Intelligent system for spoken term detection using the belief combination

Wasiq Khan[1], Kaya Kuru[2]

School of Engineering, University of Central Lancashire, PR1 2HE, United Kingdom

Emails: [1]wkhan4@uclan.ac.uk; [2]kkuru@uclan.ac.uk

Abstract

Spoken Term Detection (STD) can be considered as a sub-part of the automatic speech recognition which aims to extract the partial information from speech signals in the form of query utterances. A variety of STD techniques available in the literature employ a single source of evidence for the query utterance match/mismatch determination. In this manuscript, we develop an acoustic signal processing based approach for STD that incorporates a number of techniques for silence removal, dynamic noise filtration, and evidence combination using Dempster-Shafer Theory (DST). A 'spectral-temporal features based voiced segment detection' and 'energy and zero cross rate based unvoiced segment detection' are built to remove the silence segments in the speech signal. Comprehensive experiments have been performed on large speech datasets and consequently satisfactory results have been achieved with the proposed approach. Our approach improves the existing speaker dependent STD approaches, specifically the reliability of query utterance spotting by combining the evidences from multiple belief sources.

Keywords: Spoken term detection, Acoustic keyword spotting, Query-by-example, Dempster-Shafer's theory, Speech recognition, Speech processing.

Acknowledgement: A special gratitude we give to Prof. Daniel Neagu, University of Bradford, whose contribution in stimulating suggestions helped us to coordinate this research in terms of statistical analysis, and performance evaluation methods.

I. Introduction

There is a long-standing interest in STD with regard to both theoretical and practical issues. Nowadays, it is receiving much importance due to the large volume of multimedia information. Research and technology improvements in automated speech recognition successfully achieved the information retrieval by using the transcribed textual form of the spoken contents [1]. Similarly, due to the exponential growth of internet and multimedia contents, the STD methods have been achieving much popularity. However, dynamic properties of speech signal make the STD task more challenging. Literature contains a variety of STD techniques that use different approaches to match the query utterance with reference speech. Template matching based utterance spotting has been recently proposed as one of the most commonly used methods [2]. For instance, speech recognition using Vector Quantization (VQ) and Dynamic Time Warping (DTW) models is the most relevant

example of these systems. However, there are some challenges associated with DTW approach that are needed to be resolved [3], [4], and [5].

In relation to acoustic keyword spotting, Query-by-Example (QbyE) methods, keyword/filler methods, and large vocabulary continuous speech recognition methods have also been used in the literature. Most of the existing QbyE methods [4], [6], [7], [8], [9], [10] and STD approaches [11], [12], [13], [14] use DTW and its variations [3], [15], [16]. Over the past decade, mass of the related research is focused on novelty of the template representation methods [17], [18], [19], [20], [21]. An acoustic segmentation model based STD is presented in [10] that amalgamates the self-organising models, query matching, and query modelling processes. Similarly, [14] introduced a template combination based STD method that deploys segmental DTW and a self-similarity matrix comparison between speech utterances. In addition to QbyE and STD methods, isolated word recognition is also related to STD however; it is less complicated as compared to STD due to the discreteness and isolation of the speech signal. Literature consists of several variations of isolated word matching that exploits different approaches for pattern recognition. For instance, isolated word recognition is presented by [22] where extracted features for test and reference utterances in the form of Mel-Frequency Cepstrum Coefficients (MFCCs) vectors are forwarded to DTW model that measure the warping distance. Similarly, a signal dependent matching for isolated word recognition is proposed in [23] producing a better performance using fast Fourier transform for feature extraction and enhanced version of DTW. Likewise, an improved DTW technique is proposed in [24] based on cross correlation for digit recognition. It uses a new approach of slacked start and end point which depends upon the performance of end point detection.

Despite of fact that the existing methods have been improving the DTW based STD to deal the time warping phenomenon more effectively, the trade-off between distance matrix pruning and DTW performance in terms of warping distance accuracy is still challenging [3], [4], and [5]. The boundary constraints on distance matrix improve the computation cost but sacrifice a significant amount of DTW performance [3]. In addition, the uni-source information used in DTW to measure the warping distance provides an unreliable spotting decision. Because of the unsupervised model of DTW, it would be much better to use multi-source information for distance calculation to make spotting decision which would increase the system reliability. This manuscript introduces a novel STD approach which amalgamates a number of techniques to improve the existing STD template matching based methods. For the first time, a temporal-spectral feature based silence removal is deployed along-with the DST to fuse the evidences from multiple information resources to produce a reliable spotting decision for query utterance. A detailed mathematical formulation of the DST for the proposed task along with the experimental results and performance analysis is presented in the following sections.

II. Material and Methods

The proposed STD methodology entails data collection, mathematical modelling, experimental setup, and analysis of statistical results to evaluate the performance of proposed approach. Experiments are conducted using a large dataset available online as described in Table I. For the long speech phrase STD experiments; two speech corpuses, Mobio [25] and Wolf [26] are requested from IDIAP research institute. These dataset consist of very large-scale spoken contents recorded by variety of speakers as a composition of single, binary, and group discussions. In addition, a case study is conducted on speech dataset acquired from 30 speakers from

diverse ethnic background, age, and gender. The data is recorded in noiseless Lab environment using a vocal dynamic microphone with built-in noise filter (SENNHEISER e935).

Table I: Speech Corpuses used for experimental results and performance evaluation

Corpus name	No. of Speakers	Gender	Length	Availability
Mobio	152	M, F	135 GB	Licence agreement
Wolf	12	M, F	100 GB, 81 hours	Licence agreement
CMU ARCTIC	4	M, F	1150 utterances	Open source
Online Children Stories	65	F	65 stories & poems	Open source
Case Study Dataset	30	M, F	Connected words (500), Short phrases (450), Spoken paragraphs (210)	Authorised user only

A. Formulation of the Spoken Term Detection

A composite of techniques are sequentially combined to build the proposed STD system. The input to the system are query and reference speech utterances which are then processed by a sequence of speech enhancement, framing, feature representation, similarity belief calculation, and probabilistic modelling approaches to make the final decision of query utterance match/ mismatch. Figure I show the workflow for proposed STD approach followed by the detailed formulation of all sub-components.

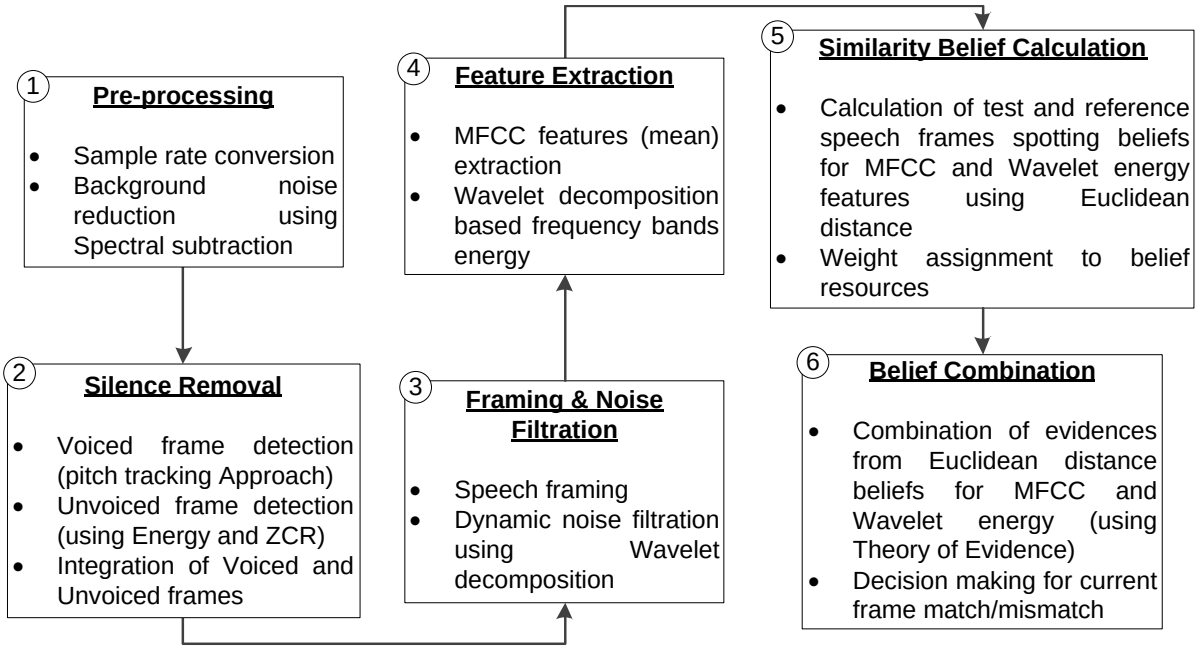

Figure I: Processes workflow in the proposed spoken term detection approach

a) Pre-processing

Existence of silence segments and background noise interference in speech signal cause misidentification and therefore resolved in a pre-processing step. In the first step, background noise is reduced to a minimum level of signal-to-noise ratio by using the spectral subtraction [27] that is performed independently in the frequency bands corresponding to the auditory critical bands. Next step is to remove the silence segments from speech signal. Literature contains several methods for silence removal that are based on signal energy, spectral

centroid, and Zero Cross Rate (ZCR) [28], [29]. For the 'Voiced' segments we used a robust pitch tracking method [30] to estimate the fundamental frequency (F_0) using the temporal-spectral information. 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 ZCR features are used as proposed by [28]. 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. Fig. II shows the sequential steps used for the silence segments removal and reconstruction of the silence free speech signal.

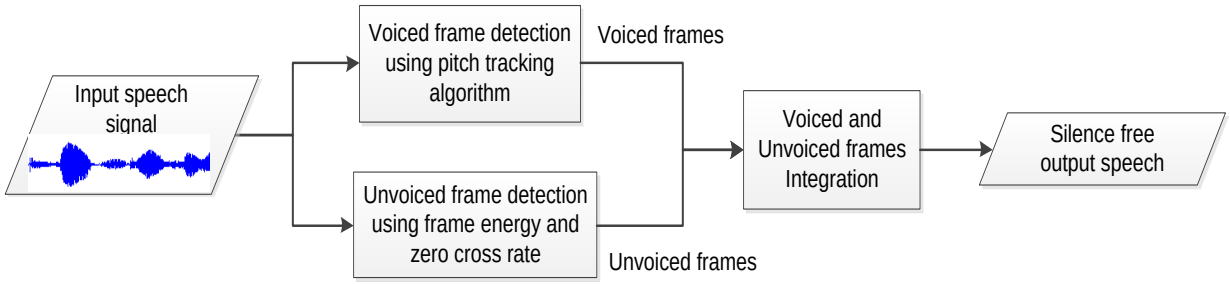

Figure II: Silence removal from speech signal using the spectral-temporal pitch estimation, ZCR, and signal energy

b) Dynamic Speech Filter and Feature Representation

The silence free speech signal is then decomposed into overlapped frames of 30 milliseconds duration and forwarded to a dynamic noise filter that uses the wavelet decomposition to filter out unnecessary frequency bands and temporal information. Wavelet decomposition has successfully been used as a powerful spectral analysis tool which can effectively compress the information about the non-stationary signal into a piece of local information. Moreover, it reveals the scale-wise organization of singularities, thus allowing for the selection of the interesting strongest events using a simultaneous time-frequency domain representation [31].

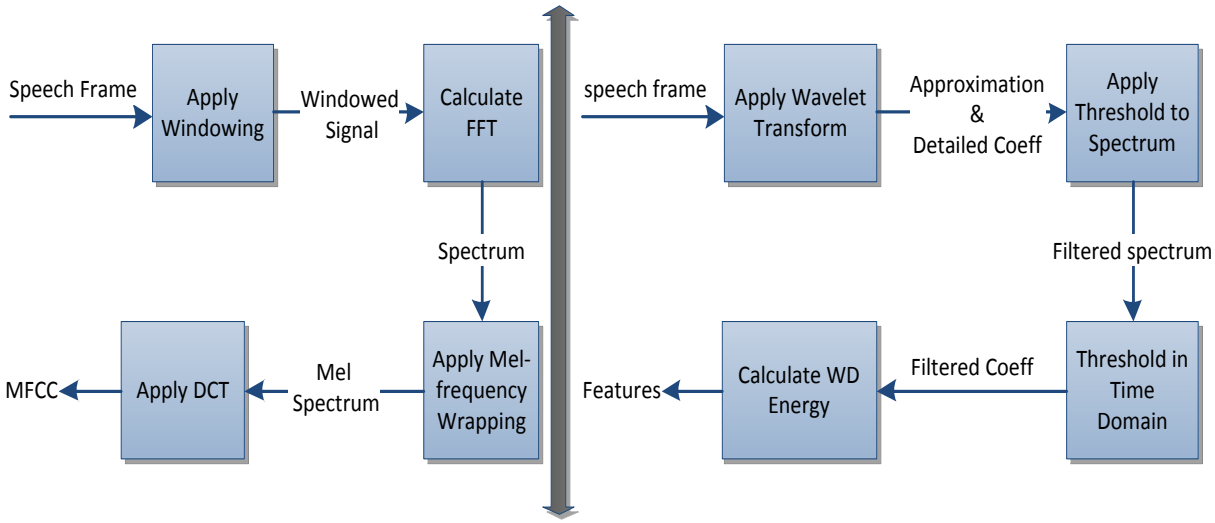

Figure III: Block diagram for MFCC and wavelet decomposition based features extraction

A block diagram for the MFCC and wavelet energy feature extraction process is presented in Fig. III. Energy in a frequency level is measured by integrating the intensity magnitudes over time and can be represented as:

$$E_{Scale} = \frac{\sum_{i=1}^{n} \left(y_i^{Scale} \right)^2}{n} \qquad 1$$

Where 'n' represents the total number of coefficients in a frequency scale, E_{Scale} is the total energy measure for a scale, and 'y' represents the output coefficients produced by wavelet decomposition for a scale. The above procedure is applied to each scale and corresponding energy vector is measured.

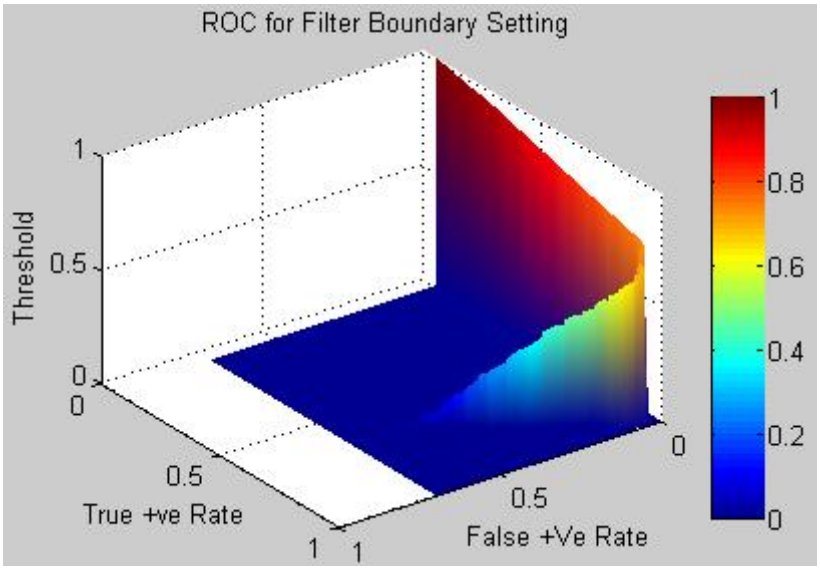

Figure IV: Optimal threshold value selection for dynamic speech filter

Scales containing energy magnitude less than a pre-set threshold are eliminated. The optimal threshold value is chosen by conducting experiments on a large dataset presented in Table I. A Receiver Operating Characteristic (ROC) curve is achieved (Fig.IV) by varying threshold values from 0 to 1 with a lag of 0.01. It is observed that the best compromise between sensitivity and specificity for STD performance is achieved with a threshold value of 0.7. The energy magnitude of noise free approximation and detailed coefficients is then used as one feature set. Simultaneously, the MFCCs features are extracted from the query and reference speech utterance that has been used as the most powerful and distinctive in terms of human speech representation [32, 22]. In the next step, feature vectors (i.e. MFCCs and energy) for query and reference frames are normalised and used by Euclidean distance to measure the degree of similarities between query and reference frames. Output similarity scores from Euclidean distance represent the evidences (i.e. beliefs) provided by MFCCs and energy based features that are further processed by DST for belief combination.

c) Mathematical Formulation of Multiple Belief Combination

The similarity beliefs from previous step are then forwarded to the evidence combination process that uses the DST to provide a combined spotting belief while taking into account the corresponding weights and model incomprehension. One of the interesting advantages of the DST is the model simplicity for the complex multi-layered situations where the system can be decomposed into many layers of simpler states and then the beliefs

can be propagated upwards, combined with sibling layer states to get overall belief. A detailed study on DST advantages, disadvantages, and its application areas is presented in [33].

For the proposed STD method, let $E = \{\mu(mfcc), WDe\}$ represent the set of belief resources in the form of MFCC and wavelet spectral energy. The exclusive assessment classes consist of two elements, i.e. $H = \{match, mis\_match\}$. For each belief resource in 'E' and assessment class'H', a degree of belief β_n is allocated by the Euclidean distance described earlier that indicates the confidence measure when evaluating the degree of fulfilment of a certain feature. The relative weights for belief resources are set by conducting offline experiments (discussed in results section) such that $0 \le \omega_i \le 1$ and $w_{mfcc} = 0.75$ $w_{wav} = 0.25$.

Basic Probability Assignments for Each Belief Resource

Let $m_{n,i}$ represent the basic probability mass indicating the level to which the i^{th} belief resource is assessed. The assumption that the general belief resource is evaluated to the n^{th} assessment class H_n will be:

$$m_{n,i} = \omega_i \beta_{n,i} \tag{2}$$

The remaining probability mass $m_{H,i}$ un-allocated to belief resources can be represented as:

$$m_{H,i} = 1 - \sum_{n=1}^{N} m_{n,i} = 1 - \omega_i \sum_{n=1}^{2} \beta_{n,i} \tag{3}$$

Where, 'N=2' represents the total number of assessment classes and $m_{H,i}$ can be further dissolved into $\bar{m}_{H,i}$ and $\tilde{m}_{H,i}$ as:

$$\bar{m}_{H,i} = 1 - \omega_i \tag{4}$$

$$\tilde{m}_{H,i} = \omega_i \left(1 - \sum_{n=1}^{2} \beta_{n,i} \right) \tag{5}$$

Eq^n. 4 calculates the amount to which final belief resources have not yet been evaluated to separate classes due to the relative significance of belief resources after their aggregation. Eq^n. 5 calculates the amount to which belief resources cannot be evaluated to separate classes due to the imperfect evaluation of belief resources.

Combined Probability Assignments

The next step is to aggregate the probability masses of $E = \{\mu(mfcc), WDe\}$ to compose a combined evaluation for query utterance match/mismatch decision that is formalised by the following equations.

$$\{H_n\}:$$
$$m_{n,i+1} = K_{i+1}[m_{n,i}.m_{n,i+1} + m_{H,i}.m_{n,i+1} + m_{n,i}.m_{H,i+1}]$$
$$n = 1, \ldots, N \tag{6}$$

Where $i = \{1, 2\}$ denotes the number of belief resources and $N = 2$ represents the total count for assessment classes. The term $m_{n,1}.m_{n,2}$ in Eq[n].6 estimates the probability of $E = \{\mu(mfcc), WDe\}$ endorsing the output decision to be evaluated to H_n. The probabilities of $\mu(mfcc)$ and WDe supporting the final decision to be evaluated to H_n are denoted by $m_{n,1}.m_{H,2}$ and $m_{H,1}.m_{n,2}$ respectively.

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \quad (7)$$

$$\tilde{m}_{H,i+1} = K_{i+1}[\tilde{m}_{H,i}.\tilde{m}_{H,i+1} + \bar{m}_{H,i}.\tilde{m}_{H,i+1} + \bar{m}_{H,i+1}.\tilde{m}_{H,i}] \quad (8)$$

$$\bar{m}_{H,i+1} = K_{i+1}[\bar{m}_{H,i}.\bar{m}_{H,i+1}] \quad (9)$$

$$K_{i+1} = \left[1 - \sum_{t=1}^{N=2} \sum_{\substack{j=1 \\ j \neq t}}^{N=2} m_{t,i}.\, m_{j,i+1}\right]^{-1}, \text{ for } i = \{1, \ldots, L-1\} \quad (10)$$

In Eq[n]. 8, $\tilde{m}_{H,1}.\tilde{m}_{H,2}$ estimates the probability of final decision cannot be evaluated to any distinct class $match, mis\_match$ because of the imperfect evaluation for $E = \{\mu(mfcc), WDe\}$. Term $\bar{m}_{H,1}.\tilde{m}_{H,2}$ and $\bar{m}_{H,2}.\tilde{m}_{H,1}$ estimate the probabilities of decision cannot be evaluated because of the imperfect evaluation for $\{WDe\}$ and $\mu(mfcc)$ respectively. In Eq[n]. 9, $\bar{m}_{H,1}.\bar{m}_{H,2}$ estimates the probability of final decision has not yet been evaluated to separate classes because of the relative significance of $\mu(mfcc)$ and $\{WDe\}$ after both belief resources have been integrated. The m_n, m_H are normalised using K as a normalization factor such that $\sum_{n=1}^{N=2} m_n + m_H = 1$

Calculation of the Combined Degree of Belief

Finally, the amount of combined belief β_n for the query utterance evaluation to class H_n can be assessed by integrating the evaluation for all associated belief resources $E = \{\mu(mfcc), WDe\}$ as:

$$\{H_n\} : \beta_n = \frac{m_{n,L}}{1 - \bar{m}_{H,L}} \qquad n = 1, .., N \quad (11)$$

$$\{H\} : \beta_H = \frac{\tilde{m}_{H,L}}{1 - \bar{m}_{H,L}} \quad (12)$$

The degree of belief that remained un-allocated during the evaluations is represented by β_H. The set of equations (Eq[n]. 2 to Eq[n]. 12) provides a combined degree of belief for the evaluation grades that are further used for the decision making of query utterance spotting.

III. Results and Discussions

Performance of the proposed STD approach has been evaluated using different statistical metrics used for binary classification [34, 35] in the form of query utterance match/mismatch decision. These metrics include sensitivity, specificity, accuracy, likelihood ratios, absolute error, execution time, and F-score.

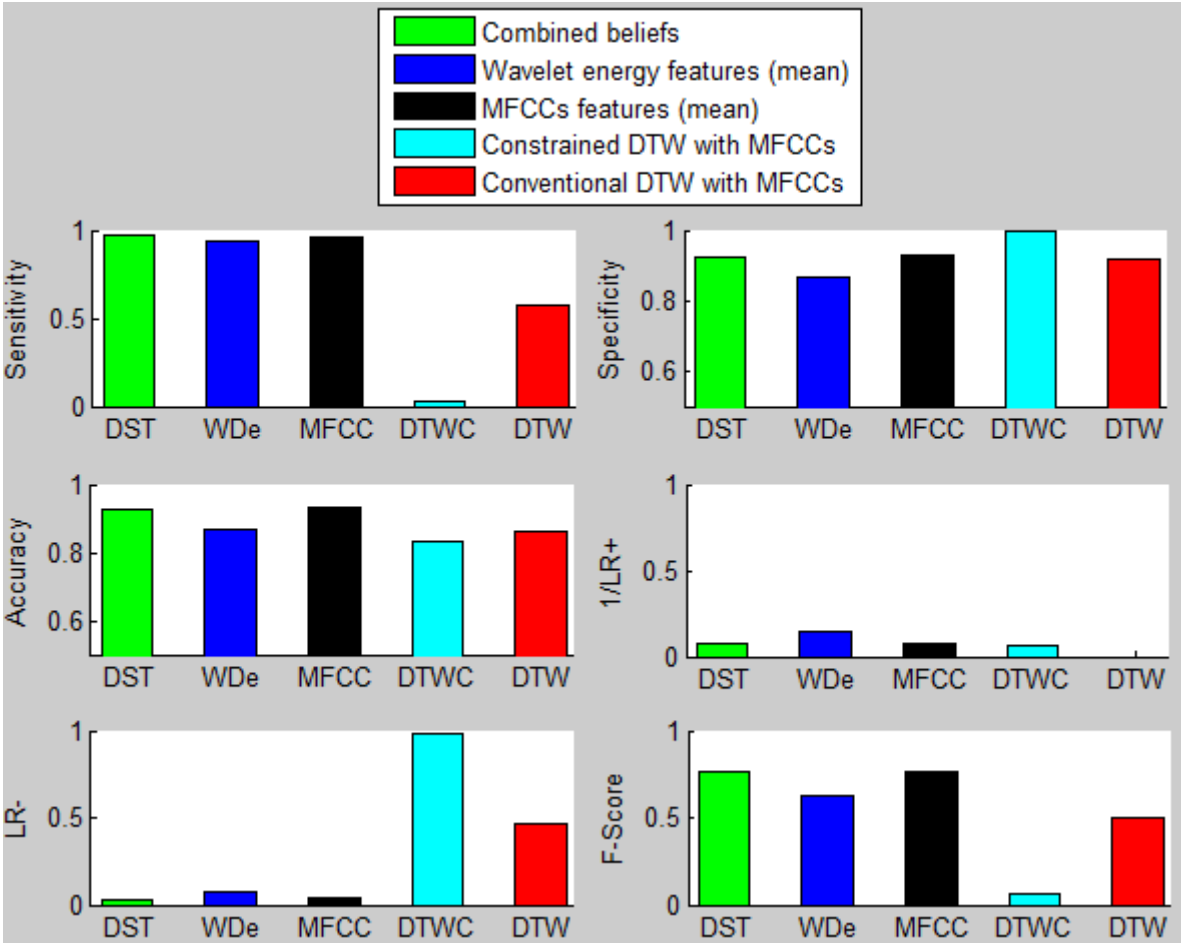

Figure V: Statistical results comparison for spoken term detection by different approaches using case study dataset

Individual performances of the proposed STD approaches are compared with the existing constrained DTW and conventional DTW based STD techniques as shown in Fig.V. It is observed that in terms of sensitivity, the performance of DST based STD is better than the individual performances of MFCC and wavelets based approaches by a factor of 2% and 5% respectively. This implies that the deployment of DST increases the STD as well as it empowers the performance in terms of decision making. Despite of the fact that the search space in DTWC is less than traditional DTW; yet the traditional DTW is better than DTWC in terms of STD outcomes. Similarly, the likelihood ratios (LR+, LR-) are used to measure the diagnostic accuracy which indicate that LR- for the DST based approach is negligible (i.e. 0.03) as compared to 0.2 for DTW and 0.9 for DTWC. These statistics also empowers the superiority of the proposed STD over the existing DTW based STD approaches. Furthermore, the F-score is measured that indicated the effectiveness of proposed STD.

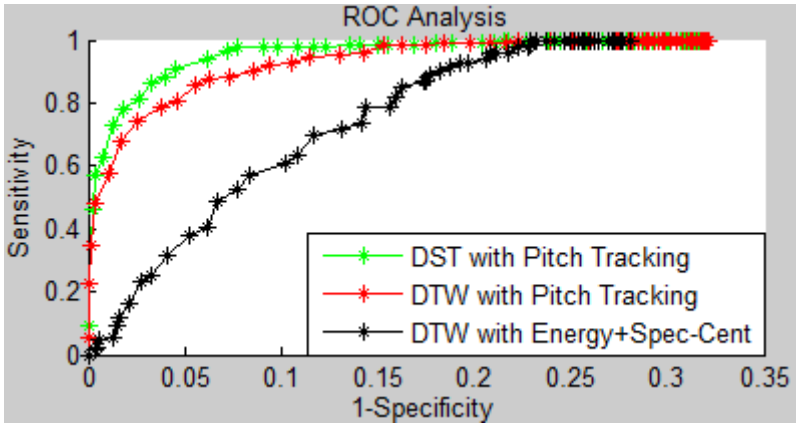

Figure VI: The ROC curves for varying threshold values for query utterance match/mismatch decision

Decision Boundary Selection

The spotting decision for query utterance varies with respect to decision boundary setting. An optimal threshold value for decision boundary is selected using the ROC curve that shows the trade-off between true positive and false positives rate for various threshold settings. Smaller the threshold for decision boundary means higher sensitivity and vice versa. Experiments are conducted on dataset described earlier and ROC curves are achieved by changing threshold scales (from 0 to 1) for various approaches as shown in Fig.VI. It is observed that the decision boundary at 0.85 thresholds value provides the optimistic trade-off between true positive and false positive rate. In addition to optimal threshold value selection, the ROC curves in Fig.VI manifest the superiority of the DST based STD as compared to state of the art DTW. Another aspect of the ROC curves shown in Fig.VI is validation of the silence removal approach introduced in this manuscript. It is clear that area under the curve for energy and spectral centroid based silence removal is far less than proposed pitch detection, ZCR, and energy based approach.

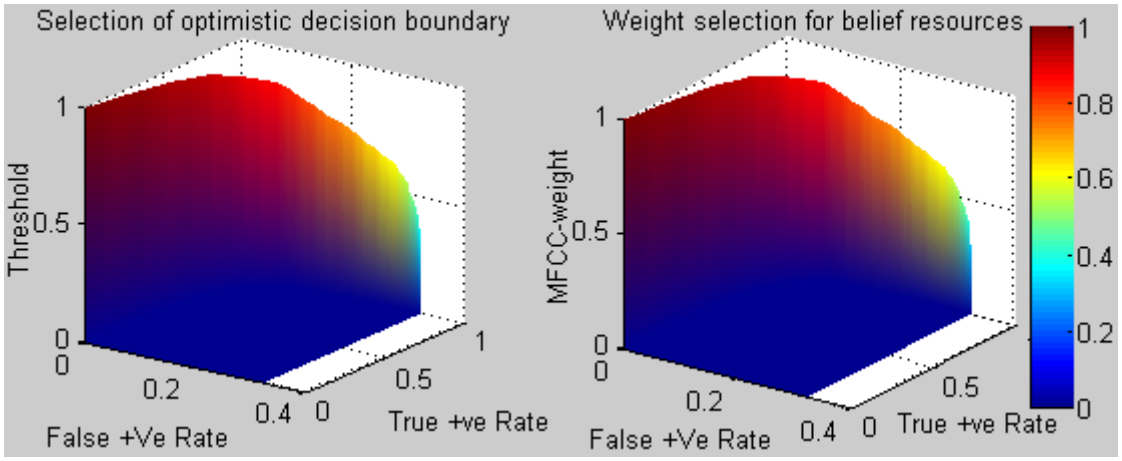

Figure VII: Setting the weights for belief resources and decision boundary value using ROC curves

Weight Allocation to Basic Attributes

Equation 2 and Eq^{n}.3 indicate the dependencies of basic probabilities in DST upon the relative weights allocated to the belief resources $\{\mu(mfcc), WDe\}$. To set up the optimal weights, experiments are conducted on the aforementioned dataset with discrete values for attribute weights (from 0 to 1 with a lag of 0.01) to retrieve the ROC curve as shown in Fig. VII. The ROC indicates the optimal compromise between false

positive rate and sensitivity is achieved at $w_{mfcc} = 0.75;\ w_{wav} = 0.25$. The higher weight for MFCC feature set indicate higher dependency of spotting decision as compared to wavelet based spectral features. However, the individual performances of both attributes $\{\mu(mfcc), WDe\}$ are less than the combined beliefs as shown in Fig.V that validates the importance of DST for the STD.

Table II: Results comparison of proposed STD approach with existing methods for multi-lingual utterances

Key-Words	Total No. of Key-Words	Speaker Gender	Combined Evidence (Mean-MFCC+Wavelets)					DTW+MFCC					DTW-Restricted+MFCC				
			Sensitivity	Specificity	Accuracy	1/LR+	LR-	Sensitivity	Specificity	Accuracy	1/LR+	LR-	Sensitivity	Specificity	Accuracy	1/LR+	LR-
'most'	14	1M/2F	0.75	1	0.95	0	0.25	0.75	0.95	0.91	0.12	0.27	0	1	0.83	NaN	1
'today'	15	2M/1F	1	0.90	0.92	0.09	0	0.8	0.93	0.92	0	0.6	0	1	0.86	NaN	1
'fish'	8	1M/1F	1	0.87	0.88	0.12	0	0.75	0.87	0.85	0.18	0.28	0.25	1	0.91	0.18	0.78
'again'	12	1M/1F	1	0.95	0.95	0.05	0	1	0.80	0.81	0.16	0.28	0.5	1	0.95	0	0.75
قُل	15	3M/2F	1	0.81	0.82	0.18	0	0.6	0.84	0.82	0.25	0.47	0.2	0.98	0.92	0.08	0.81
'dog'	4	1M/1F	1	0.93	0.93	0.06	0	0.25	0.89	0.81	0	0.75	0.25	1	0.9	NaN	1
'john'	15	2M/1F	1	0.86	0.88	0.13	0	0.8	0.81	0.8	0.16	0.23	0	1	0.8	NaN	1
'wood'	9	1M/2F	1	1	1	0	0	1	0.91	0.92	0.2	0.71	0	1	0.88	NaN	1
'collect'	9	1M/2F	1	0.96	0.96	0.03	0	0.6	0.96	0.93	0	0.33	0	1	0.9	NaN	1
'found'	8	2M/2F	1	0.97	0.97	0.02	0	1	0.85	0.86	0.1	0.27	0.75	1	0.97	NaN	1
'enough'	8	2M/2F	1	0.97	0.97	0.02	0	1	0.97	0.97	0.8	1.1	0.75	1	0.97	NaN	1
'cap'	12	2M/1F	0.75	0.94	0.91	0.07	0.26	0.75	1	0.95	0.7	0.91	0	1	0.82	NaN	1
اَلنّاس	10	2M	0.8	1	0.96	0	0.2	0.8	0.955	0.92	0.05	0.21	0	1	0.81	NaN	1
'bed'	4	2M/1F	1	0.87	0.88	0.12	0	1	0.90	0.91	0.05	0.26	0	1	0.91	NaN	1
'throw'	4	1M/2F	1	0.95	0.95	0.04	0	1	0.93	0.93	0	0.5	0	1	0.91	NaN	1
'tim'	12	2M/1F	1	0.80	0.81	0.19	0	1	0.83	0.85	0.25	0.57	1	1	1	NaN	1
'thought'	16	1M/1F	0.75	0.91	0.90	0.11	0.27	0.5	0.85	0.82	0.19	0.78	0.25	1	0.94	NaN	1
'dog'	12	1M/1F	1	0.91	0.92	0.08	0	1	0.89	0.9	0.26	0.8	0.5	1	0.96	NaN	1
'decline'	12	1M/2F	1	0.86	0.87	0.13	0	1	0.83	0.85	0	0	0.25	1	0.92	NaN	1
'said'	10	1M/2F	1	0.74	0.76	0.25	0	0.6	0.86	0.84	0.13	0.43	0.4	1	0.95	NaN	1
آپ	12	2M/1F	1	1	1	0	0	0.75	0.84	0.83	0.21	0.29	0.25	1	0.9	0	0.75
'fish'	10	2M/2F	1	0.87	0.88	0.12	0	0.6	0.94	0.9	0.13	0.22	0.2	1	0.9	0	0.8
'albert'	12	2M/1F	1	1	1	0	0	1	1	1	0.11	0.27	0.25	1	0.88	NaN	1
'threw'	16	1M/2F	1	0.79	0.81	0.20	0	1	0.83	0.84	0	0.5	0.5	1	0.96	NaN	1
'spect'	8	1M/1F	1	1	1	0	0	1	0.92	0.93	0.4	0.62	0.25	1	0.9	NaN	1
'collect'	12	1M/1F	1	1	1	0	0	0.75	1	0.96	0.05	0.26	0	1	0.86	NaN	1
'pilled'	12	1M/1F	1	0.86	0.88	0.13	0	1	0.73	0.76	0.2	0.55	0.5	1	0.95	NaN	1
'please'	12	2M/2F	1	0.97	0.97	0.02	0	1	0.94	0.94	0.08	0	0.5	1	0.94	NaN	1
'main	16	2M/2F	1	1	1	0	0	0.75	1	0.93	0.15	0	0.25	1	0.81	NaN	1
دھماکہ	20	3M/1F	1	1	1	0	0	1	1	1	0	0	0.20	1	0.85	0	0.8
اَللہ	25	4M/1F	0.80	1	0.94	0	0.2	0.80	1	0.94	0	0.2	0	1	0.72	NaN	1
'scare	8	2M/1F	1	0.90	0.91	0.1	0	0.75	0.9	0.87	0.12	0	0.25	1	0.87	NaN	1
'port'	16	2M/2F	1	1	1	0	0	1	0.88	0.9	0.37	0.61	0.25	1	0.9	NaN	1
'quickly'	12	1M/1F	1	0.90	0.92	0.09	0	0.75	0.95	0.92	NaN	1	0	1	0.84	NaN	1
'short'	9	1M/1F	1	0.90	0.91	0.09	0	1	0.83	0.84	0.45	0.78	0	1	0.93	NaN	1
Avg. Sensitivity, Specificity, Accuracy, 1/LR+, LR-			**0.97**	**0.93**	**0.92**	**0.06**	**0.03**	**0.83**	**0.90**	**0.88**	**0.17**	**0.43**	**0.24**	**0.99**	**0.89**	**NaN**	**0.96**

Detailed experimental results for a case study conducted over multi-lingual utterances are presented in Table II. Information about the dataset is shown in terms of total number of occurrences, length, gender, language, and query utterances. As the proposed STD is based on the feature based template matching without the model training, the system performance is independent of query and reference phonemes structure and spoken language. It can be observed from the output statistical metrics (e.g. sensitivity, specificity, accuracy, LR+, and LR-) that the query utterance detection rate is achieved consistently regardless the spoken keywords. The statistical results also demonstrated the robustness of the proposed STD approach as compared to DTW and constrained DTW based methods.

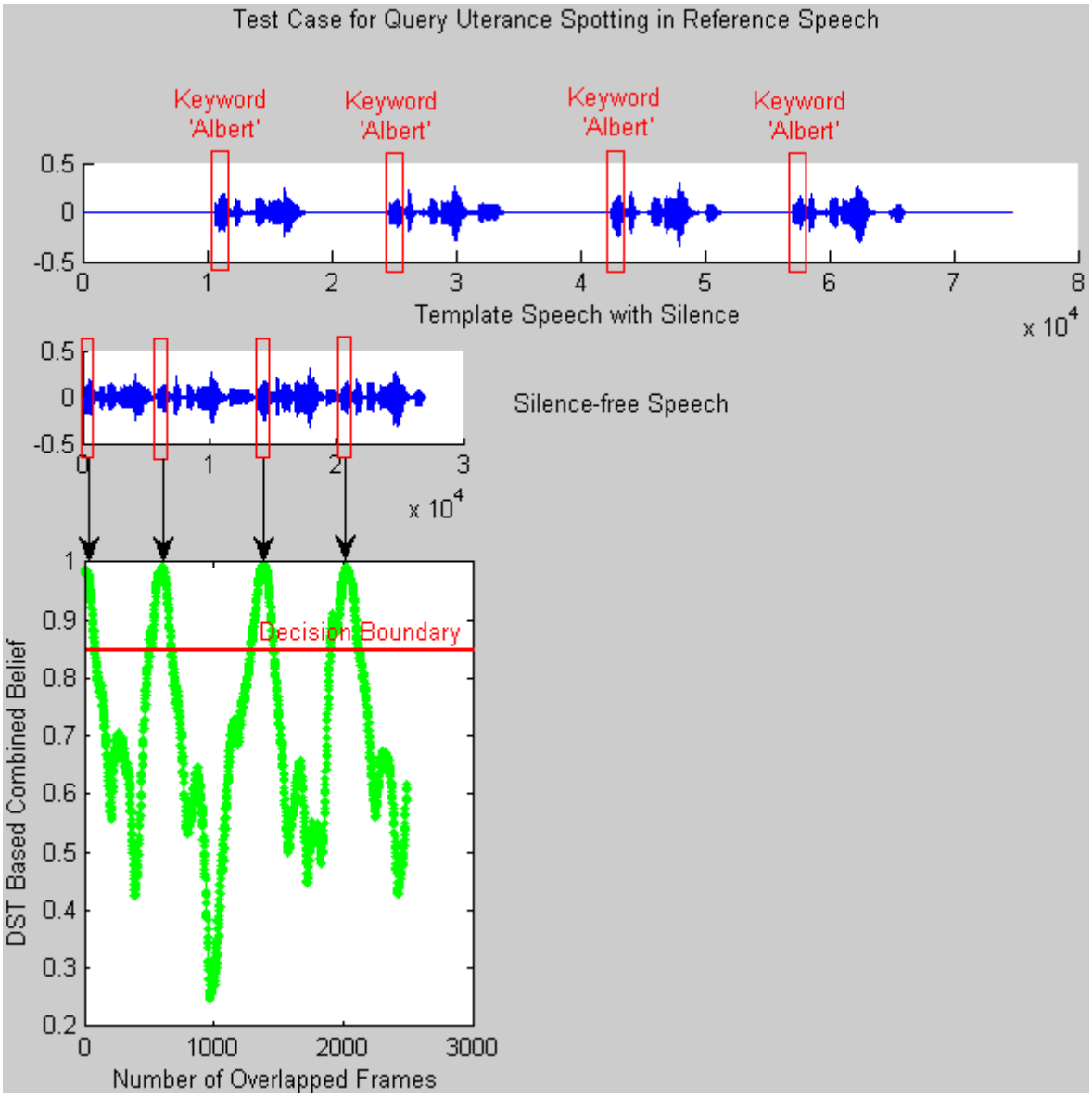

Figure VIII: Test case using the proposed STD for multiple occurrences of query utterance 'Albert'

Figure VIII demonstrates the proof of concept using a test case for query utterance 'Albert' that indicates the robustness of the proposed approach in terms of synchronised spotted locations (peaks) in the reference speech corresponding to each query utterance position (i.e. ground truth).

Table III: Impact of silence removal techniques on the performances of STD approaches in terms of accuracy, sensitivity, and mean square error using the dataset described in Table I

		Proposed approaches			Existing approaches	
		Combined Evidence	**μ (MFCC)**	**Wavelet Energy**	**DTW**	**DTW Constrained**
Performance without Silence Removal	**Accuracy**	0.8044	0.8265	0.8265	0.7616	0.7624
	True +Ve Rate	0.5544	0.7806	0.7806	0.2928	0
Energy, ZCR & Pitch Detection Based Silence Removal	**Accuracy**	0.9266	0.8695	0.8695	0.7616	0.9110
	True +Ve Rate	0.9750	0.9322	0.9322	0.2928	0.2617
Energy & Spectral Centroid Based Silence Removal	**Accuracy**	0.8687	0.8608	0.8608	0.8597	0.8340
	True +Ve Rate	0.5656	0.7617	0.7617	0.5728	0.0233
Type I Error	**μ**	0.0105	0.0258	0.0258	0.0105	6.8871e-05
	σ	0.0148	0.0260	0.0260	0.0110	3.7722e-04
Type II Error	**μ**	0.0063	0.0300	0.0300	0.2596	0.9588
	σ	0.0191	0.1042	0.1042	0.2762	0.1262

Table III demonstrates 'μ' (mean) and 'σ' (standard deviation) values of Type I and Type II errors for five different approaches that indicate the superiority of proposed DST based STD over the existence approaches. Detailed results are presented that demonstrate the silence removal methods impacts on different STD approaches. It can be observed that the sensitivity and accuracy increased from 56% and 80% to 97.5% and 93% respectively by using the proposed pitch tracking, energy, and ZCR based silence removal as compared to existing techniques presented in [28], [36]. Furthermore, the pitch detection based silence removal not only enhances the DST based performance but improves the performances of other approaches also that validate the effectiveness of the proposed silence removal approach.

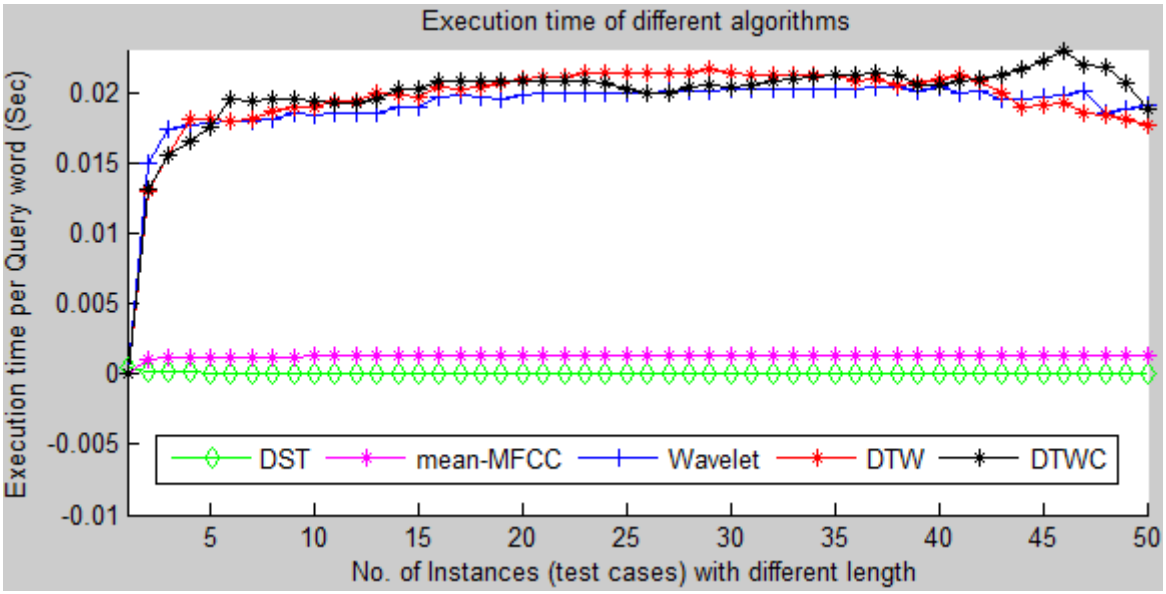

Figure IX: Trade-off between query utterance detection rate and computational cost

The execution time for aforementioned STD approaches is calculated for 50 query utterances as shown in Fig.IX. The efficient computation cost (0.0013 sec) is achieved by using the mean MFCC based approach. However; a very small overhead in terms of computation time (i.e. 0.00005 sec) needed for DST implementation produced a significant improvement in STD performance as shown in Table III. The accuracy rate increased to 92% by using the proposed DST based approach comparing to 86% of MFCC, wavelet, and 76% of DTW based approaches. Despite of the efficient accuracy rate (i.e. 91%) produced by constrained DTW, the true detection of query utterances dramatically decreased to only 26% which fails the main objective of STD. Also, the conventional DTW and constrained DTW use high dimensional features set which increases the search space [37] resulting high execution time (0.02 sec). Similarly, the simultaneous time-frequency analysis in wavelet decomposition needs comparatively higher computation time (0.019 sec). These statistics validate the significance of information combination from multiple belief resources to make a reliable decision for query utterance spotting.

IV. Conclusions and Future Directions

In this manuscript, a comprehensive overview of the research contribution towards the query utterance spotting in continuous speech is presented. An experimental setup was built up comprising speech enhancement using a newly introduced silence removal method, dynamic noise filtration, feature extraction, belief combination and reasoning based decision making. A novel approach for 'pitch tracking based voiced segment detection' and 'energy and zero cross rate based unvoiced segment detection' is introduced to remove the silence segments from speech signal. The statistical results were obtained that validated the proposed approach and its contributions towards the spoken term detection. In the future, these outcomes serve to explore different aspects in the related area. For instance, vocal tract normalisation can be used for speaker independent spoken term detection. Similarly, quantity of the information sources can be increased to analyse the impact on system performance.

References

[1] L. Lee, J. Glass, H. Lee and C. Chan, "Spoken content retrieval-beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 09, Pp. 1389-1420, 2015.

[2] A. Saxena, *Significance of knowledge-based representation of speech for spoken term detection*. PhD [Dissertation]. Hyderabad, India, International Institute of Information Technology, 2015. [Online]. Available: IIIT Hyderabad Publications.

[3] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems, Springer*, vol. 7, no. 3, Pp. 358-386, 2015, DOI:10.1007/s10115-004-0154-9.

[4] A. Abad et al., "On the calibration and fusion of heterogeneous spoken term detection systems," *Conference of the International Speech Communication Association*, *INTERSPEECH*, Pp. 20-24, France, Aug. 2013.

[5] C. A. Ratanamahatana and E. Keogh, "Three myths about dynamic time warping data mining," in *Proceeding of SIAM International Conference on Data Mining*, 2005, Pp. 506-510.

[6] X. Anguera et al., "Query by example search on speech," in *Proc. of Media Evaluation*, Oct. 16-17, 2014, Spain, Pp. 1–2.

[7] H. Joho and K. Kishida, "Overview of the NTCIR-11 spoken query & doc task," In *Proc. of the 11th NTCIR Conference*, Dec. 9-12, 2014, Tokyo, Japan, Pp. 1–7.

[8] J. Tejedor et al., "Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion," *Journal of Audio, Speech, and Music Processing, EURASIP*, vol. 23, Pp. 1–17, 2013.

[9] J. Tejedor et al., "Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 21, Pp. 1-27, 2015.

[10] F. Metze et al., "Spoken web search CEUR Workshop Proceedings," in *Proc. of Media Evaluation,* Aug. 02, 2012, Aachen, Germany, Pp. 1–2.

[11] A. Mandal, K. R. Kumar and P. Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology*, vol. 17, no. 02, Pp. 183-198, 2014.

[12] X. Anguera et al., "The spoken web search task," in *Proc. of Media Evaluation, CEUR Workshop Proceedings*, Oct. 13, 2013, Aachen, Germany, Pp. 1–2.

[13] C. Chan and L. Lee, "Model-based unsupervised spoken term detection with spoken queries," *IEEE Trans on Audio, Speech, and Language Processing,* vol. 21, no. 07, Pp. 1330-1342, 2013.

[14] A. Muscariello, G. Gravier and F. Bimbot, "Zero-resource audio-only spoken term detection based on a combination of template matching techniques," in *Proc. INTERSPEECH'11 (ISCA)*, Aug. 06-08, 2011, Italy, Pp. 921-924, 2011.

[15] Y. Zhang and J. R. Glass, "A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping," in *Proc. of ISCA, INTERSPEECH*, Aug. 28-31, 2011, Florence, Italy, Pp. 1909–1912.

[16] Y. S. Lin and C. P. Ji, "Research on improved algorithm of DTW in speech recognition," *Int. Conf. on Computer Application and System Modelling*, Oct. 20-24, 2010, Taiyuan, vol. 9, Pp. 418-421.

[17] W. Shen, C. M. White and T. J. Hazen, "A comparison of query-by-example methods for spoken term detection," in *Proc. of INTERSPEECH' 09, ISCA*, 2009, United Kingdom, Pp. 2143-2146.

[18] H. Lin, A. Stupakov and J. Bilmes, "Spoken keyword spotting via multi-lattice alignment," in *proc. of INTERSPEECH' 08*, 2008, Pp. 2191-2194, [Online]. Available: http://melodi.ee.washington.edu/~bilmes/mypubs/lin2008-skeylat.pdf. [Accessed: 10 Sep. 2013].

[19] H. Marijn, M. Mitchell and V. L. David, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 22-27, 2011, Prague, Pp. 4436–4439, DOI:10.1109/ICASSP.2011.5947338.

[20] T. Hori et al., "Open-vocabulary spoken utterance retrieval using confusion networks," in *proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* April 15-20, 2007, Honolulu, HI, vol. 04, Pp. 73-76, DOI:10.1109/ICASSP.2007.367166.

[21] H. Lin, A. Stupakov and J. Bilmes, "Improving multi-lattice alignment based spoken keyword spotting," in *proc. Of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 19-24, 2009, Taiwan, Pp. 4877- 4880, DOI:10.1109/ICASSP.2009.4960724.

[22] S. Dhingra, G. Nijhawan and P. Pandit, "Isolated speech recognition using MFCC and DTW," *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering*, vol. 02, no. 08, Pp. 1-8, 2013.

[23] B. Yegnanarayana and T. Sreekumar, "Signal dependent matching for isolated word speech recognition system," *Journal of Signal Processing*, vol. 07, no. 02, Pp. 161- 173, 1984.

[24] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 01, Pp. 145–151, 1991.

[25] C. McCool et al., "Bi-modal person recognition on a mobile phone: using mobile phone data*," IEEE Int. Conf. on Multimedia and Expo Workshops*, July 9-13, 2012, Melbourne, Pp. 635-640.

[26] H. Hung and G. Chittaranjan, "The idiap wolf corpus: exploring group behaviour in a competitive role-playing game," *ACM Multimedia*, 2010, Italy [Online]. Available: http://homepage.tudelft.nl/3e2t5/mmsct22567-hung.pdf. [Accessed: 27 Jan. 2012].

[27] S. F. Bolls, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 02, Pp. 113-120, 2003.

[28] R. G. Bachu et al., "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," *American Society for Engineering Education Zone Conference Proceedings*, 2008, Pp. 1-7.

[29] S. Srikanth, B. K. Kumar and C. Nagababu, "A novel method for silence removal in sounds produced by percussive instruments," *International Journal for Modern Trends in Science and Technology*, vol. 02, no. 05, Pp. 72-76, 2015.

[30] S. A. Zahorian, P. Dikshit and H. Hu, "A spectral-temporal method for pitch tracking," 9th *Int. Conf. on Spoken Language Processing, INTERSPEECH'06* , Sep. 2006, Pittsburgh, PA, Pp. 1710-1713.

[31] S. Young, *HMMs and related speech recognition technologies.* Handbook of Speech Processing, Heidelberg, Berlin, Springer, 2008, Pp. 539-583.

[32] I. Mohino-Herranz et al., "MFCC based enlargement of the training set for emotion recognition in speech," *Signal & Image Processing: An International Journal (SIPIJ)*, vol. 05, no.01, Pp. 29-40, 2014.

[33] Q. Chen et al., "Data classification using the Dempster-Shafer method," *Journal of Experimental & Theoretical Artificial Intelligence*, Taylor & Francis Online, vol. 26, no. 04, Pp. 493-517, 2014, DOI:10.1080/0952813X.2014.886301.

[34] G. Chen, C. Parada and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *Proc. Of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 19-24, 2015, South Brisbane, Pp. 5236-5240, DOI:10.1109/ICASSP.2015.7178970.

[35] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian Journal of Internal Medicine,* vol. 04, no. 02, Pp. 627–635, 2013.

[36] T. R. Sahoo and S. Patra, "Silence removal and endpoint detection of speech signal for text independent speaker identification," *Int. Jour. Image, Graphics and Signal Processing*, vol. 06, Pp. 27-35, 2014.

[37] C. Chung, C. Chan and L. Lee, "Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity," in *Proc. Of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 4-9, 2014, Florence, Pp. 7814-7818, DOI:10.1109/ICASSP.2014.6855121.

Biographies

Wasiq Khan is a PostDoc Research Associate at School of Engineering, University of Central Lancashire, UK. He is research active in the intelligent decision and reasoning based decision support systems, computer vision, machine learning, and signal processing. Khan completed his Ph.D in speech processing and time warped speech similarity measurement at Bradford University, UK. He is Fellow of Higher Education Academy, UK. Contact him at wkhan4@uclan.ac.uk.

Kaya Kuru is a PostDoc Research Associate at School of Engineering, University of Central Lancashire, UK. He is interested in developing autonomous intelligent and decision support systems based on machine learning and image processing algorithms. Kuru completed his Ph.D in Information Systems from the Middle East Technical University, Turkey. He is a member of Medical Informatics Association, Contact him at KKuru@uclan.ac.uk.