International Journal of Speech Technology (2018) 21:343–354
https://doi.org/10.1007/s10772-018-9511-z

Adaptive framing based similarity measurement between time warped speech signals using Kalman filter

Wasiq Khan[1] · **Keeley Crockett**[1] · **Muhammad Bilal**[2]

Received: 21 November 2017 / Accepted: 9 April 2018 / Published online: 17 April 2018

Abstract
Similarity measurement between speech signals aims at calculating the degree of similarity using acoustic features that has been receiving much interest due to the processing of large volume of multimedia information. However, dynamic properties of speech signals such as varying silence segments and time warping factor make it more challenging to measure the similarity between speech signals. This manuscript entails further extension of our research towards the adaptive framing based similarity measurement between speech signals using a Kalman filter. Silence removal is enhanced by integrating multiple features for voiced and unvoiced speech segments detection. The adaptive frame size measurement is improved by using the acceleration/deceleration phenomenon of object linear motion. A dominate feature set is used to represent the speech signals along with the pre-calculated model parameters that are set by the offline tuning of a Kalman filter. Performance is evaluated using additional datasets to evaluate the impact of the proposed model and silence removal approach on the time warped speech similarity measurement. Detailed statistical results are achieved indicating the overall accuracy improvement from 91 to 98% that proves the superiority of the extended approach on our previous research work towards the time warped continuous speech similarity measurement.

Keywords Adaptive speech segmentation · Speech processing · Dynamic time warping · Spoken term detection · Kalman filter

1 Introduction

There has been a steady improvement in speech signal matching methods for past three decades. The variety of methods has been devolved from isolated word matching to continuous speech recognition (Akila and Chandra 2013; Lawrence et al. 1989; Pour and Farokhi 2009; Olivier 1995). In the template based speech modelling, recognition is performed by matching the test word (utterance) with the all stored template of words and calculating the matching score based on acoustic features (Akila and Chandra 2013). Dynamic Time Warping (DTW) and Vector Quantization (VQ) based speech recognition is the best examples of such systems. As the speech signal contains dynamic features, development of a robust speech signal matching approach is the challenging task. Noisy speech, time warping phenomenon, and connected words and phonemes in continuous speech are common examples of such dynamics that make the speech similarity measurement task more challenging. Among these issues, time warping in speech signals has been a challenge to deal with. Multiple speech recordings having the same contents (i.e. words and phonemes) by a speaker may produce different time durations. Consequently, the devolution of the time warping factor to sub-word and hence phonemes level degrades the similarity matching performance (Wasiq and Rob 2015).

Most of the recent improvements are made in the related area in terms of Spoken Term Detection (STD) that is based on the partial information extraction (keyword) from a continuous speech signal (Timothy et al. 2009; Chun-An and Lin-Shan 2013; Wasiq and Kaya 2017; Anguera et al. 2013).

✉ Wasiq Khan
W.Khan@MMU.ac.uk

Keeley Crockett
K.Crockett@MMU.ac.uk

Muhammad Bilal
M.Bilal@UCLA.edu

[1] School of Mathematics, Computing and Digital Technology, Manchester Metropolitan University, Manchester M15 7XP, UK

[2] Institute of the Environment and Sustainability, University of California, Los Angeles, USA

Similarly, Query-by-Example (QbyE) methods, keyword/filler methods, and large vocabulary continuous speech recognition methods are some other approaches that represent some sort of similar work (Anguera et al. 2014; Joho and Kishida 2014; Tejedor et al. 2013; Javier et al. 2015). Over the past decade, most of the STD based research is focused on novelty of template representation methods (Abad et al. 2013; Marijn et al. 2011; Haipeng et al. 2011). Recent work introduced in (Abad et al. 2013) addresses the fusion of heterogeneous STD system. In the first step, a number of heuristics are hypothesized for the similarity score estimation and then a linear logistic regression method is used for the combination of these scores. The performance is measured using eight different techniques individually as well as by fusing them together using linear regression. Similarly, Wasiq and Kaya (2017) proposed an effective approach for STD based on acoustic segment models. This method amalgamates the self-organizing models, query matching, and query modelling processes to construct an efficient STD approach. Despite the effective performances of aforementioned approaches for STD, a further improvement is needed to deal with time warping issues in continuous speech similarity measurement as presented in the proposed research.

The literature contains a number of approaches in relation to QbyE and STD that use some sort of variations in DTW (Yaodong and James 2011a, b; Chan and Lee 2010; Thambiratmann and Sridharan 2007). However, in DTW based approaches, the computation time is linear to the number of frames (i.e. signal length) to be searched through (Cheng-Tao et al. 2014). Extensive efforts were made to enhance the DTW performance in terms of computation time such as segment-based DTW proposed by Chun-An and Lin-Shan (2011), lower-bound estimation for DTW (Yaodong and James 2011a, b; Yaodong et al. 2012), and a locality sensitive hashing technique for indexing speech frames presented by Jansen and Van Durme (2012). Speech matching presented by Chotirat and Eamonn (2005) reported the lack of ability of the conventional DTW to deal with the time warping phenomenon. Likewise, a signal dependent word recognition system is presented by Yegnanarayana and Sreekumar (1984) where an enhanced DTW is proposed based on a weight factor. The query speech signal is partitioned into voiced, unvoiced and silence segments using the weight factor resulting better performance for the time warped speech matching. Despite extensive research efforts, existing approaches are unable to handle the time warping phenomenon robustly because of static frame length. Alternatively, time warped distance measurement between test and template frames may be improved using a varying frame size corresponding to dynamically changing speed of spoken words. A state estimator is needed to dynamically predict the query frame position in reference speech pattern. The Kalman filter (KF) seems a good candidate for such a state estimation by modeling variable speed and noise covariance in an effective way. The KF is a recursive state estimator with diverse application areas that include object tracking, navigation systems, multi-sensor data fusion, control systems, manufacturing, noise reduction in signal, and free-way traffic modeling Mohinder and Angus (2001).

This manuscript propose a time warped similarity measurement approach in extension to our previous study Wasiq and Rob (2015). A number of significant improvements are made in terms of silence removal by integrating multiple approaches for voiced and unvoiced detection, state modeling and continuously varying frame size using acceleration phenomenon of object linear motion, offline tuning of the KF to retrieve the optimal parameters, enhanced feature extraction, evaluation methods and detailed statistical analysis of the results as discussed in following sections. As speech signals are analyzed frame by frame, each frame of test speech is considered as an individual unit moving along the time-axis of reference speech with a certain speed. The object linear motion is modeled to estimate the test and reference frames corresponding positions within the test and reference speech signals respectively. Simultaneously, feature based best matched reference frame position corresponding to query frame is also calculated. Both position estimates are then forwarded to a KF along with the noise covariance which recursively predicts the final position estimate for query frame. At each time step, template frame size changes according to the acceleration/deceleration calculated using the KF state estimate.

2 Methods and materials

2.1 Speech corpuses and dataset

A variety of proprietary and open source speech datasets are used to conduct experiments in the proposed research. The dataset consists of recorded speeches as short sentences, isolated utterances, long phrases and paragraphs; that were acquired from different genders, age groups, and ethnic background people. To conduct a case study, we have recorded a speech dataset of 50 speakers (37 male, 13 female) that consists of connected words in the form of digits (five recordings for each digit by each speaker), short phrases of up to 10 s (five sentences by each speaker) and long phrases of up to 20 s (five paragraph bay each speaker). For recording purpose, the SENNHEISER e935 is used which is a vocal dynamic microphone that consists a built in noise filter. The dataset is recorded in a noise free research lab environment. Although, the proposed approach is purely based on acoustic features without the transcribed data, however; to prove the concept of language independence, the dataset is recorded for multiple languages that include English, Arabic and Urdu. For the long speech phrase tracking

experiments; a speech corpus from American Rhetoric's (top 100 speeches) (Michael 2013) is used. It is based on hours of speeches recorded by different people on different topics. Moreover, two speech corpuses, Mobio (McCool et al. 2012) and Wolf (Hung and Chittaranjan 2010) are obtained from IDIAP research institute which consist of huge amount of speech data recorded by different speakers.

2.2 Pre-processing and search window

A number of techniques are amalgamated in the proposed method to deal with silence removal, frame size adaptation, and time warping challenges. In the first step, speech signals are forwarded to a pre-processing unit to enhance the quality in terms of silence removal as presented in Fig. 1. The silence removal is composed of two different approaches for voiced, unvoiced and silence segments detection. A robust pitch tracking method proposed by Zahorian and Hu (2008) is used to estimate the fundamental frequency (F_0) using multiple information resources.

Acquired speech is segmented and forwarded to the pitch tracking algorithm proposed by Zahorian and Hu (2008) that searches for the existence of F_0 components in each segment of input speech. As the F_0 doesn't exist in the silence part of speech, these frames can be eliminated. All frames having the F_0 components are produced as 'voiced' segments. For the 'unvoiced frame detection' energy and Zero Cross Rate (ZCR) features are used as proposed by Sharma and Rajpoot (2013) shown in Fig. 2.

Output 'voiced and unvoiced' frames produced from aforementioned approaches are combined together to reconstruct a silence free speech signal which is used for further processing. Figure 3 shows the sequential steps used for the silence segments removal and reconstruction of the silence free speech signal. The silence free speech signals are then forwarded to a speech framing process that recursively selects a fixed length frame and forwards it for further processing until the end of test or reference speech. Because of the slowly varying nature of the speech signal, it is common to process speech in blocks (also called "frames") of 10–50 ms over which the speech waveform can be assumed as a stationary signal (Ravindran et al. 2010). In the proposed approach, the test speech signal is partitioned into 30 ms frames and forwarded to feature extraction as shown in Fig. 1.

Mel Frequency Cepstral Coefficients (MFCC) has been used as the most dominant features of human speech (Dave 2013; Dhingra et al. 2013; Shahzadi and Azra 2013; Ezzaidi and Jean 2004) are extracted for test and reference frames and a Euclidean distance based degree of similarity is calculated between the means of MFCCs vectors of test and reference frames. A search window is defined within the reference speech and the query frame is coasted by overlapped steps to measure the likelihood at each step as shown in Fig. 4. The length of search window is set twice to template frame size '*w*'. As a result, the maximum likelihood estimate (MLE) of the position is calculated from the similarity distribution. For each overlapped frame shift *i* (i = 1, 2 ...*n* and *n* = number of shifts), the likelihood positions $f(l_i)$ are calculated using:

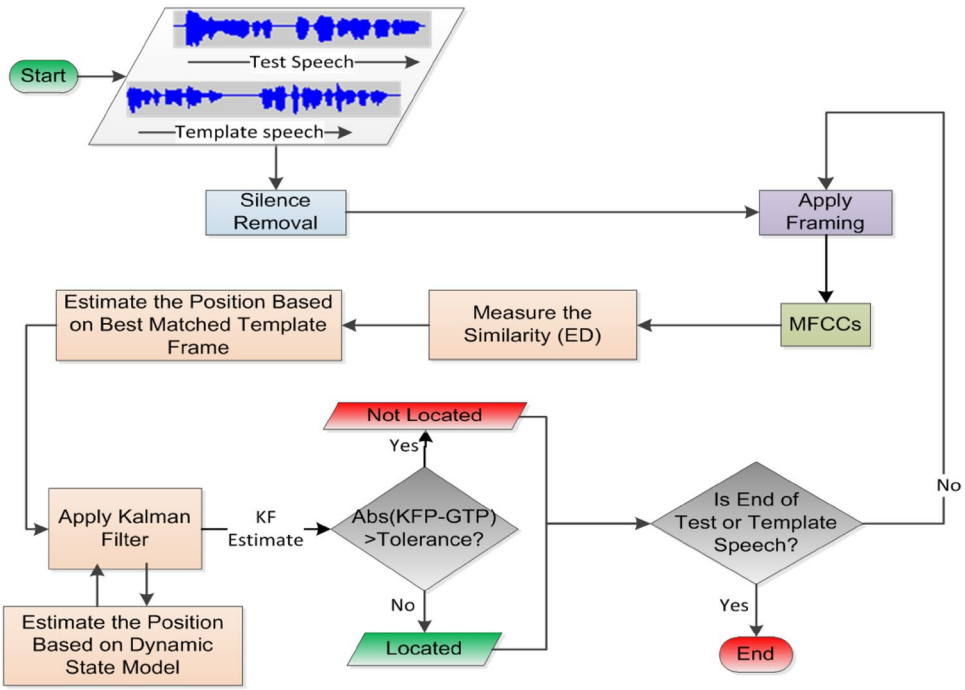

Fig. 1 Overall sequential processing of proposed speech signal matching using state model and Kalman filter. *GTP* ground truth position, *KFP* Kalman filter estimated position

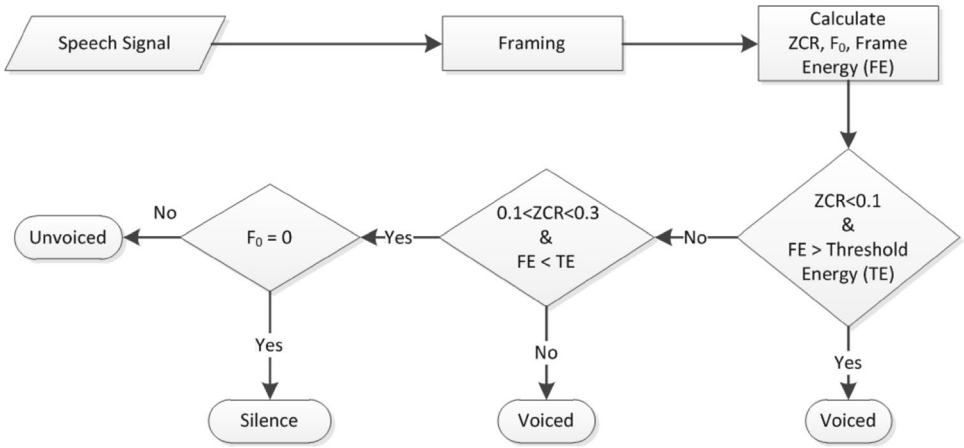

Fig. 2 Silence removal process using Energy, ZCR and F_0 (Sharma and Rajpoot 2013)

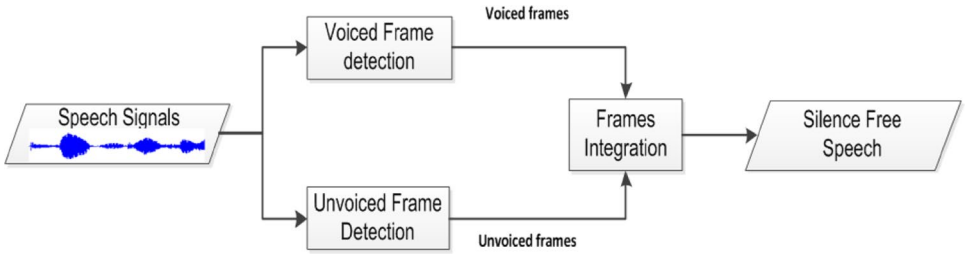

Fig. 3 Block diagram for silence removal from speech signal using pitch tracking and time domain features

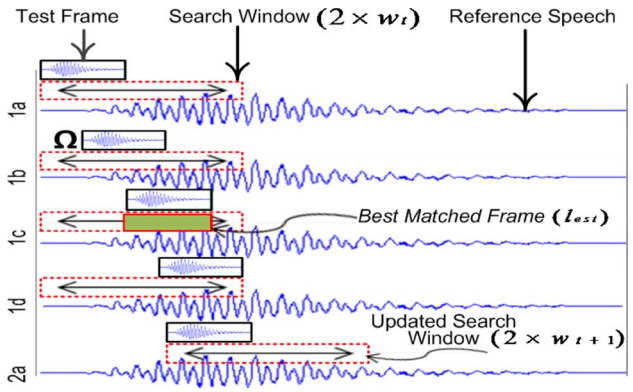

Fig. 4 Overlapped progression (1a–1d) of query frame along the reference speech pattern and search window. Second cycle starts from 2a and process repeated until end of reference speech signal

$$f(l_i) = l_{cur} + i(w - \Omega) \tag{1}$$

where

$$l \in \begin{Bmatrix} [l_{cur} - w/2, l_{cur} + w/2] & \text{if } l_{cur} - w/2 \geq 1, \\ & l_{cur} + w/2 \leq L \\ [l_{cur} - w/2, l_{cur}] & \text{if } l_{cur} + w/2 > L \end{Bmatrix}$$

where 'L' is the total length of search window, and l_{cur} is the current position (position before the MLE calculation) of test frame in the search window while considering the boundary constraints 'l' for each frame shift. The 'Ω' is overlap interval within the search window. In order to obtain a normalized probability distribution, we used:

$$\Phi_l = \frac{\Phi_{l_i}}{\sum_{i=1}^{n} \Phi_{l_i}} \tag{2}$$

where Φ_l is the vector containing similarities scores measured by Euclidean distance for $f(l_i)$ with 'n' elements.

$$l_{est} = MLE = \arg \max_i \Phi_{l_i} \tag{3}$$

Position of the best matched reference frame is calculated using the 'l_{est}' and 'f_l' vectors which is then used by KF as input parameter for observed position at current state.

2.3 Linear motion based state model and Kalman filter

The proposed method for a speech similarity measure uses the recursive process of KF to estimate the instantaneous positions of test speech frames in the reference speech patterns. At each time step, the query frame position is estimated using a linear motion model which is fused with the position observed by aforementioned feature based similarity measure. Thus the process of linear motion based test frame position estimate is needed to be described by a linear system such that:

$$x_{t+1} = Ax_t + Bu_t + f_t \tag{4}$$

$$z_t = Hx_t + g_t \tag{5}$$

In above equations, 'x' represents the state at time 't' 'A' is the state matrix, 'B' is input matrix, and 'H' is the output matrix. Dynamically changing size of reference frame is a known input 'u' to the system. System output is represented by 'z' along with the process noise 'f' and measurement noises 'g'. In terms of speech frame matching, time warping phenomenon causes the process noise generation. Following these facts, the linear motion model is implemented for query frame position estimate. Let 'v_t' and 'v_{t+1}' represent the initial and final velocities of query frame progression along the reference speech pattern at time 't' and 'Δw' is the difference between successive reference frames sizes then, velocity at current state will be:

$$v_{t+1} = v_t + \Delta wT \tag{6}$$

In the start, sample rate (8000 samples/s) of query frame is set to 'v_t' which is updated according to 'Δw' recursively and 'T' represents the static interval (i.e. 30 ms) of query frame. Position of the query frame in reference speech pattern is then represented as:

$$p_{t+1} = p_t + Tv_t + \frac{1}{2}\Delta wT^2 \tag{7}$$

where 'p_{t+1}' is the query frame position at current time. Equation 7 indicates the query frame position dependency on varying template frame size 'Δw' which reflects the time warping phenomenon in speech signals. Combining the above set of equations (Eqs. 4, 5, 6, and 7) the linear motion system can be represented as:

$$x_{t+1} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} x_t + \begin{bmatrix} T^2/2 \\ T \end{bmatrix} \Delta w + f_t \tag{8}$$

$$z_t = \begin{bmatrix} 1 & 0 \end{bmatrix} x_t + g_t \tag{9}$$

An efficient position and velocity estimation for the query frame would produce a feedback control system which is achieved by incorporation of the KF. The process and measurement noises are assumed to be independent variables which are set by the offline tuning of KF using a Receiver Operating Characteristics (ROC) curve points discussed later (Sect. 2.3.2). The process and measurement noise covariance matrices can be represented as:

$$\begin{aligned} Q =& E(f_t f_t^T) \\ =& E\left(\begin{bmatrix} \mathrm{p} & \mathrm{v} \end{bmatrix} \begin{bmatrix} p \\ v \end{bmatrix} \right) = E\left(\begin{bmatrix} p^2 & pv \\ vp & v^2 \end{bmatrix} \right) \\ =& f_t \times \begin{bmatrix} \frac{T^4}{4} & \frac{T^3}{2} \\ \frac{T^3}{2} & \mathrm{T} \end{bmatrix} \end{aligned} \tag{10}$$

$$R = E(g_t g_t^T) = E(g_t^2) \tag{11}$$

where 'Q' represent the standard deviation in the estimated state (i.e. position and velocity) and 'R' represents the standard deviation in the measurement of noise covariance. The query frame state predicted by aforementioned setup (Eqs. 8, 9, 10, and 11) is forwarded to the KF update (correction) step which corrects the estimate by means of 'Q', 'R', Kalman gain 'k', and state estimate 'x'. Mathematically, the update step can be modeled as:

$$k_t = P_t^- H^T (HP_t^- H^T + R)^{-1} \tag{12}$$

$$\hat{x}_t = \hat{x}_t^- + k_t(l_{est(t)} - H\hat{x}_t^-) \tag{13}$$

$$P_t = (I - k_t H)P_t^- \tag{14}$$

where

$$P_t^- = AP_{t-1}A^T + Q; \quad P_t = (I - k_t H)P_t^-;$$

$$A = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}; H = [1\ 0];$$

$$Q = E(f_t * f_t^T); R = E(g_t^2)$$

Equations 12, 13, and 14 represent the correction step in KF which uses the manipulation of matrices. Test frame position estimate by feature based similarity measure 'l_{est}' and predicted position by linear motion model 'x' at current time 't' are used in Eq. 13 to get an updated state estimate. The Kalman gain 'k' in Eq. 12 shows the dependency of the state estimate upon noise covariance. The entire setup (predict and correct) runs recursively to measure the current test speech frame position within the search region of reference speech pattern while adapting the template frame size 'Δw' at each time step as described in Sect. 2.3.1.

2.3.1 Frame size and search region adaptation

Once the position for the current state is estimated by KF model, the template frame size is updated according to the calculated difference 'Δw' using:

$$\Delta w = \frac{v_{t+1} - v_t}{T} \tag{15}$$

$$w_{t+1} = w_t + \Delta w \tag{16}$$

where 'w_{t+1}' is the estimated template frame size for next time step corresponding to acceleration/deceleration amount ('Δw') at current time which represents the time warping phenomenon. Figure 5 presents an example of the recursive frame size adaptation phenomenon in the proposed speech similarity measurement approach. It can be analyzed that

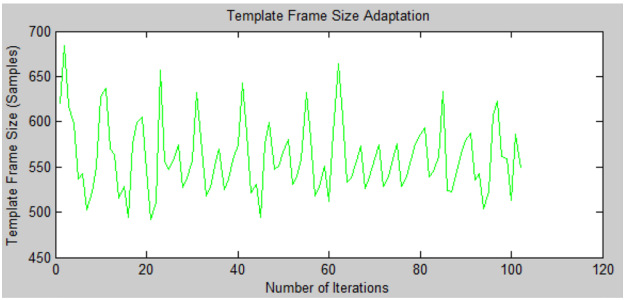

Fig. 5 Reference speech frame size adaptation at each time step for corresponding time warping phenomenon in test speech

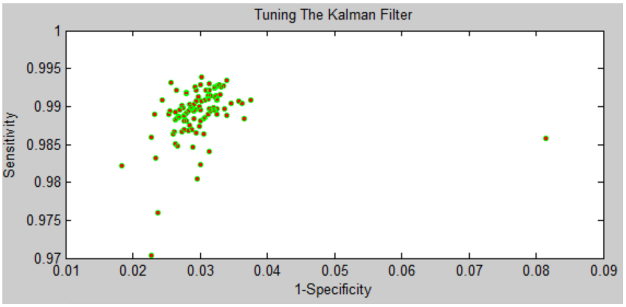

Fig. 6 KF tuning in terms of noise variance selection based on best compromise between sensitivity and specificity

the template frame size changes instantly using Eq. 16 and reflects the natural time warping phenomenon in a speech signal.

2.3.2 Tuning the Kalman filter

Practically, the measurement noise covariance 'R' can be measured prior to filter operation. This is because during the filter operation, as the process is needed to be measured, the process noise covariance 'Q' measurement can be a challenge. Generally speaking, a robust filter performance may be achieved by tuning the filter parameters 'Q' and 'R'. Most of the time; this tuning is performed offline, frequently with the help of another (distinct) KF in a process generally referred to a system identification (Mohinder and Angus 1993; Greg and Gary 2006). It can be analyzed from measurement update that in case of constant values of Q and R, estimation error covariance P_t and Kalman gain k_t will stabilize quickly and then remain constant.

This means that these parameters can be computed prior to filter operation by offline tuning or by determining the steady state values. Variations in the values of 'Q' and 'R' indicate the dependency (level of trust) of the system. Greater value for a variance means less dependency on the corresponding measure and vice versa. In the proposed method, values for measurement and process variances are validated using the ROC curve points that are retrieved by varying them from 0 to 1 with a lag of 0.01 as shown in Fig. 6. The entire setup is tested on sample speech dataset described earlier and the best values for process and measurement noise variances are selected based on the best compromise between sensitivity and specificity.

2.4 Experimental setup and performance evaluation

To achieve the optimal performance in terms of time-warped speech matching, a number of factors are set by iteratively analyzing the experimental results and updating the setup values. These factors consist of KF tuning, recording devices, processing tools, and recording environment. Table 1 shows the simulation settings for the experimentation of speech tracking performance. Because of the template frames overlapping, a tolerance of half frame size for the matching decision is set throughout the experiment conduction. The evaluation methodology entails experiments for multiple settings that involve KF variables setting, silence removal approaches, and similarity measurement methods.

A number of metrics have been used in the literature for the validation of query term similarity measurement. However, the most relevant are the gold standards used for the performance evaluation of a binary classifier (Soluade 2010). This is because the output of test and reference speech frames is in the binary form (i.e. match or mismatch). Table 2 presents the detailed metrics that are used for the validation of the proposed speech similarity measurement approach.

3 Results and discussion

Experimental results for proposed approach and state-of-the-art segmented DTW based approach are achieved using sample dataset described earlier. The performance difference between KF based adaptive framing and the search window based non-adaptive approach is presented using the gold standard metrics addressed in Table 2. It is observed that the similarity matching and speech tracking performance degrades while using the non-adaptive framing. This is because in contrast to static frame size, the adaptive framing handles the time warping phenomenon better way. Also, use of the KF and linear motion model provide the substitute tracking information that never loses the tracking path when a mismatch or false positive occurs. This proves the reliability of the proposed approach as compared to segmented DTW method as well as our previous approach (Wasiq and Rob 2015) that are considered as a baseline in our experiments. In the proposed approach, the concept of

Table 1 Initial variables setting of Kalman filter and motion based state model in the proposed method

Hardware specifications	
Processor: Intel® Core™ i5 CPU Installed memory: 4 GB System type: 32 bit operating system OS: Window 7 Home Premium H-Disk: 500 GB Microphone: SENNHEISER e935	
Simulation tools	Matlab R2009a, PRAAT, SFS, Audacity
Sampling frequency	8000 Hz
Initial frame size	240 samples
Overlap amount	50%
Search region	2 × frame size = 480 samples
Tolerance for position estimate	½ template frame
Kalman filter variables	
T	0.03 s
p_t	0
p_{t+1}	240 (30 ms)
Δw	0 at 1st step then changes dynamically
f_t	0.72
g_t	0.28

Table 2 Statistical metrics used for performance evaluation of the proposed approach

Condition as determined by gold standard		
Total population	Condition positive	Condition negative
Positive match	True positive	False positive
Negative match	False negative	True negative
Accuracy (ACC) = (Σ true positive + Σ true negative)/Σ total population	True positive rate (TPR), sensitivity, recall = Σ true positive/Σ condition positive	(Type I error), false positive rate (FPR) = Σ false positive/Σ condition negative
	(Type II error), false negative rate (FNR) = Σ false negative/Σ condition positive	True negative rate (TNR), Specificity (SPC) = Σ true negative/Σ condition negative
F1 score = 2 × precision × recall/(precision + recall)	Positive likelihood ratio (LR+) = TPR/FPR	Negative likelihood ratio (LR−) = FNR/TNR

acceleration/deceleration amount (Eq. 15, 'Δw') from the motion model produces a continuously varying template frame size as shown in Figs. 4 and 5. This phenomenon makes the entire system more effective as compared to our previous study (Wasiq and Rob 2015) for which 'Δw' was updated based on estimated frame position. Thus, overall accuracy is increased from 91 to 98.01% that indicates the significance of acceleration based 'Δw' in the proposed method.

A substantial decrement of 27% in the sensitivity and 32% in tracking accuracy of the proposed approach is observed in Table 3 when adaptive framing is changed with non-adaptive frame size. Likewise, 15 and 25% decrements in the performances of DTW in terms of sensitivity and tracking accuracy respectively, is measured. Consequently, the desired speech tracking accuracy is significantly affected and is unable to maintain the tracking path information because of mismatches and false negatives. This proves the novelty of the dynamic frame size concept as introduced in this research study.

The likelihood ratios (LR+, LR−) are considered one of the useful metrics to measure the diagnostic accuracy. In terms of test and template frames matching, LR presents the probability of a test with test frame match divided by the probability of the same test with test frame mismatch. Larger LR+ consist more information than smaller LR+. On the other hand, smaller LR− consist more information than larger LR−. To simplify the LR values, a relative magnitude is considered by taking the reciprocal of LR+. It can be analyzed from Table 3 that the LR− for KF is negligible (0.008) as compared to 0.12 for DTW which indicates the robustness of the proposed approach. Similarly, F-score is

Table 3 Statistical results for proposed KF based adaptive/static framing and segmented DTW approaches

Evaluation metrics	KF approach	Segmented DTW approach
Adaptive framing		
Sensitivity	0.9918	0.8999
Specificity	0.9726	0.8702
Matching accuracy	0.9801	0.8958
1/LR+	0.0276	0.1578
LR−	0.0084	0.1229
F-score	0.9713	0.8487
Tracking accuracy (%)	1	0.9484
Avg. execution time (s)	1.3747	3.0419
Type I error		
μ	0.0011	0.0332
σ	0.0016	0.0960
Type II error		
μ	3.1270e-04	0.0282
σ	8.8200e-04	0.0735
Non-adaptive framing		
Sensitivity	0.7299	0.7517
Specificity	0.9856	0.9421
Matching accuracy	0.9415	0.9121
1/LR+	0.0182	0.0909
LR−	0.2717	0.2560
F-score	0.8957	0.8698
Tracking accuracy (%)	0.6822	0.6991
Type I error		
μ	0.0011	0.0047
σ	0.0016	0.0052
Type II error		
μ	3.1270e-04	0.1462
Σ	8.8200e-04	0.2237

a measure that considers both precision and recall to measure the system performance. Table 3 demonstrates that the F-score value for KF based approach is increased to 0.97 as compared to 0.84 of the DTW method. Type I and Type II errors indicate the recognizer failure rates related to FP and FN respectively. These metrics have been represented in a number of ways in the related area including mean square error and absolute erros as most commonly used. Table 3 shows the 'μ' (mean) and 'σ' standard deviation for both types of error for different approaches while using the same speech dataset.

Computation time is also an important factor that is analyzed for the proposed KF based approach and DTW approaches. Table 3 shows the average computational costs for these approaches that demonstrate the variations in computational time for varying lengths of test and reference speech signals. It is observed that the KF based approach have a lead over the conventional approaches. This is because of computation time in DTW is linear with the number of frames to be searched through (Cheng-Tao et al. 2014). Despite of the fact that constrained DTW reduces the average time complexity to 1.6 s, yet its computation time is greater than the proposed approach (1.3 s) that measures the Euclidean distance based on selected features.

3.1 Decision boundary selection

The trade-off between true positive hits and true negative rejection rate is based on the threshold value that is used as a decision boundary for test and template frames match/mismatch. Figure 7 presents a three dimensional relation showing the true positive rate, false positive rate and threshold values. To set a threshold value for the match/mismatch decision boundary, an ROC curve is achieved by varying the threshold from 0 to 1 with a lag of 0.01. It means that the template frame was rejected if its matching score with the

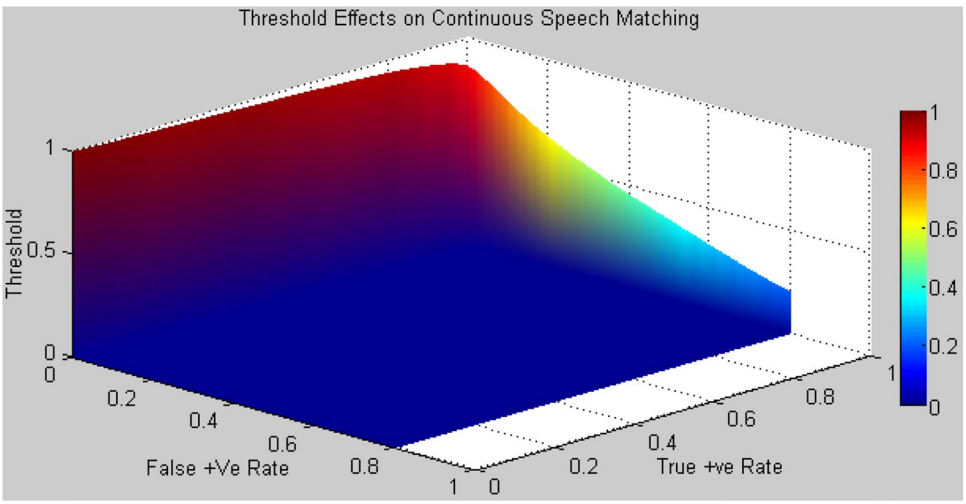

Fig. 7 Measuring the optimal threshold value for query and reference frames match/mismatch decision using ROC curve

corresponding test frame is less than the threshold value. The best threshold value (0.85) for the proposed approach is selected based on a best compromise between sensitivity and specificity as shown in Fig. 7.

In addition to adaptive framing, appropriate selection of a silence removal approach also affects the performance. There are a number of techniques in the literature that use time and frequency domain features (e.g. Energy, zero cross rate, spectral centroid) and pattern recognition methods to remove the silence part of speech utterance (Sharma and Rajpoot 2013; Tushar et al. 2014; Sen and Graduate 2006; Liscombe and Asif 2009; Saha et al. 2005; Giannakopoulos 2014). However, in the proposed work, a more effective approach for silence removal is used that combines a robust pitch tracking algorithm for the voiced segment detection and a ZCR and energy based approach for the unvoiced segment detection. Figure 8 shows the performance statistics for different silence removal approaches. It can be observed that the sensitivity, specificity, frames matching accuracy, and speech tracking accuracy are decreased 2, 5, 3, and 2% respectively using traditional silenced removal approach as compared to the proposed approach that use multiple resources to remove the silence part of speech.

Figure 9 demonstrates a test case for a time warped speech signal matching using the speech data presented in Fig. 10. The concept of adaptive framing and search window is shown in three-dimensional representation of time warped speech similarity measurement. Intensity of the color in

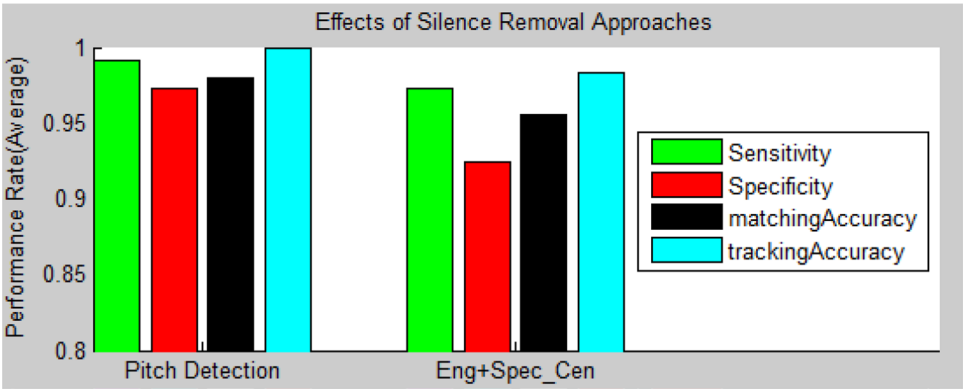

Fig. 8 Statistical performance analysis of the time warped speech signal matching using conventional and proposed silence removal approaches

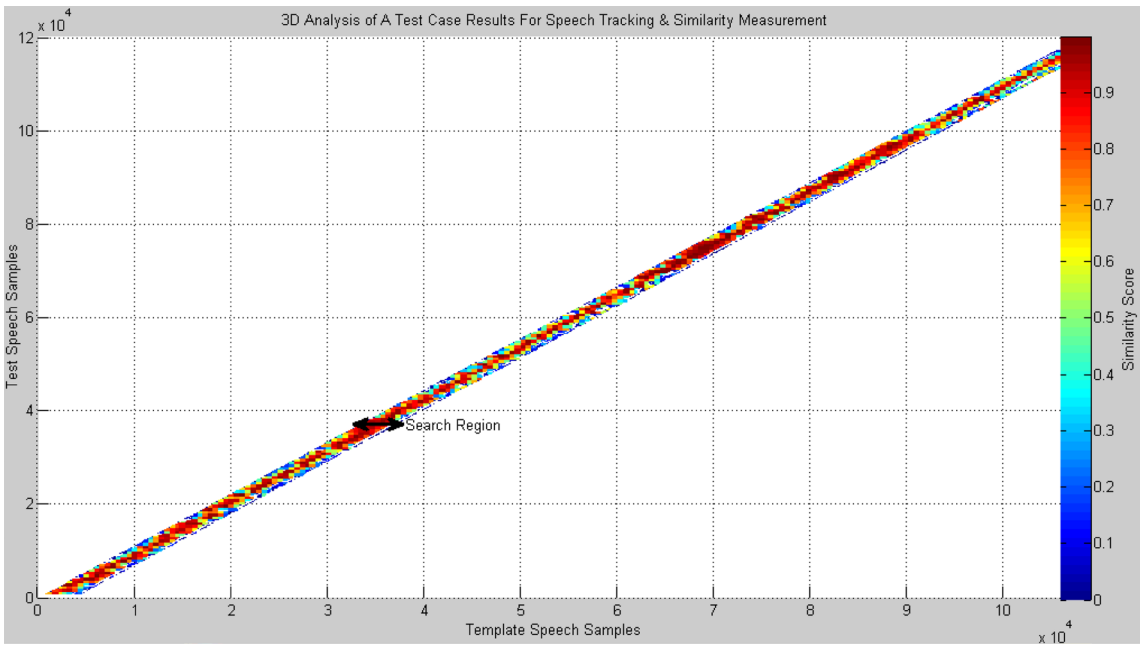

Fig. 9 Three-dimensional representations of frame size adaptation in proposed speech similarity measurement for test case data in Fig. 10. For 10 k samples of reference speech, the varying width of vertical entries (tiles) indicates frame level time warping; leading to produce overall time warped path based on corresponding frames in test speech of 12 k samples

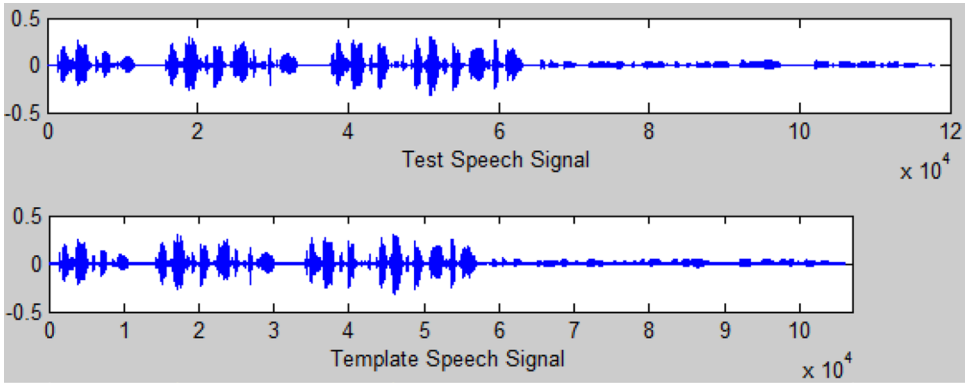

Fig. 10 Speech data used in Fig. 9: test and reference speech signals acquired from same speaker, with same spoken contents but different speaking rate

Fig. 9 from blue to red indicates the match/mismatch score respectively. The dynamics in the template frame size can be analyzed in Fig. 9 that indicates the adaptive framing with respect to the varying speaking rate of input speech. It is also observed that there exists at least one frame in the search window of the reference speech pattern exceeding the match/mismatch decision threshold that indicates the robustness in sensitivity leading to consistency in the tracking path. In results, the proposed approach is able to deal the time warping phenomenon in an efficient way. Thus, for 120,000 samples of test speech in contrast to 106,000 samples of reference speech, KF based adaptive framing provided the robust similarity matches between test and reference speech frames without losing the tracking path.

4 Conclusions and future directions

In general, speaking rate variations produce the time warped speech that affects the performance of similarity measure. This manuscript entails further extension of our previous research (Wasiq and Rob 2015) towards the adaptive framing based continuous speech signal matching which deals with the time warping phenomenon more efficiently. Deployment of acceleration/deceleration based frame size adaptation and standard MFCC features produced significant improvements in our previous model. Likewise, the offline tuning of the KF produced the optimistic parameters for the statistical model. Silence removal approach is improved by integrating multiple algorithms for voiced and unvoiced segment detection in continuous speech signal. Experiments demonstrated that the proposed approach outperforms our previous approach and other existing methods in terms of speech similarity matching. Although, the existing enhanced version of DTW combines multiple distance matrixes to generate more reliable decision boundary however; this approach is unable to recover the path for a continuous speech signal matching/tracking in case of mismatches. This issue is resolved in the proposed approach by using a dynamic state model that deploys the equations of motion to predict the temporal information for the next state. The use of Kalman filter makes it more robust in terms of match/mismatch frame position estimation.

There is always an uncertainty in the model (process) which indicates the error in process and the aim is to minimize this error. The beauty of KF is that it recursively updates the states according to noise covariance which plays multiple roles. Firstly, it measure the quantity of error exists in terms of process and measurement noise. Secondly, the noise variance provides the degree of dependency on both observations. In other words, noise covariance assigns initial weights to each observation on the basis of which a KF tunes itself recursively. On the other hand, noise variance is not considered in the existing time warping techniques which can deal better with process and measurement uncertainties. There are some possible extensions that can be considered as future research direction. For instance, system reliability can be further improved using multiple information resources (e.g. similarity measure, feature set) which can be fused together using theory of evidence.

Acknowledgements A special gratitude I give to Prof. Daniel Neagu whose contribution in stimulating suggestions helped me to coordinate this manuscript in layout, statistical analysis, and performance evaluation methods. The author also thanks to Prof. Ping Jiang for his time and support to discuss several alternatives to deal with time warping issue.

References

Abad, A., Rodriguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., & Bordel, G. (2013). On the calibration and fusion of heterogeneous spoken term detection systems. *Conference of the International Speech Communication Association, Interspeech*, France, 25–29 August 2013.

Akila, A., & Chandra, E. (2013). Slope finder—A distance measure for DTW based isolated word speech recognition. *International Journal of Engineering and Computer Science, 2*(12), 3411–3417.

Anguera, X., Metze, F., Buzo, A., Szoke, I., & Rodriguez-Fuentes, L. J. (2013). The spoken web search task. In *Proceedings of MediaEval* (pp. 1–2), Aachen, Germany: CEUR Workshop Proceedings.

Anguera, X., Rodriguez-Fuentes, L. J., Szoke, I., Buzo, A., & Metze, F. (2014). Query by example search on speech. In *Proceedings of MediaEval* (pp. 1–2). Spain

Chan, C.-A., & Lee, L. S. (2010). Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. In *Proceedings of Interspeech* (pp. 693–696). Prague

Cheng-Tao, C., Chun-an, C., & Lin-Shan, L. (2014). Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity. *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 7814–7818), 4–9 May 2014. https://doi.org/10.1109/ICASSP.2014.6855121.

Chotirat, R., & Eamonn, K. (2005). Three myths about dynamic time warping data mining. In *The Proceedings of SIAM International Conference on Data Mining* (pp. 506–510).

Chun-An, C., & Lin-Shan, L. (2011). Unsupervised hidden markov modeling of spoken queries for spoken term detection without speech recognition. In *Proceedings of Interspeech* (pp. 2141–2144).

Chun-An, C., & Lin-Shan, L. (2013). Model-based unsupervised spoken term detection with spoken queries. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(7), 1330–1342.

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology, 1*(6), 1–4.

Dhingra, S., Nijhawan, G., & Pandit, P. (2013). Isolated speech recognition using MFCC and DTW. *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering, 2*(8), 1–8.

Ezzaidi, H., & Jean, R. (2004). Pitch and MFCC dependent GMM models for speaker identification systems. *Canadian Conference on Electrical and Computer Engineering* (Vol. 1, pp. 43–46).

Giannakopoulos, T. (2014). *A method for silence removal and segmentation of speech signals, implemented in Matlab, 2014*. Retrieved May 13, 2014 from http://cgi.di.uoa.gr/~tyiannak/Software.html.

Greg, W., & Gary, B. (2006). An introduction to Kalman Filter. TR 95-041. Course 8. Chapel Hill: University of North Carolina at Chapel Hill.

Haipeng, W., Tan, L., & Cheung-Chi, L. (2011). Unsupervised spoken term detection with acoustic segment model. In *IEEE Proceedings of the International Conference on Speech Database and Assessments* (Oriental COCOSDA) (pp. 106–111).

Hung, H., & Chittaranjan, G. (2010). *The Idiap wolf corpus: Exploring group behaviour in a competitive role-playing game*. Florence, Italy: ACM Multimedia. Retrieved January 27, 2011 from http://homepage.tudelft.nl/3e2t5/mmsct22567-hung.pdf.

Jansen, A., & Van Durme, B. (2012). *Indexing raw acoustic features for scalable zero resource search*. In *Proceedings of Interspeech*

Javier, T., Doroteo, T. T., Paula, L., Laura, D., Carmen, G., Antonio, C., Julian, D., Alejandro, C., Julia, O., & Antonio, M. (2015). Spoken term detection ALBAYZIN 2014 evaluation: Overview, systems, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing, 21*, 1–27.

Joho, H., & Kishida, K. (2014). Overview of the NTCIR-11 SpokenQuery&Doc task. In *Proceedings of NTCIR-11* (pp. 1–7). Tokyo, Japan: National Institute of Informatics (NII).

Lawrence, R. R., Jay, G. W., & Frank, K. S. (1989). High performance connected digit recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*(2), 1214–1225.

Liscombe, M., & Asif, A. (2009). A new method for instantaneous signal period identification by repetitive pattern matching. In *IEEE 13th International Multitopic Conference*, INMIC (pp. 1–5).

Marijn, H., Mitchell, M., & David, V. L. (2011). Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 4436–4439).

McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matějka, P., Černocký, J., Poh, N., Kittler, J., Larcher, A., Lévy, C., Matrouf, D., Bonastre, J., Tresadern, P., & Cootes, T. (2012). *Bi-modal person recognition on a mobile phone: Using mobile phone data*. IEEE ICME Workshop on Hot Topics in Mobile Multimedia.

Michael, E. (2013). *Top 100 speeches, American Rhetoric, 2001*. Retrieved December 12, 2013 from http://www.americanrhetoric.com/top100speechesall.html.

Mohinder, S. G., & Angus, P. A. (1993). *Kalman filtering: Theory and practice*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Mohinder, S. G., & Angus, P. A. (2001). *Kalman filtering: Theory and practice using MATLAB* (2nd ed., pp. 15–17). New York: Wiley).

Olivier, S. (1995). On the robustness of linear discriminant analysis as a pre-processing step for noisy speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, 9–12 May 1995 (Vol. 1, pp. 125–128).

Pour, M. M., & Farokhi, F. (2009). An advanced method for speech recognition. *International Scholarly and Scientific Research & Innovation, 3*(1), 840–845.

Ravindran, G., Shenbagadevi, S., & Salai, S. V. (2010). Cepstral and linear prediction techniques for improving intelligibility and audibility of impaired speech. *Journal of Biomedical Science and Engineering, 3*(1), 85–94.

Saha, G., Sandipan, C., & Suman, S. (2005). A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. In *Proceedings of the NCC*.

Sen, Z., & Graduate, S. (2006). An energy-based adaptive voice detection approach. *8th International Conference on Signal Processing* (Vol. 1). Beijing: Chinese Academy of Science

Shahzadi, F., & Azra, S. (2013). Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization. *International Conference on Computer, Control & Communication (IC4)* (pp. 1–5).

Sharma, P., & Rajpoot, A. K. (2013). Automatic Identification of silence, unvoiced and voiced chunks in speech. *Journal of Computer Science & Information Technology (CS & IT), 3*(5), 87–96.

Soluade, O. A. (2010). Establishment of confidence threshold for interactive voice response systems using ROC Analysis. *Communications of the IIMA, 10*(2), 43–57.

Tejedor, J., Toledano, D. T., Anguera, X., Varona, A., Hurtado, L. F., Miguel, A., & Colas, J. (2013). Query-by-example spoken term detection ALBAYZIN 2012 evaluation: Overview, systems, results, and discussion. *Journal on Audio, Speech, and Music Processing, EURASIP, 23*, 1–17.

Thambiratmann, K., & Sridharan, S. (2007). Rapid yet accurate speech indexing using dynamic match lattice spotting. *IEEE Transactions on Audio, Speech and Language Processing, 15*(1), 346–357.

Timothy, J. H., Wade, S., & Christopher, W. (2009). Query-by-example spoken term detection using phonetic posteriorgram templates. In

IEEE Proceedings of the Automatic Speech Recognition & Understanding (ASRU) Workshop, 17 December 2009 (pp. 421–426).

Tushar, R. S., Ranjan, S., & Sabyasachi, P. (2014). Silence removal and endpoint detection of speech signal for text independent speaker identification. *International Journal of Image, Graphics and Signal Processing, 6*, 27–35.

Wasiq, K., & Kaya, K. (2017). An intelligent system for spoken term detection that uses belief combination. *IEEE Intelligent Systems, 32*(1), 70–79.

Wasiq, K., & Rob, H. (2015). Time Warped continuous speech signal matching using Kalman filter. *International Journal of Speech Technology, 18*(1), 1381–2416.

Yaodong, Z., & James, R. G. (2011a). A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping. In *Proceedings of Interspeech* (pp. 1909–1912).

Yaodong, Z., & James, R. G. (2011b). An inner-product lower-bound estimate for dynamic time warping. In *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5660–5663).

Yaodong, Z., Kiarash, A., & James, G. (2012). Fast spoken query detection using lower-bound dynamic time warping on graphical processing units. *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5173–5176).

Yegnanarayana, B., & Sreekumar, T. (1984). Signal dependent matching for isolated word speech recognition system. *Journal of Signal Processing, 17*(2), 161–173.

Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *Journal of Acoustic Society of America, 123*(6), 4559–4571.