Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

AL-QURAN RECITATION SPEECH SIGNALS TIME SERIES SEGMENTATION FOR SPEAKER ADAPTATION USING DYNAMIC TIME WARPING

N. Shafie*, M. Z. Adam and H. Abas

Advanced Informatics School, Universiti Teknologi Malaysia

Published online: 01 February 2018

ABSTRACT

The transformation of the whole traditional process of Al-Quran Recitation into automated system application could raise the issues of robustness and integrity of recitation and correction quality and acceptance. There are several existing variations especially involve of rhythm, tone and length of recitation for different speaker. Dynamic Time warping (DTW) is used as dynamic programming to normalize the recitation speech length of speaker which warp the speech spectrum amplitude in time series for the experts and learners. DTW is used to gain the same length of recitation which warp the amplitudes, rhythm and tone into same length in time series segments based on formant frequency frame for different recitation speakers. There are 8 experts and 10 users from the Malay Muslim community had three selected surahs for Al-Quran recitation session. The aim of the paper is to normalize Al-Quran recitation speech signals as speaker adaption between experts and learner then represent each recitation speech signal at same vocal tract formant frequency that can be used in robust Automatic Recitation Recognition (ARR) to evaluate the performance of recitation evaluation.

Keywords: Al-Quran speech recitation, Automatic Recitation Recognition, speaker adaptation, vocal tract length normalization.

Author Correspondence, e-mail: noraimi.kl@utm.my

doi: http://dx.doi.org/10.4314/jfas.v10i2s.11

1. INTRODUCTION

In confronting the variability and complexity of the continuous speech recitation signals, the proposed approach introduced in this paper has manipulated the availability of signal properties to determine the performance of recitation between experts and learners. Before the performance can be evaluated. The both speech recitation of the expert and learner should be parameterized by a single warp factor for each learner based on expert recitation to use as dynamical parameters. The learner recitation speech signals are warping the energy in the same length of recitation in the time series frame. The main purpose of the paper is to get the same rhythm, tone and length between expert and learner recitation speech signals that can be compared in same word/utterance articulation from the vocal tract. The stage of the experiments was divided into studios data preparation, compensation of recitation speech signal and dynamic programming approach.

2. PROBLEM FORMULATION

The length of expert recitation speech signals was used as a template and the learner recitation speech signals was used as input. The inconsistency and non-uniform of time series in comparison on both signals is shown in figure 1 is the major problematic. Before doing evaluation, both of recitation signals must nevertheless be able to say that the distance between both recitations speech signals should be small and able to correlate based on string or phoneme.

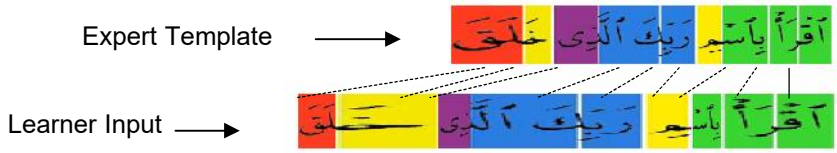

Fig.1. Different input and template between expert and learner recitation

The sequences of feature vectors have to be placed at same varying speech rate by stretching and shrinking of speech segments. The meaningful comparison parameters in (1) can be produced to ensure the both signals must be kept at the same length. The meaningful comparison is succeed when it is able to maintain the characteristics in both original speech signals time series without losing information while stretching and shrinking. Three

parameters that should be considered are minimum path cost, horizontal alignment and variability of recitation. Minimum path cost is used to obtain optimal alignment by computing all the possible cost path and determine the lowest overall cost from the path to achieve high similarity between two temporal recitation speech signals. Horizontal alignment refers to the alignment in time series axis only which is to maintain the characteristic of energy, rhythm and tone in the frame. Then, the variability parameters represents the variation of recitation that can be reduced if the reciter accurately follow the rules of Al-Quran recitation. The proportional equation that shows the relationship between all parameters is given by

$$Meaningful\ comparison \propto \frac{minimum\ path\ cost \times horizontal\ alignment}{variability} \quad (1)$$

3. TIME SERIES ALIGNMENT

Speech is a variable process and stochastic [4] in which the duration of a word and its sub-words vary randomly. Time alignment or normalization can be applied in recitation which required to find the best alignment between the sequences of experts and learners recitation vector features can be compare similarly on time occurrence without concerning of gender, age or health condition of the reciter. There are two approaches. Firstly, by adjusting the frame shift rate in the calculation of the cepstrum features vectors sequence of each input name to yield a pre-specific number of uniformly-spaced cepstrum feature vectors across the duration of the spoken name. Secondly, by using dynamic time warping (DTW) to normalize two feature vector sequences in time. The technique of minimum Euclidean distance approach was used in DTW. The DTW is measuring similarity between two temporal sequences which may vary in time. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linear by stretching or shrinking it along its time axis [4]. This technique will be used to re-map the recitation, such that all utterance was generated from same vocal tract. DTW already well known in speech recognition to cope with different speaking speeds [5]

4. DYNAMIC TIME WARPING

Quranic speech signals have different length of signals in time domain. Time normalization/alignments is used to normalize the recitation signals to get the same length of the signals for speaker adaptation. After that, the comparison at same segment of expert's and learner's Quranic can be done based on Quranic Speech Model (QSM). In [12] defined QSM is the combination of process recording with guided materials, compensation, frame segmentation and warping. In this paper, Dynamic time warping will be used as non-linear sequence alignment for Quranic speech signals. Illustration of two series of Quranic speech signal that using a warp function of DTW is shown on figure 2.

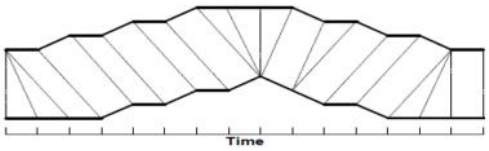

Fig.2. Illustration of two series of Quranic speech signal using warp function

Recitation speech signal of learner will be warped with respect to Quranic speech signal of expert by using a DTW algorithm [14][13]. Quranic speech signal sequences of expert reference X with lengths |X| and learner Y with length |Y| in (2)

$$X = x_1, x_2, ..., x_i, ..., x_{|X|} \quad \text{and} \quad Y = y_1, y_2, ..., y_j, ..., y_{|Y|} \tag{2}$$

Construct the warp path , W

$$W = w_1, w_2, ..., w_K \qquad \max(|X|, |Y|) \le K < |X| + |Y| \tag{3}$$

Where K is the length and kth element of the warp path is $w_k = (i, j)$

Where i = index of time series of X and j = index of time series of Y. The optimal warp path is presented as 'local match' scores matrix by getting the lowest-cost distance warp path and the first frame optimal warp start at $w_1 = (1,1)$ where the distance of a warp path is given by

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj}) \tag{4}$$

$Dist(W)$ is the cosine distance of warp W, and $Dist(w_{ij}, w_{kj})$ is the distance between the two data frame indexes of X and Y in the kth element of the warp path. The lowest-cost path for the first frame is $D(1,1) = 0$ and can be calculate by [6][13][2]

$$D(i, j) = Dist(i, j) + \min[D(i-1, j), D(i, j-1), D(i-1, j-1)] \quad (5)$$

The minimum –cost alignment is determined from optimal warp path that end at $D(|X|,|Y|)$ by looking at lowest-cost warp path. The search grid is used one-to-one monotonic transformation of the time axis [7] which all the movements have equal weight [8].

5. EXPERIMENT DESIGN

The Quranic recitation are recorded in studio and transform into Quranic recitation speech signals that are clearly from the unwanted audio signals such as noise, surrounding and unpredictable audio sound. The format of audio data recorded are wav format with 16bits, 44100 samples [9] and mono channel. Expert and learner are selected from Malay community to recite Al-Quran. Expert's recitation are recorded from qualified teachers of Pusat Islam Universiti Teknologi Malaysia. They also conducted Al-Quran Clinic for Al-Quran recitation evaluation class to assist the learners. The common styles of Qiraat hafs [10] are used and 3 selected surah such as surah Al-Fatihah, Al-Alaq and Ad-Dhuha in rasm uthmani version are provided in this recording session. The step of warping [15] between the template expert and learner input of recitation speech signals.

1. Load two speech waveforms of the same utterance/word/ayah.
2. Calculate short-time Fourier transform (STFT) features for both sounds (25% window overlaps).
3. Construct the 'local match' scores matrix as the cosine distance between the STFT magnitudes.
4. Use dynamic programming to find the lowest-cost path between the opposite corners of the cost matrix.
5. Find the cost of minimum-cost alignment of the two recitation speech signals.

6. Calculate the frames in template that are indicated to match each frame in learner input to resynthesize a warped.
7. Interpolate Learner's recitation speech signal STFT under the time warp.
8. Invert learner recitation back to time domain.

6. COMPENSATION AND PREPROCESSING

The loss of information in Quranic speech signals is also depending on variations of the reciter. The complexity of variability will be compensated to gain the clean speech signals from each ayah of Quranic speech signal without losing the important features. The main aims of pre-processing are to select techniques should be able to represent the transformation speech production signals to Quranic speech raw signals with variability compensation. In the silent trimming technique, the selected threshold is used to remove the silent. Therefore, the amplitude normalization is used to normalize the reciter's speech signals. While the end point detection is used to define the start point and end point of Quranic speech signals. Each of learners or experts have different start point and end point while do recitation. Combined zero crossing and short term energy function are used to determine start point and end point based on paper [11]

7. EXPERIMENT AND RESULTS

The standard DTW is basically using the idea of deterministic Dynamic programming (DP)[15][13] to find the shortest path algorithm. Dynamic Time Warping (DTW) is well-studied as non-linear sequences alignment algorithm. It seeks an optimal mapping from the learner recitation speech input signal to the expert recitation speech template signal. Although DTW are allowing non-linear alignment it also allow monotonic distortion (warping).

Expert 1 recitation of Al-Fatihah- ayah 02

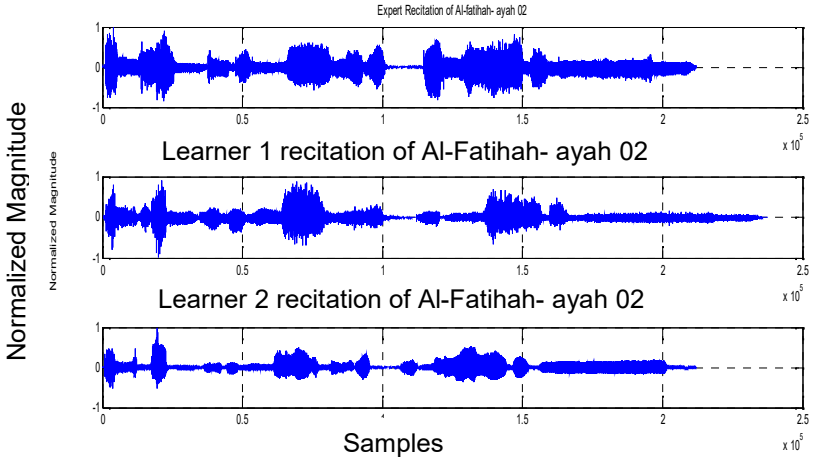

Fig.2. Different length of recitation between expert 1, learner 1 and learner 2 of Al-Fatihah (ayah 02)

The red line is an optimal cost path from the beginning to the end of both expert and learner recitation signal. Figure 3 shows DTW path in similarity matrix, which denotes the correlation of two recitation speech signals. Hence the path will tend to pick darker blocks since it will maximize the matching performance. Note that the minimizing the distance is identical to maximizing the similarity. Figure 4 shows the matrix at minimum cost to arrive with the same DTW path. Optimal DTW path in similarity matrix ottom right as it becomes darker since the cost is monotonically increasing meanwhile optimal DTW path is taken as lowest-cost as possible.

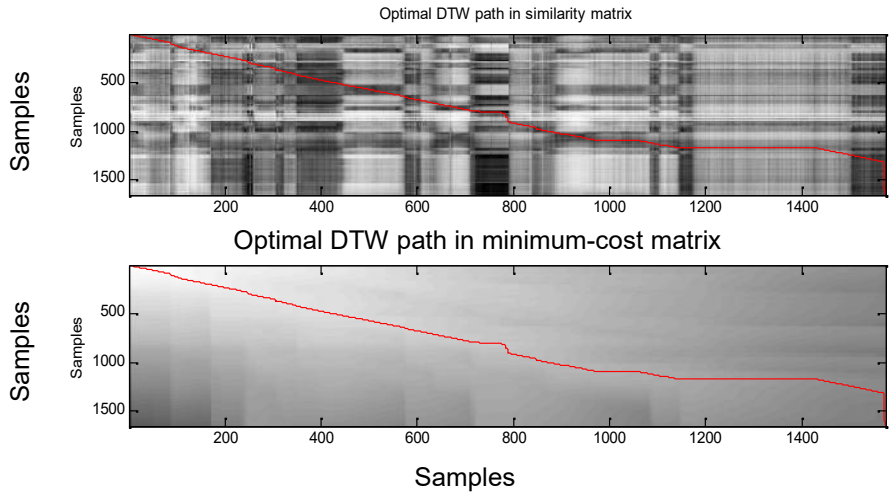

Fig.3 and 4. DTW between expert 1 and learner 2 of Al-Fatihah (ayah 02)

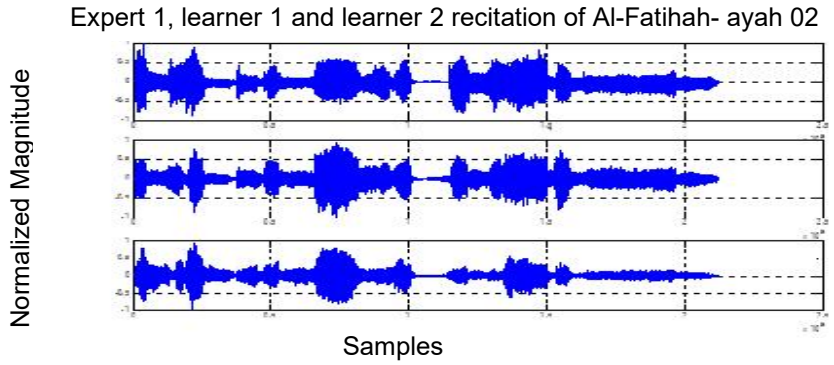

Fig.5. Same length time series of recitation of expert 1, Learner 1 and Learner 2 for Al-Fatihah-ayah 02

8. MEAN SQUARE ERROR

The mean square error is used to determine the success of the warping between expert template and learner. This technique is used for samples frame matching, where it measures the average of the squares [3] of the errors for each frame. This technique is used to determine how close a regression line samples frame a set of points between expert and learner recitation speech signals. The smaller the means squared error value, the closer to find the best fit between two recitation speech signals. Mean squared error performance function with fraction between 0 and 1 indicating the proportion of performance attributed to

scalar weight. The equation of MSE is given as $MSE = \frac{\left\| x - x_{ref} \right\|^2}{N_s}$, where MSE = a scalar value, x = sequences sampled of learner recitation, x_{ref} = sequences sampled of expert recitation and N_s = number of samples

Table 1. The comparison of each early 10 frame DTW (ayah 02 Al-Fatihah) recitation between expert 1 and learners before time warping

Segment by frame	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
L1 (% MSE)	0.59	2.5	2.03	3.35	9	7.31	1.35	0.72	1.72	2.5
L2 (% MSE)	0.005	0.61	3.8	16.45	13.29	6	0.73	0.56	1.48	2.16
L3 (% MSE)	0.17	0.77	1.9	7.58	7.07	0.88	0.22	0.87	1.68	2.05

Table 2. The comparison of each early 10 frame DTW (ayah 02 Al-fatihah) recitation between expert 1 and learners after time warping

Segment by frame	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
L1 (% MSE)	0.68	3.61	2.04	0.99	0.42	0.45	0.28	0.4	0.43	0.88
L2 (% MSE)	0.56	3.46	2.36	1.05	0.53	0.47	0.31	0.4	0.46	0.74
L3 (% MSE)	0.6	3.97	2.73	2.28	0.37	0.37	0.26	0.31	0.36	0.71

9. CONCLUSIONS

Automated recitation evaluation is important as a guided tool to enhance the Al-Quran recitation to the Malay community and the lives of Muslims in particular. With the experience of the experts, the learner can improve their Al-Quran recitation. To compare recitation between expert and learner, new approaches has been develop to compensate the variability for differences recitation especially speaker adaptation techniques. In terms of speaker adaptation, the pronunciation of Al-Quran recitation mostly affected especially related to the word articulation and phoneme utterance that occurred at different time series. By using DTW, all the utterance and articulation of Al Quran recitation can be represented

by the same point on time series in formant frequency segment frame to all reciters. In others word, by using DTW time series segmentation alignment, the recitation between expert and learner can be compared in the same frame and produce the same energy and prosody weight in time series for each utterance or phonetic in the recitation to evaluate the performance between learners and experts recitation.

10. FUTURE WORKS

After the length of recitation between experts and learners recitation signals are same. It is possible to evaluate the recitation by processing the recitation speech signals to the features vector such as cepstral co-efficient or probabilistic features vector that can be derive from speech production. Most of modern speech recognition or pronunciation evaluation used cepstral coefficient as acoustic features vector. The acoustic features can be derived from the speech parameterization based on formant frequency. Formants frequency are the resonant frequencies in the vocal tract which form the characteristic shape of the speech spectrum. Formant-like features can be used as acoustic model for Al-Quran recitation evaluation. Acoustic model of Al-Quran recitation can be aligned in same frame which representing the energy, rhythm and tone. The formant-like features can be represented the Makraj and Tajweed at same frame and will be modelled by Gaussian Mixture model as a static pattern based on formant frequencies and as sequential pattern based on rhythm and tone as modelled by Hidden Markov Model.

11. ACKNOWLEDGEMENTS

The research was supported by Aroma Research Center, Universiti Teknologi Malaysia

12. REFERENCES

12.1. Journal Article

[1] L. Muda, M. Begam, and I. Elamvazuthi. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing.2, 3 **(2010)**

[2] L. Muda, M. Begam, and I. Elamvazuthi. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW)

Techniques. Journal of Computing. **(2010)**

[3] N. W. Arshad, S. M. Sukri, L. N. Muhammad, H. Ahmad, R. Hamid, F. Naim, and N. Z. A. Naharuddin, Makhraj Recognition for Al-Quran Recitation using MFCC. International Journal of Intelligent Information Processing.**(2013)**

12.2. Conference Proceedings

[4] T. Gunawan and M. Kartiwi. Development of high quality speech compression system for Quranic recitation based on modified CELP algorithm. Proceeding of the IEEE International Conference on Smart Instrumentation , Measurement and Application **(2013)**. November.

[5] B. J. Mohan and R. Babu. Speech Recognition using MFCC and DTW. Proceeding of the International Conference on Advances in Electrical Engineering.**(2014)**

[6] V. N. Truong. Vietnamese speech recognition using Dynamic Time Warping and Coefficient of Correlation. Proceeding of the 2013 International Conference on Control, Automation and Information Sciences **(2013)**

[7] S. M. Jayanthi, Divide-and-Warp Temporal Alignment of Speech Signals between Speakers:Validation Using Articulatory Data. Proceeding of the IEEE International Conference on Acoustic, Speech and Signal Processing.**(2017)**

[8] E. Signal and P. Conference. Query by Example Search with Segmented Dynamic Time Warping for Non-Exact Spoken Queries. Proceeding of the 23rd European Signal Processing Confererence **(2015)**

[9] T. S. Gunawan and M. Kartiwi. On the Characteristics of Various Quranic Recitation for Lossless Audio Coding Application. Proceeding of the 6th International Conference on Computer and Communication Engineering: Innovative Technologies to Serve Humanity. **(2016)**

[10] S. F. Ishak, Z. M. Zaki, K. A. Mohamad, M. A. M. Bahrin, N. H. A. Roni, and M. A. Musa. MyQiraat: An interactive Qiraat mobile application. Proceedings of the 6th International Conference on Computer and Communication Engineering: Innovative Technologies to Serve Humanity. **(2016)**

[11] N. N. Lokhande, N. S. Nehe, and P.Vikhe. Voice activity detection Algorithm for Speech Recognition Applications. Proceeding of the International Conference in

Computational Intelligence **(2011)**

12.3. Others

[12] N. Shafie, M. Z. Adam, and H. Abas, The model of Al-Quran recitation evaluation to support in Da'wah Technology media for self-learning of recitation using mobile apps. 3rd International Seminar on Da'wah, National University of Malaysia.**(2017)**

[13] C. Fang. From Dynamic Time Warping (DTW) to Hidden Markov Model (HMM). University of Cincinnati. **(2009)**

[14] Y. X. Luo J. The speech evaluation method of English phoneme mobile learning system. IEEE Workshop on Advanced Research and Technology In Industry Application. **(2014)**

[15] D. Ellis, "Dynamic Time Warp (DTW) in Matlab," Online Web Resources. **(2003)** http//www.ee.columbia.edu/~dpwe/resources/matlab/dtw.

How to cite this article:
Shafie N, Adam M Z, Abas H. Al-quran recitation speech signals time series segmentation for speaker adaptation using dynamic time warping. J. Fundam. Appl. Sci., 2018, *10(2S), 126-137.*