Experimental Studies on Effect of Speaking Mode on Spoken Term Detection

Kallola Rout, Pappagari Raghavendra Reddy, K Sri Rama Murty
Department of Electrical Engineering
Indian Institute of Technology Hyderabad, India-502205
Email: {ee12m0003, ee12m1023, ksrm}@iith.ac.in

***Abstract*—The objective of this paper is to study the effect of speaking mode on spoken term detection (STD) system. The experiments are conducted with respect to query words recorded in isolated manner and words cut out from continuous speech. Durations of phonemes in query words greatly vary between these two modes. Hence pattern matching stage plays a crucial role which takes care of temporal variations. Matching is done using Subsequence dynamic time warping (DTW) on posterior features of query and reference utterances, obtained by training Multilayer perceptron (MLP). The difference in performance of the STD system for different phoneme groupings (45, 25, 15 and 6 classes) is also analyzed. Our STD system is tested on Telugu broadcast news. Major difference in STD system performance is observed for recorded and cut-out types of query words. It is observed that STD system performance is better with query words cut out from continuous speech compared to words recorded in isolated manner. This performance difference can be accounted for large temporal variations.**

I. Introduction

In the present era, on account of data deluge over the Internet, it often becomes cumbersome to find relevant information. In a rural and uneducated setting, dissemination of information via some public kiosk or any personal device, where the query can be submitted by merely speaking out a keyword might become a promising solution. In this case, spoken term detection (STD) technique assumes significance. The goal of STD is to retrieve the occurrences of the user-spoken-term from the given speech database. It is instrumental in applications such as content retrieval, voice command detection etc [1]. In STD the input is given in audio form.

The major point of emphasis in the task of STD is the detection of accurate instance of query word in a given reference sample. Conventional STD make use of Dynamic time warping (DTW) technique to match the two time series sequences [2], [3], [4], [5]. But implementation of DTW is less efficient for continuous speech, which is the case with STD. Many researchers had applied speech recognizer to solve this problem, where they transformed the speech signal to acoustic units such as word, syllable or phoneme [6], [7]; which in turn were used to develop vocabulary independent STD model. These methods are highly dependent on speech recognizer and are also difficult to build for large number of speakers.

The authors would like to acknowledge the Department of Electronics and Information Technology (DeitY), Ministry of Communications & Information Technology, Government of India for sponsoring this work

978-1-4799-6619-6/15/$31.00 © 2015 IEEE

In [8], a template matching technique based on conventional DTW was used to match the posterior probabilities of speech frames computed from frame-wise acoustic likelihoods to locate the query word. In [9], Zhang et. al. employed Gaussian mixture model (GMM) to compute posterior probabilities of the query utterance and reference utterances, which were then matched using DTW. Lee et. al. proposed an unsupervised acoustics segment model (ASM) for STD [11]. Speech frames were represented by frame label ASM posteriorgrams and segmental DTW was used to match the query and reference utterances. These frame based methods do not carry speech information for long utterance. The methods proposed in [8], [9], [11] are unsupervised techniques, which delivered poor efficiency for large database system.

Speech signals generally exhibit both temporal and spectral variability. Temporal variability can be accounted by DTW, but spectral variability can not be properly handled. Recently a certain class of artificial neural network particularly multilayer perceptron (MLP) was used as a discriminative classifier and was applied to a variety of problem domains in order to handle spectral variability [12], [13]. MLP can be used in supervised framework and has the ability to learn the underlying classes from training dataset.

In this paper, a supervised technique for STD based on posteriors obtained from an MLP is analyzed with respect to different modes of query words and different groupings of phoneme classes. For STD, the MLP is trained with respect to continuous Telugu news bulletins database, using mel-frequency cepstral coefficients (MFCCs) as features. After obtaining posterior features from the trained MLP, the matching of sequences of features from query and reference utterances is done using subsequence DTW (SubDTW) [15]. Most of the research on STD were done with the query words extracted from continuous speech. But, in real world scenario, query words are spoken in isolated manner. In this paper, we analyze the effects of two kinds of query words- those recorded in isolation and those cut out from continuous speech. It is found that the isolated query detection performed worse than detection of query cut out of continuous speech, owing to large differences in duration of recorded query and those spoken in continuous speech. We have also performed classical phoneme recognition task to study the effects of number of classes chosen, considering classes with 45, 25, 15 and 6 phoneme groupings. It is noticed that the phoneme recognition accuracy decreases as number of classes increases. In terms of P@N performance measure, the class with 25 phonemes had

delivered the best performance and is used for studying the effects of different modes of query word detection.

The rest of the paper is organized as follows: Extraction of posterior features by training MLP is explained in section II. Section-III explains matching of two time-series sequences. Experimental setup and evaluation results of STD are discussed in section-IV. Section-V features with conclusion and future work.

II. Extraction of Phoneme Posteriors

Conventional features of speech such as MFCC, perceptual linear prediction (PLP) etc. exhibit speaker variability across the utterances. Fig. 1(a) illustrates the similarity matrix based on MFCC features obtained between reference and query utterances spoken by same speaker and Fig.1 (b) illustrates the same for different speakers. Black color indicates more similarity and white indicates less similarity. It can be seen from Fig. 1(a) that the match between the query and the segment of reference utterance is clearly visible. However in Fig. 1(b) this match is not noticeable because of the speaker variability.

A stabilized set of features, which is invariant to speaker is required for efficient STD system. Features based on posterior probabilities of acoustic classes are known to be robust to speaker variability [16]. In this paper, MLP has been employed to capture nonlinear relationships in speech utterances, thereby computing posterior features of utterances.

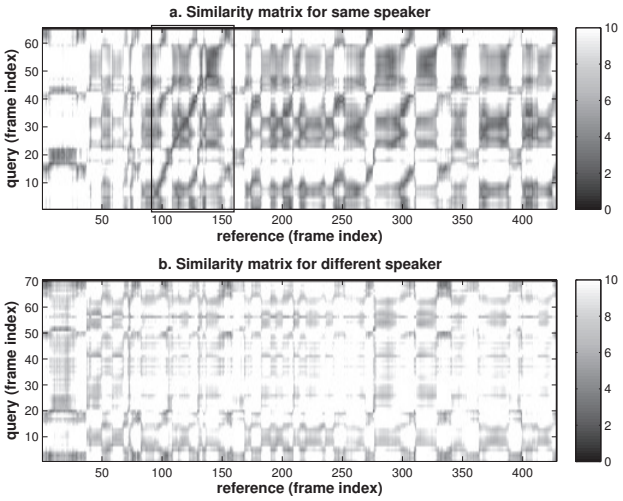


Fig. 1. Similarity matrices using MFCC as feature vector. Query spoken by (a) same speaker in a different context, (b) different speaker

A. Training MLP

The framework of MLP training used in this work is shown in Fig.2. MFCC vectors are fed into MLP at the input layer and training is carried out in supervised manner. Back propagation (BP) algorithm is used to maximize the cross entropy function [12], [14]. The MLP contains 3 layers, namely input, output and hidden layer. Context based information is exploited by concatenating adjacent speech frames. Combining P dimensional MFCC features from L consecutive frames, the M dimensional input vectors are obtained. Z is the dimension of the hidden layer and N is the dimension of output layer, where N is the number of phonemes used for classification. Sigmoid function is used as the activation function in hidden layer. To get the posterior probability we use Softmax function at the output layer [14]. Required phoneme labels (to train the MLP in supervised manner) are obtained by force alignment, which in turn is done using hidden markov model(HMM) training.

The number of phonemes varies from 20 to 50 for most of the languages. For training this network, we investigated the usage of different number of phonemes at the output layer. Initially, 45 classes were used for training MLP and for comparative study, these 45 classes are quantized depending upon the manner of articulation into 25 classes, 15 classes and 6 classes as shown in TABLE I.

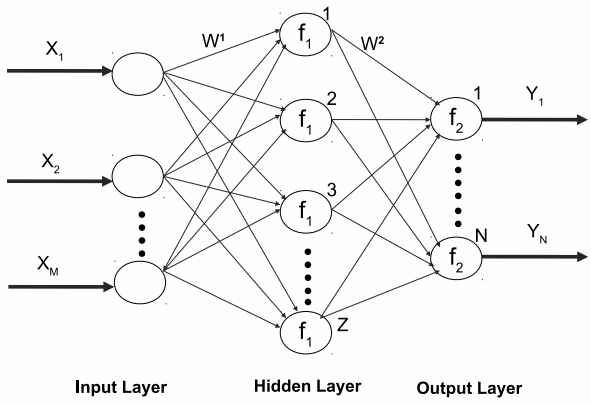


Fig. 2. MLP Network, where W^1 is the weight matrix at the input layer and W^2 is the weight matrix at the output layer. X and Y are the input and output vectors respectively. M, Z and N denote the number nodes at the input, hidden and output layer respectively. f_1 and f_2 denotes the activation function at hidden and output layer respectively.

B. Posterior feature

A word can be represented as a sequence of phonemes. Given a speech frame, posterior probabilities represent the posterior distribution over the defined class of phonemes. The sequence of frames for a speech utterance is defined as

$$F = [f_1, f_2, ..., f_t, ..., f_T] \tag{1}$$

and the sequence of posterior features are

$$G = [g_1, g_2, ..., g_t, ..., g_T] \tag{2}$$

Each posterior feature is represented by

$$g_t = [P(C_1|f_t), ..., P(C_k|f_t), ..., P(C_N|f_t)] \tag{3}$$

where $\{C_k\}_{k=1}^{N}$ represents set of phoneme class and $P(C_i|f_t)$ represents the posterior probability of i^{th} class given frame f_t. Each speech frame can be written as posterior vector of given phoneme class size. Thus, speech utterance containing T frames can be written as a $N \times T$ matrix, where each column represents the posterior probabilities of the corresponding

TABLE I. GROUPING OF THE PHONEMES INTO DIFFERENT CLASSES

<table>
<tr><td>45 classes</td><td>a</td><td>aː</td><td>i</td><td>iː</td><td>j</td><td>u</td><td>uː</td><td>e</td><td>eː</td><td>o</td><td>oː</td><td>ʋ</td><td>ʃ</td><td>ʂ</td><td>s</td><td>h</td><td>f</td><td>m</td><td>n</td><td>k</td><td>kʰ</td><td>g</td><td>gʰ</td><td>c</td><td>cʰ</td><td>z</td><td>jʰ</td><td>ȷʰ</td><td>ʒ</td><td>ɟ</td><td>ʈ</td><td>ʈʰ</td><td>ɖ</td><td>ɖʰ</td><td>t</td><td>tʰ</td><td>d</td><td>dʰ</td><td>p</td><td>pʰ</td><td>b</td><td>bʰ</td><td>r</td><td>l</td><td>sil</td></tr>
<tr><td>25 classes</td><td colspan="2">a</td><td colspan="2">i</td><td>j</td><td colspan="2">u</td><td colspan="2">e</td><td colspan="2">o</td><td>ʋ</td><td colspan="3">s</td><td>h</td><td>f</td><td>m</td><td>n</td><td colspan="2">k</td><td colspan="2">g</td><td colspan="2">c</td><td colspan="5">j</td><td colspan="2">ʈ</td><td colspan="2">ɖ</td><td colspan="2">t</td><td colspan="2">d</td><td colspan="2">p</td><td colspan="2">b</td><td>r</td><td>l</td><td>sil</td></tr>
<tr><td>15 classes</td><td colspan="2">a</td><td colspan="3">i</td><td colspan="2">u</td><td colspan="2">e</td><td colspan="3">o</td><td colspan="5">F(Fricatives)</td><td colspan="2">N(Nasal)</td><td colspan="4">G(Glottal)</td><td colspan="7">P(Palatal)</td><td colspan="4">R(Retroflex)</td><td colspan="4">D(Dental)</td><td colspan="4">B(Bilabial)</td><td>r</td><td>l</td><td>sil</td></tr>
<tr><td>6 classes</td><td colspan="12">V(Vowel)</td><td colspan="5">F(Fricatives)</td><td colspan="2">N(Nasal)</td><td colspan="23">C(Consonants)</td><td colspan="2">T(Trill and Liquid)</td><td>sil(Silence)</td></tr>
</table>

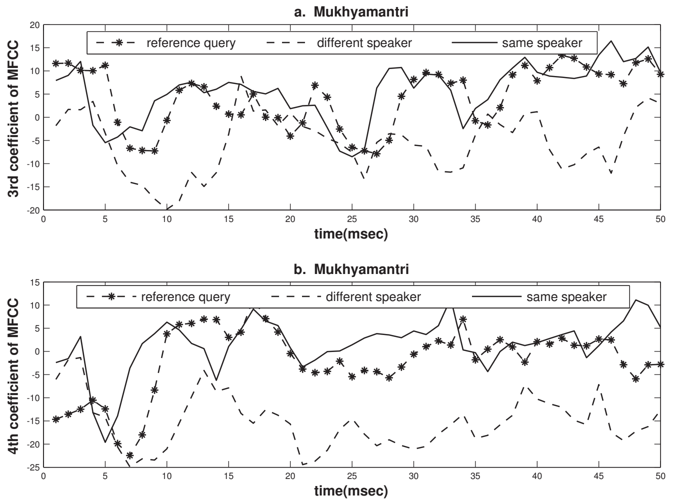


Fig. 3. Comparison of variability of MFCC coefficients in a word “Mukhyamantri” spoken by two different speaker

frame. The variation in the 3^{rd} coefficient of P dimensional MFCC vectors of query word spoken by the same speaker and a different speaker as that of reference utterance are shown in Fig. 3. It can be seen that, when the query word and reference utterance are spoken by the same speaker, the MFCC coefficients are closely following each other. But in the case of distinct speakers for query and reference utterances, a large deviation in MFCC coefficients can be observed. Fig. 4 shows the posterior probabilities of phoneme from query word uttered by the same as well as different speaker as of reference utterance. The robustness of posterior features towards speaker variabilities are portrayed in this figure by the close matching of posteriors of query and reference utterance irrespective of speakers.

To visualize the significance of posteriorgrams, the phoneme posteriorgrams for 15 phoneme classes are shown in Fig. 5. Posteriorgram of reference is plotted in Fig. 5(a) and the highlighted region shows the presence of query word. Fig. 5(b) and 5(c) represent the posteriorgrams of query spoken by a different speaker and same speaker as that of the reference. It can be seen from this figure that, same phonemes are getting activated upon time frames irrespective of the speaker. Thus the posterior features were able to overcome the speaker dependency in MFCC features, and hence are more stable than MFCCs. Therefore we move further by considering posterior features to represent the speech frames.

Similarity matrices using posterior probabilities as input vectors are shown in Fig. 6, in cases when query and reference utterances spoken by same and different speakers. An unambiguous DTW path can be observed in both Fig. 6(a) and Fig. 6(b) at the marked area. This depicts the effectiveness of posterior features in bringing out the speaker independent acoustic information from utterances, thereby delivering better

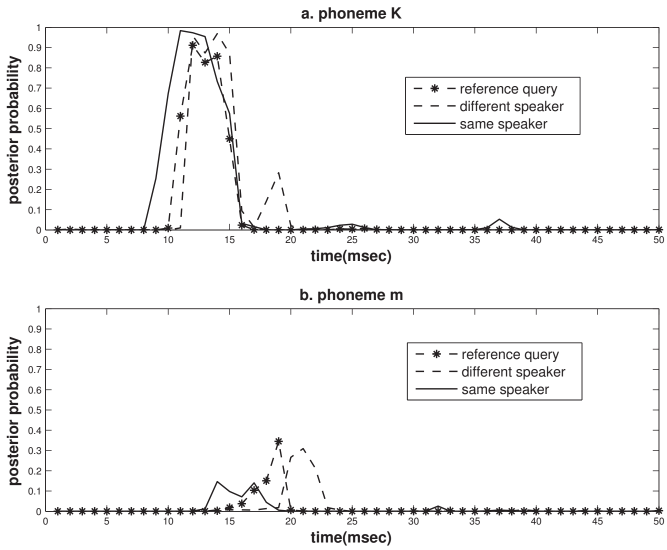


Fig. 4. Comparison of variability in posteriors of phonemes in a word “Mukhyamantri” spoken by two different speaker

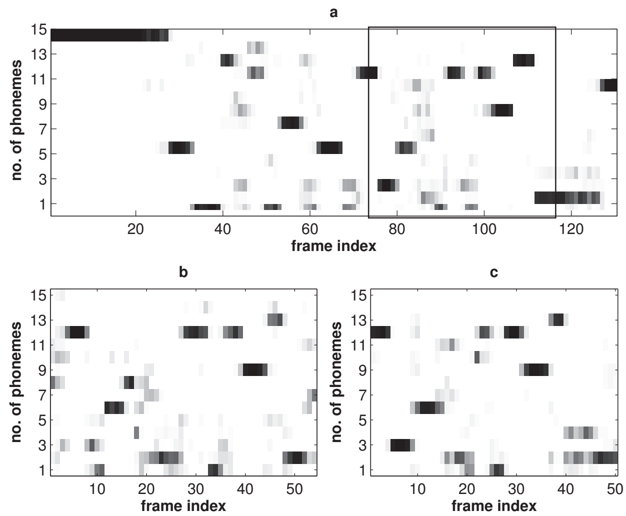


Fig. 5. This figure illustrate stability of posterior representation. (a) Reference posteriorgram. Posteriorgrams of same query word spoken by (b) different speaker, (c) same speaker in a different context

STD performance.

III. TEMPLATE MATCHING

Template matching is expected to bring out pairs of most similar frames from two templates. In STD task template represents speech utterances. Since, durations of phonemes vary from one template to another, the warping path is nonlinear. DTW is a widely used technique to capture temporal alignment between two speech utterances. DTW performs well if both

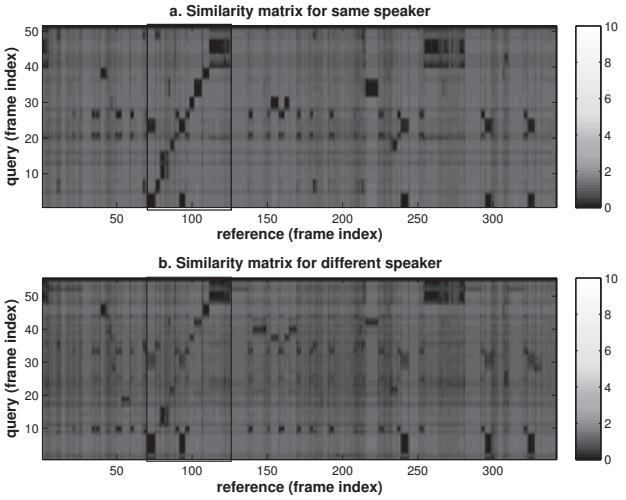


Fig. 6. Similarity matrices using posterior probability as feature vector. Query spoken by (a) same speaker with a different context,(b) different speaker

query and reference utterances are of comparable lengths, so it is not suitable for the task of detecting spoken term in continuous speech.

Since the query word can start from anywhere in reference we keep individual distances in the first row of accumulated distance matrix i.e., we do not accumulate distances along the first row as in the case of conventional DTW. The concept of SubDTW renders the requirement of STD system, where the best matching paths between query and reference utterances can be traced by backtracking from local minimum points along the last row of accumulated distance matrix. In Fig. 7(a) shows the path followed to calculate the accumulated matrix in case of conventional DTW and Fig. 7(b) shows the path followed for the same in case of SubDTW with local weights a, b, c, d, e. These local weights can be adjusted to account for the incomparable durations of query and reference utterances.

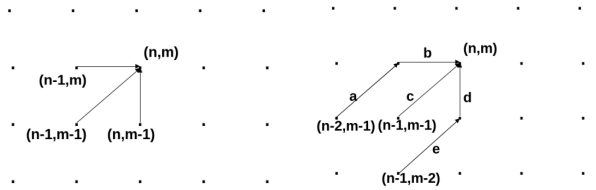
To perform the DTW task on posterior feature vectors, an efficient distance metric to calculate the similarity matrix is needed. Authors have shown in [10] that Kullback-Leibler divergence works well for posterior features compared to dot product and euclidean distance. Let r and s are posterior vectors in reference and query words respectively.
Kullback-Leibler divergence : KL-divergence will give the difference between two probability distribution r and s.

$$D_{KL}(r||s) = \sum_j r_j log(\frac{r_j}{s_j}) \tag{4}$$

This symmetric KL-divergence has been successfully applied to calculate the distance between two probability distribution functions [17].

IV. Experimental Setup and Results

Our STD system is evaluated on Telugu news data. The total available data is divided into two subsets, training set and testing set. The training dataset includes 3500 utterance of $3\frac{1}{2}$ hours read by 6 male and 6 female speakers. The testing dataset consists of 1 hour news data from 3 male and 3 female



(a) conventional DTW path (b) subDTW path with local weights a b c d e

Fig. 7. Path shown for different DTW

TABLE II. Recognition accuracy for using different number of phoneme grouping(C)

C	HMM	ANN
6	77.88	81.87
15	70.6	76.02
25	69.51	74.24
45	62.68	69.11

TABLE III. Average performance using subsequence DTW by taking different phoneme classes

Performance measure	6 classes	15 classes	25 classes	45 classes	raw MFCCs
P@N	44.05	77.65	80.13	72.36	45.68
P@2N	55.68	86.50	89.13	80.57	54.91
P@3N	59.17	88.10	90.75	82.39	60.81
P@4N	61.19	88.71	91.28	83.10	63.23
P@5N	62.13	88.99	91.61	83.41	64.29

news readers. The training and testing dataset are constituted by distinct speakers.

Each utterance is segmented into frames of 25 ms duration with a shift of 10 ms. 39 dimensional MFCC vectors (13 cepstra + 13 Δ + 13 $\Delta\Delta$) are extracted from each frame. 13 adjacent frames are concatenated to exploit context dependent information, thereby fixing input layer size of 507 (13×39) for MLP. The posterior vectors of dimension 25 specifying 25 phonetic classes are obtained for each frame from the trained MLP. SubDTW based template matching of posterior feature vectors is done for STD. Each query word is tested against 5-10 minutes of Telugu news bulletin.

A comparison study is carried out by taking different number of phoneme classes for classification. For phoneme recognition we used both HMM and MLP and the results are shown in TABLE II. In contrast to the traditional HMM, which depends on the prior probability of the states to predict the hidden states, the discriminative approach which is used in MLP, utilizes observed output values to obtain the posterior probabilities. Hence the posterior is directly obtained from the observation sequence and this essentially circumvents the issue of high model dependency and low efficiency which haunts the HMM as shown in TABLE II. Corresponding to an increase in number of phoneme classes, the nodes at the output layer of MLP increases and this results in a notable drop in phoneme recognition accuracy.

Evaluation was done using P@N according to [8], where P@N is average precision for top N hits, where N is the

TABLE IV. AVERAGE PERFORMANCE USING SUBSEQUENCE DTW BY TAKING 25 CLASSES

Performance measure	P@N	P@2N	P@3N	P@4N	P@5N
Cut queries from read speech	80.49	88.61	90.10	90.67	90.84
Isolated recorded queries	56.02	66.70	69.66	70.82	71.25

TABLE V. P@N (%) MEASURE FOR QUERY WORDS FROM CONTINUOUS SPEECH. THE SAME FOR RECORDED ISOLATED QUERIES IS GIVEN IN PARENTHESIS.

Query word	P@N	Query word	P@N
prənəbmukʰərji	99.75(97.50)	digviɟəjsiŋg	82.78(80.00)
teləngaːnə	83.50(60.00)	adjəksʰuraːlu	76.96(50.00)
səmaːveːsəm	88.97(81.73)	prəbʰutvəm	71.43(68.00)
saĭləjaːnaːtʰ	84.64(81.17)	adʰikaːrulu	85.31(69.28)
alpəpiːɖənəm	84.62(77.69)	haĭdraːbaːd	83.57(82.14)
paːrləmenţ	88.24(81.17)	kəmːţi	58.82(47.06)
bəngaːlaːkʰaːtəm	81.75(75.00)	eṉnikəlu	72.25(48.75)
kaŋgres	78.00(74.70)	erpaːţu	63.38(40.68)
raːjiːnaːma	85.19(79.62)	vaːtaːvərənəm	67.00(71.49)
neːpətʰjəm	62.69(61.94)	vibʰəɟənə	71.43(47.62)
pənɕaːjəti	76.62(73.79)	səmaikʰjə	93.55(74.19)
soːnijaːgaːndʰi	90.83(81.47)	ḍilliː	50.00(30.00)
poːliŋg	63.75(31.25)	vivəraːlu	80.00(80.00)
kirənkumaːrreḍḍi	95.53(98.57)	ruːpaːji	70.00(38.00)
nirməjəm	83.33(38.89)	məntri	32.73(28.57)

number of times query occurs in reference utterance. TABLE III shows P@N by considering different phoneme classes. TABLE II and III elucidates the independence of STD and phoneme recognition. We observed that there is significant decrease in P@N from 15 to 6 classes. As more information lies in vowels compared to consonants and in case of 6 classes we grouped all the vowels into one class caused this reduction in performance. Segregation of the different phoneme classes into further sub-classes helps to bring up phoneme accuracy index. The number of these sub-classes can be anything from 2 to 45. However even by dividing the phoneme classes into merely 2 sub-classes namely voiced and unvoiced, does not assure an increase in STD performance. Concluding 25 phoneme classes giving better results, further analysis was carried out only for this class.

The experiment was done for 30 queries cut from continuous read speech and from recordings of isolated queries spoken by the 20 Telugu regional speakers. As the channels are different in read as well as recorded speech, (it is TV channel in read speech and microphone channel in recorded speech), mean normalization was performed. The average performance of STD with 20 speakers over 30 queries is given in the TABLE IV for both read speech and recorded queries. Because of channel mismatch and huge differences in number of frames between the recorded isolated queries and cut queries from read speech, the STD performance for recorded queries was consistently worse than those with read queries. To get rid of frame length mismatch problem, local weights are updated differently for both the recorded and cut out queries. The local weights used for isolated queries $[2\ 3\ 1.5\ 1\ 2.5]$ and for recorded queries $[2\ 3\ 2\ 1\ 1\ ]$. It was observed that the isolated queries contain $1\frac{1}{2}$ times the number of frames as those in queries cut from continuous speech, on an average.

Individual query words and corresponding precision rates have been tabulated in TABLE V using subDTW. A very obvious observation to be made is the decrease in precision scores with drop in the number of syllables present in the query word. The query word "*məntri*" is an epitome of this property. məntri with only two syllables('man','tri') this query has a low P@N(%) of only 28.57. Longer spoken terms are detected with high accuracy.

V. CONCLUSION AND FUTURE WORK

An analysis of effect of speaking mode of query for STD was presented in this paper. Experiments were carried on two types of speaking modes of queries: words spoken in isolation and extracted from continuous speech. Representation of utterance was done using posteriorgrams as they were performing well in nullifying speaker variabilities for the task of STD. The required posteriorgrams for speech utterances were obtained by training MLP. The posterior features for query and reference utterances were matched for STD using SubDTW, as they outperformed classical DTW in terms of computational as well as time complexity. A comparison of different grouping of phoneme classes was studied and 25 phoneme classes was resulted a better performance among all. Experimental evaluation of the this method was carried out on Telugu News bulletin database. The query words were obtained by cutting from continuous speech as well as recording isolated queries. This SubDTW method on posterior features had delivered improved STD accuracy in comparison with conventional MFCCs. Also the performance of STD with query words cut from continuous news speech was consistently better than that of isolated recorded queries, owing to lack of disparities in terms of number of frames.

REFERENCES

[1] H. Wang, T. Lee, C. C. Leung, "Unsupervised spoken term detection with acoustic segment model," *International Conference on Speech Database and Assessments*, pp. 106–111, Oct. 2011.

[2] H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[3] H. Ney,"The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 263–271, Apr. 1984.

[4] K. Ng, V. W. Zue,"Subword-based approaches for spoken document retrieval" *Speech Communication* vol. 32, no. 3, pp. 157–186, 2000.

[5] L. Rabiner, A. E. Rosenberg, S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 6, pp. 575–582, Dec. 1978.

[6] J. Mamou, B. Ramabhadran, O. Siohan, "Vocabulary independent spoken term detection," *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.615–622, 2007.

[7] D. R. H. Miller, "Rapid and accurate spoken term detection," *Interspeech*, pp. 314–317, 2007.

[8] T. J. Hazen, W. Shen, C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 421–426, May 2009.

[9] Y. Zhang, J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 398–403, 2009.

[10] P. R. Reddy, K. Rout, K. S. R. Murty, "Query word retrieval from continuous speech using GMM posteriorgrams," *Signal Processing and Communications (SPCOM), 2014 International Conference*, pp. 1–6, July 2014.

[11] H. Wang, C. Leung; T. Lee, B. Ma, H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5157–5160, Mar. 2012.

[12] M. R. Azimi-Sadjadi, S. Citrin, S. Sheedvash, "Supervised learning process of multi-layer perceptron neural networks using fast least squares," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1381–1384, Apr. 1990.

[13] Y. Arriola, R. A. Carrasco, "Integration of multi-layer perceptron and Markov models for automatic speech recognition," *UK IT Conference*, pp. 413–420, Mar. 1990.

[14] S. S. Haykin,"Neural Networks: A Comprehensive Foundation," *Prentice Hall*, 1999.

[15] M. Mller,"Information Retrieval for Music and Motion,"*New York, NY, USA*, Springer, 2007.

[16] G. Aradilla, H. Bourlard, M. Doss, "Posterior features applied to speech recognition tasks with user-defined vocabulary," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3809–3812, Apr. 2009.

[17] T. M. Cover, J. A. Thomas, "Elements of information theory", *John Wiley and Sons*, 2012.