16 KHz and sampling format mono-channel, 16 bits resolution. Now this speech database is divided into two parts. The first part consists of the speech signals of 5 men and 5 women and it is used for training phase of the ASR process. The other part of the database is used in the testing phase of the ASR process and a noisy set of speech signals are also generated from this part by adding five types of noises (babble noise, pink noise, white noise, volvo noise and factory noise) from NOISEX-92 database [7] to the speech signals for testing the ASR rate in noisy condition.

8. ASR EXPERIMENT

For ASR experiments, HMM is used for training and testing phase. Here each Assamese vowel is modeled by a six states HMM model.

Now experimental results are shown in Table 2. Here it is observed that the performance of the Assamese vowel recognition process with MFCC-PITCH feature vectors is little better than MFCC and MFCC-ENERGY in case of clean, babble noise, volvo noise and factory noise. On the other hand, MFCC-PITCH performs little less than the others in case of pink noise. It is also observed that in case of white noise the ASR performance is same with all the three types of feature vectors.

Table 1: Assamese vowel recognition rate (in %) with Energy, Pitch and Combination of Energy and Pitch

	Energy	Pitch	Energy+Pitch
Clean	27.50	22.50	22.50
Babble noise	25.00	15.00	23.75
Pink noise	26.25	20.00	27.50
White noise	26.25	25.00	20.00
Volvo noise	23.75	20.00	20.00
Factory noise	25.00	20.00	21.25

Table 2: Assamese vowel recognition rate (in %) with MFCC, MFCC-ENERGY and MFCC-PITCH

	MFCC	MFCC-ENERGY	MFCC-PITCH
Clean	82.50	83.75	85.00
Babble noise	66.25	70.00	71.25
Pink noise	71.25	71.25	70.00
White noise	66.25	66.25	66.25
Volvo noise	80.00	78.75	82.50
Factory noise	66.25	67.50	68.75

9. CONCLUSION

In this paper, it is observed that the average performance of Assamese vowel recognition system with MFCC-PITCH feature vectors is little better than MFCC and MFCC-ENERGY. In clean condition, the best recognition rate is achieved with MFCC-PITCH feature vectors (85%). The performance of the ASR system can be improved by considering more number of speech signals in training phase. Further in case of noisy conditions, the ASR performance can be improved by using different speech enhancement techniques.

REFERENCES

[1] Rabiner, L.R. and Schafer, R.W., *Digital Processing of Speech Signals*, Pearson Education, 2009.

[2] Dimitriadis, Dimitrios, Maragos, P. and Potamianos, A., "On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition," in *Poceedings Audio, Speech, and Language Processing, IEEE Transactions*, August 19-6 2011, pp.1504-1516.

[3] Ananthakrishnan, S. and Narayanan, S., "Improved Speech Recognition using Acoustic and Lexical Correlates of Pitch Accent in a N-Best Rescoring Framework," in *Proceedings of IEEE International Conference* on , April 4-0 2007, pp. 873-876.

[4] Bageshree V., Pathak, Sathe and Panat, Ashish R., "Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person", *IJCSI International Journal of Computer Science*, July 4-1 2012.

[5] Fang Zheng, Guoliang Zhang and Zhanjiang Song, "Comparison of different implementations of MFCC", *Computer Science & Technology*, September 16-6 2001, pp.582-589.

[6] Kakati, Banikanta, *Assamese, its Formation and Development*, 5th edition, LBS Publications, 2007.

[7] Varga. A, Steeneken, H.J.M. and Jones, D. "The noisex-92 study on the effect of additive noise on automatic speech recognition system", Reports of NATO Research Study Group (RSG.10), 1992.