International Journal on Natural Language Computing (IJNLC) Vol. 1, No.3, October 2012

EFFECT OF DYNAMIC TIME WARPING ON ALIGNMENT OF PHRASES AND PHONEMES

Padmini Rajput, Radhika Khanna, Parveen Lehana and Jang Bahadur Singh

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

*pklehanajournals@gmail.com

ABSTRACT

Speech synthesis and recognition are the basic techniques used for man-machine communication. This type of communication is valuable when our hands and eyes are busy in some other task such as driving a vehicle, performing surgery, or firing weapons at the enemy. Dynamic time warping (DTW) is mostly used for aligning two given multidimensional sequences. It finds an optimal match between the given sequences. The distance between the aligned sequences should be relatively lesser as compared to unaligned sequences. The improvement in the alignment may be estimated from the corresponding distances. This technique has applications in speech recognition, speech synthesis, and speaker transformation. The objective of this research is to investigate the amount of improvement in the alignment corresponding to the sentence based and phoneme based manually aligned phrases. The speech signals in the form of twenty five phrases were recorded from each of six speakers (3 males and 3 females). The recorded material was segmented manually and aligned at sentence and phoneme level. The aligned sentences of different speaker pairs were analyzed using HNM and the HNM parameters were further aligned at frame level using DTW. Mahalanobis distances were computed for each pair of sentences. The investigations have shown more than 20 % reduction in the average Mahalanobis distances.

KEYWORDS

Speech recognition, Dynamic time warping DTW, Mahanalobis distance, segmentation & alignment.

1. INTRODUCTION

Speech signal is generated as a consequence of exciting a dynamic vocal tract system with time varying excitation. Speech is the most innate and fastest means of human interaction. The attributes of speech have made researchers to continuously improve the man-machine communication leading to the development of efficient and intelligent techniques like speech recognition, which is continuously gaining a serious attention since more than fifty years [1] [2]. This type of communication is valuable when our hands and eyes are busy in some other task such as driving a vehicle, performing surgery, or firing weapons at the enemy. Speech recognition 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, in spite 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. The reason for this may be attributed to the lack of general knowledge in the computers. Because of these limitations, the accuracy of speech recognition is reduced. There are two main steps for speech recognition: feature extraction and feature matching. Each word in the input speech signal is isolated and then analyzed to

DOI : 10.5121/ijnlc.2012.1302

obtain the parameters such as Mel frequency cepstral coefficients (MFCCs) or line spectral frequencies (LSF). These parameters provide the information related to the dynamically changing vocal tract during speech production. These parameters are then compared with previous examples of spoken words to identify the closest match. Similar steps may be used for identity matching of a given speaker [3].

Modern architectures for Automatic Speech Recognition (ASR) are mostly software architectures that construct a sequence of word hypotheses out of an acoustic signal. In recent years the use of Multi-layer perceptron (MLP) derived acoustic feature has gradually become more popular in ASR systems [4]. 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 [5]. The most common method for training of the speaker recognition system is hidden Markov model (HMM) and its latest variant is (HMM-GMM) [6]. For better alignment or matching, normalization of the sub-band temporal modulation envelopes may be used [7]. Although, a lot of effort has been put for improving the speech recognition, until now the performances of such systems are unappealing in real world tasks [8]. The main factors responsible for the stagnation in the fields of speech recognition are environmental noise, channel distortion, and speaker variability [7] [9] [10]. For alignment, segmentation is used and it is a technique by means of which the boundaries between words, syllables, or phonemes in spoken languages are identified [1]. The segmentation is divided into two levels. The lowest level of speech segmentation is carried out by the subdivision of a sound into a sequence of phones. Another process makes use of lexical segmentation which means the splitting up of sound into words of a language
.
The manual procedure involves an approach of listening and visual judgment in order to identify the boundaries of meaningful speech segments. This process seems impractical in case of huge data bases so different techniques are developed for this principle which led to the improvement of automatic speech segmentation done by a chosen procedure or algorithms with the objective of using the results for speech synthesis, data training for speech recognizers or to build and label prosodic data basis [11]. The improvement in the alignment may be estimated from the corresponding distances between the frames of the given sentences of two speakers. The objective of this research is to investigate the amount of improvement in the alignment corresponding to the sentence based and phoneme based manually aligned phrases. The aligned sentences of different speaker pairs are analyzed using HNM and the HNM parameters were further aligned at frame level using DTW. The justification for choosing HNM is that it is a very efficient model for speech generation. The alignment is carried out only with voiced segments. For this, the harmonic magnitudes are converted to LSF. Section 2 describes the working principle of HNM and its parameter extraction procedure while Section 3 explains the methodology including dynamic time warping. The results are presented in the Section 4.

2. HNM

HNM, a variant of sinusoidal model, is an analysis/modification/synthesis model which provides high quality speech with less number of parameters, and with pitch and time scaling relatively easy compared to all existing models and seems to be more promising for speech synthesis compared to other existing models [11] [12]. Research has shown that all vowels and syllables can be produced with a better quality syllables by the implementation of HNM [13]. Results obtained from many speech signals including both male and female voices are quite satisfactory with respect to the background noise and inaccuracies in the pitch [14]. In HNM, each segment of speech can be modeled as two bands: “a lower harmonic part” can be represented using the

amplitudes and phases of the harmonics of a fundamental and an “upper “noise" part using an all pole filter excited by random white noise, with dynamically varying band boundary.

Out of the two sub bands of the speech spectrum, one is modelled with harmonics of the fundamental and the other is simulated using random noise. The harmonic part and noise part constitute the quasi-periodic components and non-periodic part respectively [11]. The frequency that separates the two bands is called maximum voiced frequency F_m. The lower band represents the signal by harmonic sine waves, slowly varying in amplitudes and frequencies:

$$s'(t) = \mathrm{Re}\sum_{l=0}^{L(t)} a_l(t)\exp\{j[\int_0^t l\omega_o(\sigma)d\sigma + \theta_l]\} \quad (1)$$

where $a_l(t)$ and $\theta_l(t)$ represent the amplitude and phase at time t of the l_{th} harmonic, while $w_o l(t)$ are the fundamental and time-varying number of harmonics included in the harmonic part. AR model represents the upper band constituting the noise part modulated by the time domain amplitude envelope. The noise part $n'(t)$ is obtained by filtering a white Gaussian noise $b(t)$ by a time varying, normalized all-pole filter $h(\tau : t)$. The result obtained is multiplied by an energy envelope function $w(t)$:

$$n'(t) = w(t)[h(\tau;t) * b(t)] \quad (2)$$

In addition to obtaining the maximum voiced frequency F_m, other parameters like voiced/unvoiced, amplitudes and phase of harmonics of fundamental frequency (pitch), glottal closure instants, parameters of noise part, and pitch are calculated for each frame. Figure 1 depicts the analysis using HNM. Speech signal is fed by the voicing detector which states the frame either voiced or unvoiced. HNM analysis is pitch synchronous so their lies the exact inference of the glottal closure instances (GCIs) [11]. GCIs can be calculated either by means of the speech signal or electroglottogram (EGG). Speech signal or EGG is given at the input side of GCI. Maximum voiced frequency F_m is calculated for each voiced frame. The analysis frame is taken twice the local pitch period. Form each GCI the voiced part is analyzed for calculating amplitudes and phase of all the pitch harmonics up to F_m.

The synthesized portion of the voice part is calculated from equation (1) for obtaining noise parameters while the remaining fraction obtained as result of the subtraction of the noise from the speech signal is the voiced part. Noise part is later analyzed for the LPC coefficients and energy envelope. For both voiced and unvoiced frames the length of the analysis window for noise part is taken as two local pitch periods. However for unvoiced frames the local pitch is the pitch of the last frame and for voiced frames the local pitch is the pitch of the frame itself and [11]. The addition of the synthesized speech.HNM based synthesis can be used for good quality output with relatively small number of parameters. Using HNM, pitch and time scaling are also possible without explicit estimation of vocal tract parameters [11]. Speaker transformation and voice conversion method has been a hot area of research in speech processing research for the last two decades [15] [16] [17] [18]. These techniques are also implemented in the framework of the HNM system, which allows the high-quality modifications of speech signals. In comparison to earlier methods based on the vector quantization, HNM based conversion scheme results in high quality modification of speech signal [17].

3. METHODOLOGY

For the analysis of the speech signal, we have carried out the recording of six speakers in Hindi language (3 males and 3 females). Speakers of different age group, from different regions of Jammu have been taken. Twenty five sentences for each speaker were recorded. The recorded material was segmented and labelled manually.

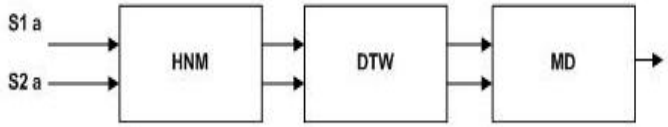

Figure 2. Estimation of Mahalanobis distances among HNM parameters

Different sentences for different speaker combinations (male-female, female-female, and male-male) were given as input to the HNM analysis module in order to obtain the HNM parameters of the recorded speech (Figure 2). The HNM parameters are converted to LSFs and applied to the next block for DTW [19]. DTW is an algorithm for measuring similarity between two sequences which may vary in time or speed. DTW finds an optimal match between two given sequences. Dynamic time warping is pattern matching based approach for finding an optimal distance between two given sequences wrapped in a non-linear fashion under certain restrictions; it is a well established technique for time alignment and comparison of speech and image patterns [20]. It is a form of dynamic programming and is extensively applied for pattern matching. We have used two techniques for alignment. In the first technique, the unlabeled sentences were directly given to the HNM. In second technique, phoneme marked sentences were given to the HNM. The Mahalanobis distance (MD) was calculated between the aligned frames. The Mahalanobis distance in parametric space was estimated using the following relation defined for feature vectors **X** and **Y** as

$$D_{\mathrm{M}}(\mathbf{X},\mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})^{\mathrm{T}} \Sigma^{-1} (\mathbf{X}-\mathbf{Y})} \,. \qquad (3)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the feature vectors used in training [21].

4. RESULTS

The average Mahalanobis distances calculated for each speaker pair is listed in Table 1. It is clear from this table that more than 20 % improvement in alignment is obtained if the phoneme marked sentences are given as input to the HNM and consequently to DTW. It may be noted that silence frames and unvoiced segments were removed before applying the frames to the DTW. Our preliminary investigations have shown that DTW does not perform satisfactory for these frames. The alignment time sequence (output of the DTW) and the Mahalanobis distances for three speaker pairs (female to female, female to male, and male to male) are shown in Figure 3. Although, the curves for Mahalanobis distances for the unlabeled sentences and the phoneme marked sentences are very near to each other, the percentage improvement is observed more for phoneme marked sentences.

Table 1. Average Mahalanobis distances.

Speaker Pair	Sentence based distance	Phoneme based distance	Improvement (%)
F2F	5.78	4.50	22
F2M	6.00	4.73	21
M2M	9.10	7.02	23

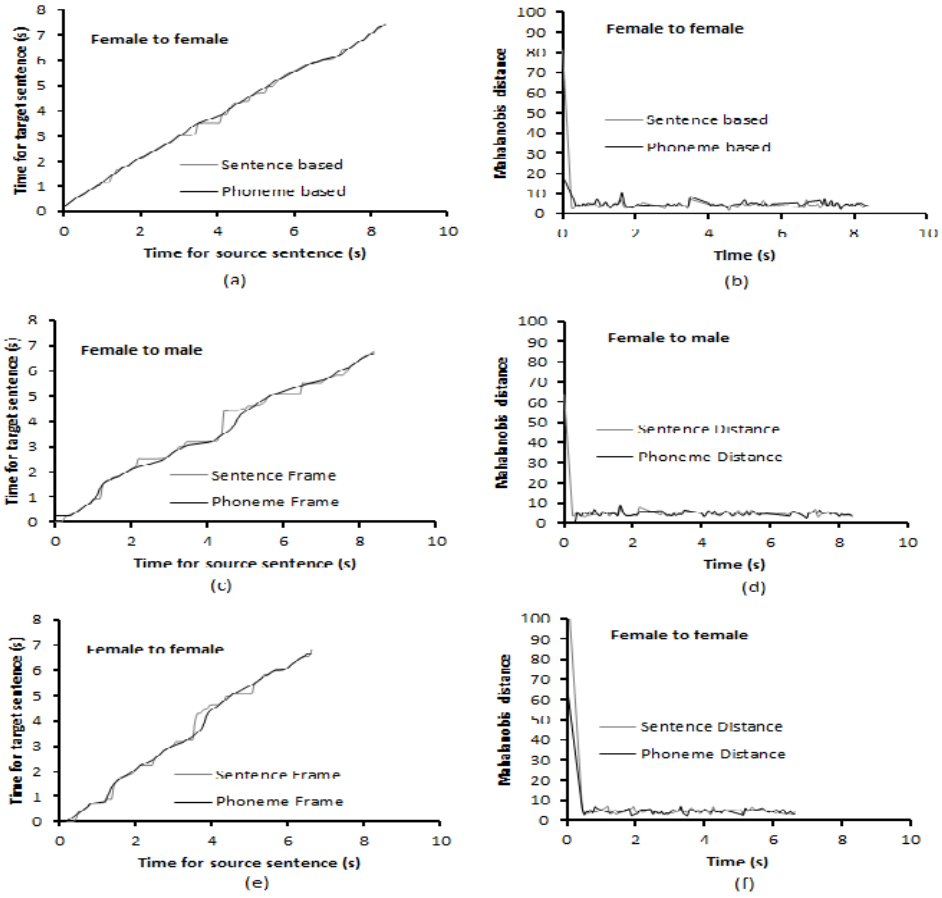

Figure 3. Plots for time relationship between DTW aligned frames (first column) and variation of Mahalanobis distances with respect to time (second column).

5. CONCLUSIONS

Investigations were carried out with three different combinations of male-male, male-female and female-female speakers for estimating the average Mahalanobis distances for unlabeled and phoneme marked sentences. From the above results we observe that phoneme based alignment is always far better than phrase level alignment. The results were quite obvious but the percentage of improvement (more than 20 %) constraints the necessity of using phoneme based alignment for developing efficient mapping for speaker recognition and speaker transformation.

REFERENCES

[1] Prasanna S R M & Zachariah J M (2002) "Detection of vowel onset point in speech", in Proc. IEEE Int Conf. Acoust Speech Signal Processing Orlando, Vol. 4, pp 4159.

[2] Ayadi M E, Kamel M S & Karray F (2010) "Survey on speech emotins recognition: Features, classificatiob schemes and databases", ELSEVIER Pattern recognitio, pp 572-587.

[3] Chaudhari U V, Navaratil J & Maes S H (2003) "Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition", IEEE Trans. on Speech and Audio Processing, Vol. 11, No. 1, pp 61-69.

[4] Park J, Delhi F, Gales M J F, Tomalin M & Woodland P C (2011) "The efficient incorporation of MLP features into automatic speech recognition system", ELSEVIER, Computer Speech and Languages, pp 519-534.

[5] Wang X & Paliwal K (2002) "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition", ELSEVIER Pattern Recognition, pp.2429 – 2439.

[6] Muller F & A Mertins (2011) "Contextual invariant-intergration features for improved speaker-independent speech recognition", Institute for Signal Processing, University of Lubeck, ELSVIER Speech communication, Feburary, pp 830-841.

[7] Lu X, Unoki M & Nakamura S (2011) "Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments", ELSVIER Computer Speech and Language, pp 571-584.

[8] Mohammadi A & Almasganj F (2011) "Reconstruction of missing features by means of multivariate Laplace disttrobution (MLD) for noise robust speech recognition", ELSVIER Expert Systems with Applications, Vol. 38, No. 4, pp 3918-3910.

[9] Mporas Iosif, Ganchev Todor, Kacsis Otilla & Fakotakis Nikos, (2011) "Context-adaptive pre-processing scheme for robust speech recognition in fast-varying noise enviornments", ELSVIER Signal Processing, Vol. 91, No. 8, pp 2101-2111.

[10] Dai Peng & Soon Ing Yann (2011) "A temporal warped 2D psychoacoustic modeling for robust speech recognition system", ELSVIER Speech Communication, Vol. 53, No. 2, pp 229-241.

[11] Lehana Parveen Kumar & Pandey Prem Chand (2003) "Effect of GCI perturbation on speech quality in Indian languages", in Proc. Convergent Technologies for the Asia Pacific (IEEE TENCON-2003), Vol. 3, pp 959-963.

[12] Lehana Parveen Kumar & Pandey Prem Chand (2004) "Harmonic plus noise model based speech synthesis in Hindi and pitch modification", in Proc. 18th Int. Cong. Acoust., ICA 2004, pp 3333-3336.

[13] Stylianou Yannis (2001) "Removing linear phase mismatches in concatenative speech synthesis", in Proc. IEEE. Vol. 9, No. 3, pp 232-239.

[14] Arslan Levent M (1999) "Speaker transformation algorithm using codebooks (STASC)," in Proc. Speech Commun. Vol. 28, No. 3, pp 211-226.

[15] Stylianou Yannis (2001) "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Trans. Speech and Audio Processing, Vol. 9, No. 1, pp 21-29.

[16] Larock J, Stylianou Yannis & Moulines E (1993) "HNM: A simple, efficient harmonic noise model for speech", in Proc. IEEE Workshop App. Signal Process., Audio, Acoust., pp 169-172.

[17] Taylor Paul, Black Alan W & Richard Caley (1998) "The architecture of the Festival speech synthesis system", The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.

[18] Stylianou Yannis, (1998) "Concatenative speech synthesis using a harmonic plus noise model", Third ESCA, Speech Synthesis, pp 261-266.

[19] Rabiner L & Juang B H (1999) Fundaments of Speech Recognition, Englewood Cliffs, NJ: Prentice Hall.

[20] Al-Manie Mohammed A, Alkanhal Mohammed I & Mansour M Al-Ghamdi (2010) "Arabic speech segmentation: Automatic verses manual method and zero crossing measurements", Indian Journal of Science and Technology, Vol. 3, No. 12, pp 1134.

[21] Mahalanobis Prasanta Chandra (1936) "On the generalised distance in statistics," in Proc. National Institute of Sciences of India, Vol. 2, No. 1, pp 49–55.