EFFECT OF MFCC BASED FEATURES FOR SPEECH SIGNAL ALIGNMENTS

Jang Bahadur Singh, Radhika Khanna and Parveen Lehana*

Department of Physics and Electronics, University of Jammu, Jammu, India

E-mail: sonajbs@gmail.com, pklehanajournals@gmail.com.

ABSTRACT

The fundamental techniques used for man-machine communication include Speech synthesis, speech recognition, and speech transformation. Feature extraction techniques provide a compressed representation of the speech signals. The HNM analyses and synthesis provides high quality speech with less number of parameters. Dynamic time warping is well known technique used for aligning two given multidimensional sequences. It locates an optimal match between the given sequences. The improvement in the alignment is estimated from the corresponding distances. ***.*** *The objective of this research is to investigate the effect of dynamic time warping on phrases, words, and phonemes based alignments. The speech signals in the form of twenty five phrases were recorded. The recorded material was segmented manually and aligned at sentence, word, and phoneme level. The Mahalanobis distance (MD) was computed between the aligned frames. The investigation has shown better alignment in case of HNM parametric domain. It has been seen that effective speech alignment can be carried out even at phrase level.*

KEYWORDS

Speech recognition, Speech transformation, MFCC, HNM, DTW

1. INTRODUCTION

Speech is one of the most dominating and natural means of communicate or express thoughts, ideas, and emotions between individuals [1]. It is a complicated signal, naturally produced by human beings because various processes are involved in the generation of speech signal. As a result, verbal communication can be varied extensively in terms of their accent, pronunciation, articulation, nasality, pitch, volume, and speed [2], [3].

Speech signal processing is one of the biggest developing areas of research in signal processing. The aim of digital speech signal processing is to take advantage of digital computing techniques to process the speech signal for increased understanding, improved communication, and increased efficiency and productivity associated with speech activities. The major fields of speech signal processing consists of speech synthesis, speech recognition, and speech/speaker transformation. In order to extract features from speech signal it has to be divided into fixed frames called windows. The types of the window function used are Hanning, Hamming, cosine, half parallelogram, not with right angles. Length of frame can be varied, depending upon nature of feature vectors to be extracted [4] Overlapping of two consecutive windows is done in order to maintain continuity. Feature extraction approaches mostly depend upon modeling human voice production and modeling perception system [5]. They provide a compressed representation of the speech signal by extracting a sequence of feature vectors [6].

DOI : 10.5121/ijnlc.2013.2205

2. SPEECH RECOGNITION

It is an alternative and effective method of interacting with a machine. It is also known as automatic speech recognition (ASR) converts spoken language in text. It has been two decades since the ASR have started moving from research labs to real-world. Speech recognition is more difficult than speech generation because of the fact that computers can store and recall enormous amounts of data, perform mathematical computations at very high speed, and do repetitive tasks without losing any type of efficiency. Because of these limitations, the accuracy of speech recognition is reduced. There are two main steps for speech recognition: feature extraction and feature matching. Isolated words of the input speech signal is analysed to obtain the speech parameters using Mel frequency cepstral coefficients (MFCC) or line spectral frequencies (LSF). These parameters provide the information related to the dynamically changing vocal tract during speech production. These parameters are then matched up to with earlier pattern of spoken words to recognize the closest match. Similar steps may be used for identity matching of a given speaker [7]. There are two types of speech recognition: speaker-dependent and speaker-independent. Speaker-dependent technique works by learning the distinctiveness of a single speaker like in case of voice recognition, while speaker-independent systems involves no training as they are designed to recognize anyone's voice. As the acoustic spaces of the speakers are multidimensional, reduction of their dimensionality is very important [8]. The most common method for training of the speaker recognition system is hidden Markov model (HNM) and its latest variant is (HMM-GMM) [9]. For better alignment or matching, normalization of the sub-band temporal modulation envelopes may be used [17]. The main factors responsible for the stagnation in the fields of speech recognition are environmental noise, channel distortion, and speaker variability [10], [11], [12].

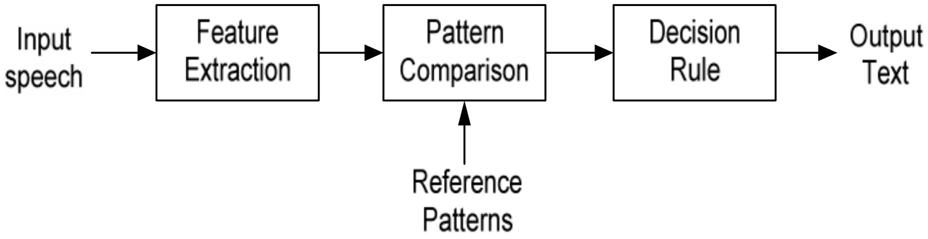


Fig. 1 Blocks of speech recognition as pattern matching problem [13].

Let us consider simple speech recognition as a pattern matching problem. As shown in fig. 1, the system accepts an input speech waveform. The feature extraction sub-block converts input speech waveform to a set of feature vectors. These vectors represent speech characteristics/features suitable for detection and then match up to it with reference patterns to find a closest sample. If the input speech produced by the alike speaker or similar environmental condition from the reference patterns, the system is likely to find the correct pattern otherwise not [13], [14]. The few valuable approaches used for feature extraction are: full-band and subband energies [15], spectrum divergence between speech signal and noise in background [16], pitch estimation [17], zero crossing rate [18], and higher-order statistics [19], [20], [21], [22]. However, using long-term speech information [23], [24] has shown considerable benefits for identifying speech signal in noisy environments. Several methods used for the formulation of the rules in decision module based on distance measures techniques like the Euclidean distance [25] Itakura-Saito and Kullback-Leibler divergence [26]. Some other method are based on fuzzy logic [27] [support vector machines (SVM) [28] and genetic algorithms [29], [30].

3. METHODOLOGY

The analysis process of the speech signals at phrase, word and phoneme levels was carried out with the raw recordings of six speakers in Hindi language. Speakers were of different age group, from different regions of Jammu. Sony recorder (ICD-UX513F) has been used for the recording of speech. It is a 4GB UX Digital Voice Recorder with expandable memory capabilities, and provides high quality voice recording. The investigation further includes the segmentation, labeling of the recorded speech at phrase, word and phoneme levels. Fig. 2 shows sub-section of speech signal waveform and spectrogram labeled at phoneme level. Feature extraction and lastly alignment of source and target feature vectors using DTW technique and then calculating Mahalanobis distance between them.

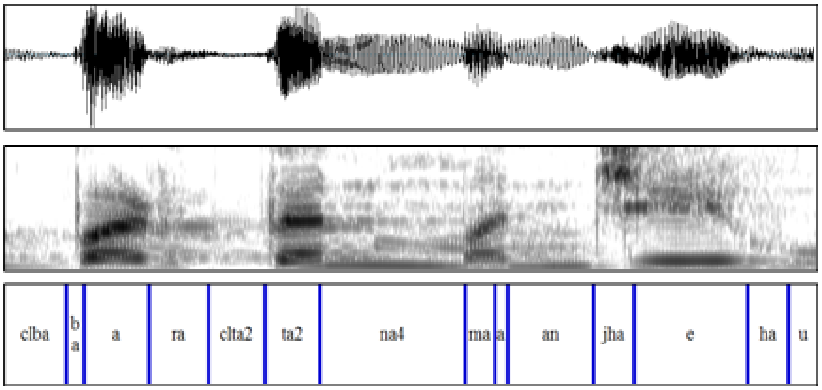


Fig. 2 Sub-section of speech signal waveform and spectrogram labeled at phoneme level.

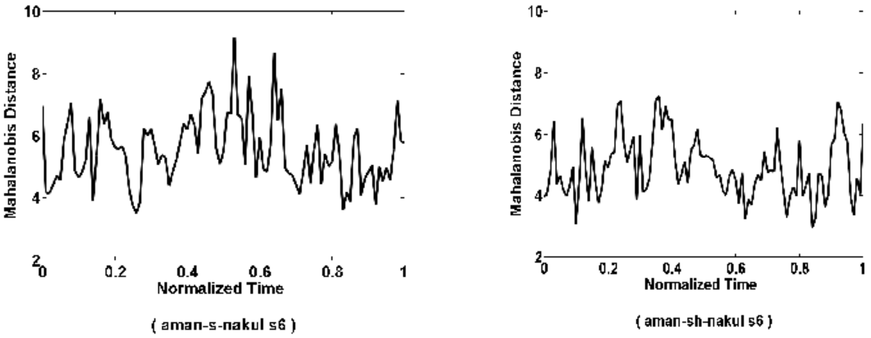
Thus main experiment is separated into two main sections. In first section, segmented source and target speech, features vectors were extracted using MFCC algorithms. The features extracted were aligned separately by means of DTW techniques and alignment error was calculated using Mahalanobis distance. In second section, segmented source and target speech were analysed first by HNM module in order to obtain the HNM parameters afterward MFCC features were calculated. The extracted features were aligned by means of DTW techniques and alignment error was calculated using Mahalanobis distance. These steps were implemented on the phrase, word and phoneme levels of the speech signal.

4. RESULTS AND DISCUSSION

In order to analyse the results based on the techniques namely MFCC, HNM, and DTW for the alignment of segmented speech at phrase, word, and phoneme levels. These investigations were carried out with male to male, female to male, and female to female speaker combinations. In order to compare alignment techniques, mean and standard deviation of Mahalanobis distances were calculated. Alignment error using Mahalanobis distances of various male-male combinations are represented from fig. 3 to fig. 5. The various combinations of speaker pair's use as source and targets were written on the bottom of each plot, e.g. aman-s-nakul represents source target alignment without using HNM at sentence level while aman-sh-nakul represents source target alignment using HNM at sentence level.

From the below graphs it can easily be analysed that alignment error at all segmented level of speech decreases by using HNM model, as Mahalanobis distances reduces. Therefore using HNM model, alignment can further be improved in comparison with the feature extraction method based on MFCC only. Decrease in the Mahalanobis distances means better alignment between the two sequences.

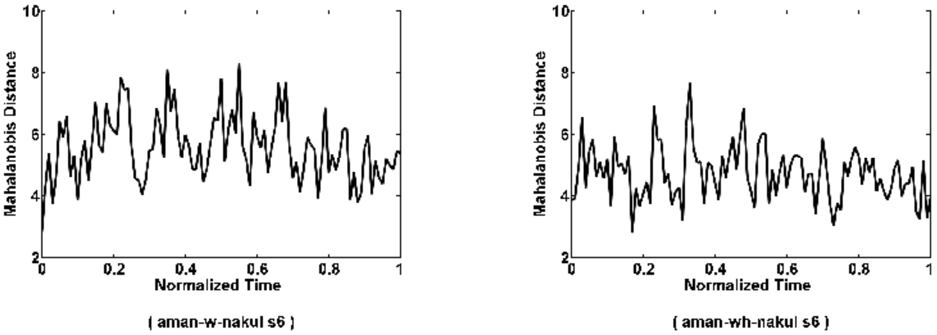
The overall comparisons of the speech alignment were shown in form of bar graphs. It can be well estimated that better speech alignment can be achieved at sentence level segmentation rather than word level or phoneme level segmentation of speech.



(aman-s-nakul s6)

(aman-sh-nakul s6)

Fig. 3 MFCC based speech alignment errors using Mahalanobis distance at sentence level with male to male combination. The first column shows the results for alignment without the use of HNM based method and second column shows the results for HNM based method.



(aman-w-nakul s6)

(aman-wh-nakul s6)

Fig. 4 MFCC based speech alignment errors using Mahalanobis distance at word level with male to male combination. The first column shows the results for alignment without the use of HNM based method and second column shows the results for HNM based method

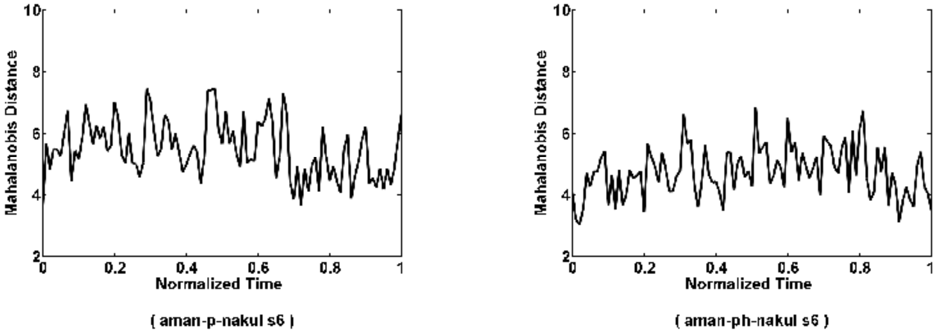


Fig. 5 MFCC based speech alignment errors using Mahalanobis distance at phoneme level with male to male combination. The first column shows the results for alignment without the use of HNM based method and second column shows the results for HNM based method.

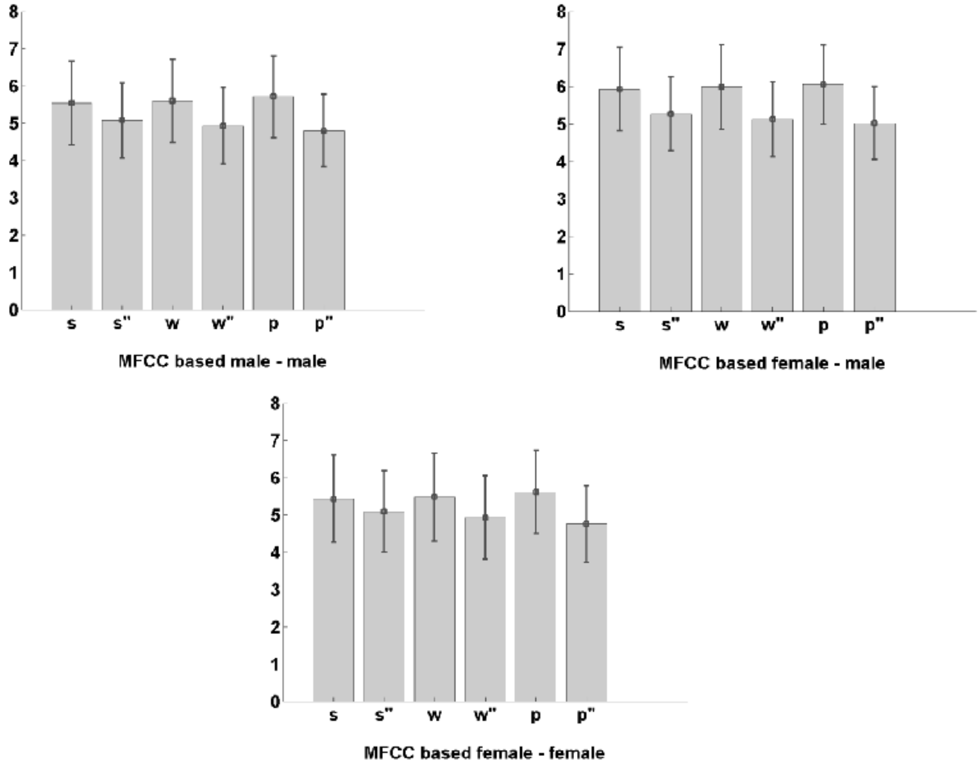


Fig. 6 Bar graph illustrate averaged MFCC based mean along standard deviation speech alignment errors using Mahalanobis distance at sentence, word and phoneme level with all male to male, female to male and female to female combinations. The symbol s, w, p shows the results for alignment without the use of HNM based method and the symbol s", w", p" shows the results for HNM based method.

5. CONCLUSION

The speech signals in the form of twenty phrases were recorded from each of six speakers. Investigations were carried out with three different combinations of male to male, female to male, and female to female speakers. The feature vectors are extracted using MFCC, and HNM techniques. Speech alignment errors using Mahalanobis distances for labeled phrases, words, and phonemes, aligned by means of DTW were calculated. From the analysis of the results it can be observed that alignment error with HNM model decreases at all the levels of labeled speech levels. Therefore implementing HNM model alignment error can further be reduced in compression with the feature extraction method based on MFCC only. Reduction in alignment error means better alignment or matching between two speech signals. Thus this research work is concluded as follows in brief:

1) HNM based alignment is more promising than other existing techniques.
2) The accuracy of speech recognition cannot be considerably increased even labeling the phrases at phoneme level. Effective speech alignment can be obtained at phrase level also which save our valuable time and unnecessary step in algorithm.
3) It is remarkable to note that speech alignment error in case of female-male is much larger than other combinations of male-male and female-female. Therefore such combination must be avoided in speech recognition and speaker transformation.

REFERENCES

[1] Zaher A A (2009) “Recent Advances in Signal Processing”, Publisher: InTech, pp. 544.

[2] Tebelskis J (1995) “Speech Recognition using Neural Networks”, Ph.D. dissertation, Dept. School of Computer Science, Carnegie Mellon University Pittsburgh, Pennsylvania.

[3] http://www.learnartificialneuralnetworks.com/speechrecognition.html

[4] Lindholm S (2010) “A speech recognition system for Swedish running on Android”.

[5] Ursin M (2002) “Triphone Clustering in Finnish Continuous Speech Recognition”, M.S thesis, Department of Computer Science, Helsinki University of Technology, Finland,

[6] Trabelsi I and Ayed D Ben (2012) “On the Use of Different Feature Extraction Methods for Linear and Non Linear kernels”, Proc. IEEE, Sciences of Electronic, Technologies of Information and Telecommunications.

[7] Rabiner L R (1989) “A tutorial on Hidden Markov Models and selected applications in speech recognition".

[8] Wang1 X & Paliwal K K (2003) “Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition,” Pattern Recognition, vol. 36, pp. 2429.

[9] Muller F & Mertins A (2011) “Contextual invariant-intergration features for improved speaker-independent speech recognition”, Speech communication, pp. 830.

[10] Lu X, Unoki M & Nakamura S (2011) “Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments”, Computer Speech and Language, pp. 571.

[11] Mporas I, Ganchev T, Kacsis O & Fakotakis N (2011) “Context-adaptive pre-processing scheme for robust speech recognition in fast-varying noise environments”, Signal Processing, vol. 91, pp. 2101.

[12] Dai P & Soon I Y (2011) “A temporal warped 2D psychoacoustic modeling for robust speech recognition system”, Speech Communication, vol. 53, pp. 229.

[13] Rabiner L R & Juang B H (1993) “Fundamentals of Speech Recognition”, Prentice-Hall.

[14] Doh S J (2000) “Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression”, PhD, Department of Electrical and Computer Engineering, Carnegie Mellon University Pittsburgh.

[15] Woo K, Yang T, Park K & Lee C (2000) “Robust voice activity detection algorithm for estimating noise spectrum”, Electronics Letters, vol. 36, pp. 180.

[16] Marzinzik M & Kollmeier B (2002) “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics”, IEEE Trans. Speech Audio Processing, vol. 10, pp. 341.

[17] Tucker R (1992), “Voice activity detection using a periodicity measure”, Proc. Institute of Electronic Engineering, vol. 139, pp. 377.

[18] Bachu R G, Kopparthi S, Adapa B. & Barkana B D (2008) "Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal", Proc., American Society for Engineering Education, Zone Conference, pp. 1.

[19] Nemer E, Goubran R & Mahmoud S (2001) "Robust voice activity detection using higher order statistics in the lpc residual domain", IEEE Trans. Speech Audio Processing, vol. 9, pp. 217.

[20] Ramirez J, Gorriz J, Segura J C, Puntonet C G & Rubio A (2006) "Speech/Non-speech Discrimination based on Contextual Information Integrated Bispectrum LRT", IEEE Signal Processing Letters, vol. 13, pp. 497.

[21] Gorriz J M, Ramirez J, Puntonet C G & Segura J C (2006) "Generalized LRT-based voice activity detector", *IEEE Signal Processing Letters*, vol. 13, pp. 636.

[22] Ramírez J, Gorriz J M & Segura J C (2007) "Statistical Voice Activity Detection Based on Integrated Bispectrum Likelihood Ratio Tests", Journal of the Acoustical Society of America, vo. 121, pp. 2946,

[23] Ramirez J, Segura J C, Benitez C, Torre A & Rubio A (2004), "Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information", Speech Communication, vol. 42, pp. 271.

[24] Ramirez J, Segura J C, Benitez C, Torre A & Rubio A (2005) "An Effective Subband OSF-based VAD with Noise Reduction for Robust Speech Recognition", IEEE Transactions on Speech and Audio Processing, vol. 13, pp. 1119.

[25] Gorriz J M, Ramirez J, Segura J C & Puntonet C G (2006) "An effective cluster-based model for robust speech detection and speech recognition in noisy environments", Journal of the Acoustical Society of America, vol. 120, pp. 470.

[26] Ramirez J, Segura J C, Benitez C, Torre A & Rubio A (2004) "A New Kullback- Leibler VAD for Robust Speech Recognition", IEEE Signal Processing Letters, vol. 11**,** pp. 266.

[27] Beritelli F, Casale S, Rugeri G & Serrano S (2002) "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors", IEEE Signal Processing Letters, vol. 9, pp. 85.

[28] Ramirez J, Yelamos P, Gorriz J M & Segura J C (2006) "SVM-based Speech Endpoint Detection Using Contextual Speech Features", IEE Electronics Letters, vol. 42.

[29] Estevez P A, Yoma N, Boric N & Ramirez J A (2005) "Genetic programming based voice activity detection", Electronics Letters, vol. 41, pp. 1141.

[30] Ramirez J, Gorriz J M & Segura J C (2007) "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness", Robust Speech Recognition and Understanding, I-TECH Education and Publishing, pp. 1.