Query-by-Example Spoken Term Detection Using Low Dimensional Posteriorgrams Motivated by Articulatory Classes

Abhimanyu Popli [#*1] and Arun Kumar [#2]

[#] *Centre for Applied Research in Electronics, Indian Institute of Technology Delhi*
New Delhi-110016, India
[1] crz108049@care.iitd.ac.in
[2] arunkm@care.iitd.ac.in

[*] *Centre for Development of Telematics*
New Delhi-110030, India

***Abstract*—This paper addresses the problem of Query-by-Example Spoken Term Detection (QbE-STD). Posteriorgrams have been widely used in the research on QbE-STD. Features based on articulatory classes are known to be robust to phonemic variations. The articulatory features like voicing and place of articulation are the main distinguishing features among some plosives and fricatives. These properties inspire the study of posteriorgrams based on articulatory classes for QbE-STD. Most of the previous works based on articulatory features have defined a large number of articulatory classes making it difficult to use them directly for pattern matching. Also, most of the works have completely ignored the uniqueness of the phonemes having transitory places of articulation eg. diphthongs and approximants. These issues have been addressed in this work while carefully selecting low dimensional articulatory motivated (LDAM) posteriorgrams on the basis of detailed experiments. This work is the first to show that the articulatory based posteriorgrams can outperform the phonemic posteriorgram significantly in a stand alone way (without any support from acoustic or phonemic features) for the task of QbE-STD.**

I. Introduction

Multimedia content is growing by leaps and bounds every day. Information retrieval technologies are required to use this multimedia content. A lot of research has gone into one such technology called Query-by-Example Spoken Term Detection (QbE-STD) which refers to retrieving a spoken query from the audio content based on acoustic properties. Some applications of this technology include searching audio queries in voluminous audio archives like telephone conversations, news or long digital lectures.

Recent research in STD has largely focussed on lattices [1],[2],[3],[4] and posteriorgram based techniques [5],[6],[7],[8] for the representation of speech. A posteriorgram is a vector of posterior probabilities of its constituent elements given the observation. Dynamic Time Warping (DTW) is often used to do the pattern matching in posteriorgram based techniques. In the Spoken Web Search (SWS), the 2013 benchmark evaluation campaign, 9 out of 13 systems used DTW [9]. A majority of the systems based on posteriorgrams have used phonemic posteriorgrams e.g. [5],[6]. Gaussian posteriorgrams have also been proposed for low resource scenarios e.g. [7], [8].

One shortcoming of the phonemic posteriorgrams is that similar sounding phonemes in a posteriorgram are correlated. This is because multiple phonemes share similar articulatory attributes. Each phoneme can be defined by a unique set of articulatory properties. Thus, multiple posteriorgrams can be used to represent all the phonemes and handle phonemic variations arising due to change in a particular articulatory property. Articulatory based posteriors are known to be robust to noise [10] and they have good classification accuracies across languages [11]. Articulatory classes have been researched for applications like speech recognition since long time. They have also been used for spoken term detection recently [12].

To the best of our knowledge, standalone articulatory based posteriors have never been reported to perform better than phonemic posteriors. They have been used as tandem features with acoustic features for getting higher accuracies [13],[14]. In this paper, we aim to establish the use of Low Dimentional Articulatory Motivated (LDAM) posteriorgrams for QbE-STD in a monolingual environment i.e. English. The rest of the paper is organized as follows. Section II describes some relevant observations from the prior frameworks which form the basis for our framework. Section III describes the proposed framework for QbE-STD. We have studied two mappings from phonemes to LDAM classes viz. LDAM1 and LDAM2. Whenever we make a statement that applies to both LDAM1 and LDAM2 we just use the word LDAM. Section IV describes the experiments and the results. Conclusions are drawn in section V. The phrase *Consonant Position* refers to the place of articulation of consonants and *Consonant Manner* refers to the manner of articulation of consonants in this work. Phoneme symbols used in this work are the same as in CMU Pronouncing Dictionary [15].

MMSP'15, Oct. 19 - Oct. 21, 2015, Xiamen, China.

II. The basis of the present work

Posteriorgrams are used in the present work to represent the audio content of the spoken terms and the test utterances. Further, Dynamic Time Warping (DTW) [5] is used to match the posteriorgrams of the spoken terms and the test utterances. Next, we propose LDAM posteriorgrams instead of gaussian posteriorgram [8] or phonemic posteriorgram [5]. Also, the classes for creating posteriorgrams differ from articulatory classes in previous works. The articulatory classes in [10], [13] and [14] are displayed in table I.

Some observations on these classifications are following:

TABLE I: Set of articulatory classes in some previous works

Mantena et. al. [14]	Kirchoff et. al. [10]	Frankel et. al. [13]
Place of articulation, Manner of articulation, Voicing, Lip rounding, Vowel frontness, Vowel height, Vowel length, aspiration and Silence	Place, Manner, Voicing, Rounding, Front-back	Place, Degree, Glottal state, Rounding, Nasality, Frontness, Height, Vowel

A. It can be gathered from table I that vowels and consonants have different articulatory features. For example, vowels are described by frontness, height and rounding. The voicing class is redundant for English vowels since all the vowels are voiced. Consonants are described by manner of articulation, place of articulation and voicing. These observations are relevant in combining the articulatory classes to reduce the total number of classes.

B. In diphthongs, the tongue and the lip positions may be dynamic. Therefore, it is not possible to describe diphthongs by a single configuration of height, frontback and rounding features. Gautam et. al. club all the diphthongs in same subclass Diphthongs under the class Vowel Length [14]. But, diphthongs like /ay/ (as in might) and /oy/ (as in *toy*) are similar neither in the beginning nor in the end. Only the phonemes which share one or more articulatory feature(s) should be combined together under a subclass.

C. Approximants are put together under the subclass Approximant under the manner of articulation class [14]. Similarly, in phoneme to articulatory feature mapping [13], phonemes /w/, /y/ and /r/ fall under the Approximant subclass in the degree class. The Approximant category comprises of many categories of sounds such as semi-vowels, rhotics and laterals [16]. Therefore, it might be useful to allot separate subclasses for semi-vowels, rhotics and laterals.

D. Vowel articulatory features are dependent on the tongue positions. They are recognized by their formants. So, acoustic features should suffice the identification of vowels. Also, it is difficult to assign a class for the place of articulation for a vowel as some vowels (diphthongs) have non-steady places of articulation. Apart from this, the semi-vowel /y/ (as in *yes*) has a transitory place of articulation between /ih/ and /ah/. Similarly, /w/ (as in *fluctuate*) has a transitory place of articulation between /uh/ and /ah/. So, the diphthongs and the semi-vowels partially overlap with more than one place of articulation of vowels. We present a novel set of classes based on articulatory classes and the above observations, in the next section.

Next, we look at the methods for finding the distance between the spoken query posteriorgram and the test utterance posteriorgram which is required in the DTW algorithm. The negative of the logarithm of the dot product has been used by some researchers to find the distance between two posteriorgrams $\mathbf{q}$ and $\mathbf{u}$ [6],[8]. But it is hypothesized in [5] that KL (Kullback-Leibler) divergence is better suited than the negative of the logarithm of the dot product in the environments where the distributions are 'flatter'. KL divergence has also been used to compute the scores between a distribution characterizing an HMM state and posterior features of an utterance in large vocabulary recognition task in [17]. We use the symmetric KL distance in Eq. (1) in this work for comparing posteriorgrams.

$$kl(\mathbf{q}||\mathbf{u}) = \sum_{i=1}^{D} \big[q(i)\log\frac{q(i)}{u(i)} + u(i)\log\frac{u(i)}{q(i)}\big], \qquad (1)$$

where $q(i), u(i)$ are the elements of the posteriorgrams $\mathbf{q}$ and $\mathbf{u}$ respectively. D is the dimension of the posteriorgrams. The usage of the symmetric KL distance in the proposed framework is described in the next section.

III. LDAM posteriorgrams based framework

A. LDAM1 posteriorgrams

We propose a three class representation in table II. Three Multilayer Perceptons (MLPs) are used to train the three classes. All the vowels are kept in one class. Consonants are represented by two classes, namely Consonant Position and Consonant Manner. The Consonant Position subclass for consonants is similar to the subclass *Place* in the representations in table I. Consonant Manner class includes lateral, trill, approximant and nasal subclasses. The phonemes /l/,/r/,/y/,/v/,/w/ are accommodated in different subclasses. It can be noticed that all the phonemes have a unique configuration in terms of subclasses in LDAM1 representation baring few exceptions. Some vowels with very close places of articulation, e.g. (/ih/ and /iy/), are kept in the same class. Another exception is of the phonemes /r/ and /er/ which are very similar in the KL distance matrix described in the next section.

This arrangement is similar to the arrangement of phonemes in the International Phonetic Alphabet (IPA) Chart [18]. In the IPA chart, the consonants and the vowels are listed in different tables. In the consonant table, the rows represent the manner of articulation of consonants while the columns represent the place of articulation. Approximants, trills and laterals have been allotted different subclasses in the rows.

B. LDAM2 posteriorgrams

Some enhancements were made to the LDAM1 configuration by looking at the LDAM transcription of several words

TABLE II: Arrangement of phonemes in LDAM1 configuration

Class:Vowel												
Subclass : Label	front-high: fup	front-low: fbo	back-high: bup	back-mid: bmi	back-low: bbo	central: cen	eh: eh	ey: eh	diph1: day	diph2: daw	diph3: doy	not vowel: NV	sil: sil
Constituent phonemes	ih,iy	ae	uh,uw	ao.ow	aa	ah	eh	ey	ay	aw	oy	all cons.	sil
Class:Consonant Position													
Subclass: Label	velar: vel	palatal: pal	dental: den		retroflex: ret	labial: lab		labiodental: lab		glott-al:glo	not cons.: NP		sil: sil
Constituent phonemes	k,g,ng	ch,j,sh,zh	th,dh,ns,z,r,er,l		t,d	p,b,m		f,w,v		hh	all vowels		sil
Class:Consonant Manner													
Subclass: Label	voiced stop: vos	unvoiced st-op:uns	voiced fric..:vos	unvoiced fric:unf	approxi-mant:apx	trill:tri	later al:lat	aspira-tion:apu	nasal: nas	not cons: NM		silence: sil	
Constitiuent phonemes	g,jh,dh,d,b	k,ch,t,p	v,w,z,zh	f,hh,s,sh	y	r,er	l	th	m,n,ng	all vowels		sil	

TABLE III: Phonemic, LDAM1 (Vowel, Consonant Position and Consonant Manner) and LDAM2 (Vowel, Consonant Position and Consonant Manner) transcriptions of the word "average".

Phoneme	ae	ae	ae	ae	ae	t	t	t	t	t	t	v	r	r	r	r	r	r	er	ih	ih	er	ih	jh	jh	jh	jh	jh	jh	jh
LDAM1	fbo	fbo	fbo	fbo	fbo	fbo	NV	NV	NV	NV	NV	daw	NV	NV	NV	NV	NV	NV	fup	cen	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV
	NP	NP	NP	NP	NP	ret	ret	ret	ret	ret	ret	ret	den	den	den	den	den	den	den	NP	NP	NP	NP	pal	pal	pal	pal	pal	pal	pal
	NM	NM	NM	NM	uns	uns	uns	uns	uns	uns	uns	vof	tri	tri	tri	tri	tri	tri	NM	NM	NM	NM	NM	vos	vos	vos	vos	vos	vos	vos
LDAM2	fbo	bmi	fbo	fbo	fbo	NV	NV	NV	NV	NV	NV	NV	tri	tri	tri	tri	tri	tri	tri	fup	fup	fup	tri	NV	NV	NV	NV	NV	NV	NV
	NP	NP	NP	NP	ret	ret	ret	ret	ret	ret	ldn	vel	vel	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	pal	pal	pal	pal	pal	pal	pal
	NM	NM	NM	NM	NM	uns	uns	uns	uns	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	vos	vos	vos	vos	vos	vos

and analysing KL distance matrix. The KL distance matrix (similar to a confusion matrix) was plotted by finding the average KL distance between the phonemic posteriorgrams of all pairs of phonemes (Fig. 1). In this figure, the colour blue indicates lower values of KL distance (greater similarity) between phonemes and red indicates higher KL distance (less similarity) between phonemes. Ideally the diagonal elements should have been blue and non-diagonal elements should have been red. The following observations are made in the phonemes distance matrix:

1) Vowels are more confused with the vowels and consonants are more confused with the consonants. Therefore, it might be better to train vowel and consonant MLPs separately rather than training a single phoneme MLP.
2) Another finer observation is that the semi-vowels (/y/,/w/), lateral (/l/) and trill sounds (/r/,/er/) have comparatively less KL distances with the vowels compared to the consonants. Therefore it might be better to train these phonemes (will be referred to as *vowel-like* phonemes in this paper) with vowels rather than consonants.
3) Nasal phonemes and the phoneme /v/ have similarity with both vowels and voiced stops. So, these phonemes should rather be trained as separate subclass in vowel class too instead of training separately only in consonant classes.

The phonemic and LDAM transcriptions of the word "average" are shown in table III. The labels of the most probable entity in the phonemic, LDAM1 and LDAM2 posteriorgrams are shown in the table. Firstly, looking at the phonemic transcription, it is noticed that /v/ is mislabelled as /t/. It can also be noted that it is mostly labelled as NV (non vowel), 'ret' (retroflex place of articulation) and 'uns' (unvoiced consonant) in the LDAM posteriorgrams. The retroflex unvoiced stop is /t/ (from table II). Therefore, LDAM1 transcription is consistent with the phonemic transcription and the misclassification of /v/ is probably a case of phonemic or co-articulatory variation. But, we want to point to another shortcoming of the LDAM1 configuration through this table. The three *greyed* columns in LDAM1 transcriptions are labelled as NV, NP and NM *(not vowel, not cons., not cons.)* in table III. The phonemic labels for these three columns are /ih/,/er/,/ih/ and the error has occurred while transition from a trill to a vowel. The confusion between vowel and vowel-like phonemes (which are trained to be classified as consonants in LDAM1 mapping) has resulted in the simultaneous occurrence of *none of these* in all the three LDAM1 posteriorgrams. This can cause ambiguity in the representation and matching of the frames. The simultaneous high probability of *none of these* in all the three LDAM1 posteriorgrams was observed in many other words too. Apart from these errors in the LDAM1 transcription, there were some confusions in Consonant Position class of LDAM1 arrangement. The lateral approximant exists in alveolar, retroflex, palatal and velar positions in the IPA chart [18]. Similarly, trill sounds are present for more than one place of articulation. In LDAM1 we take both /l/ and /r/ in the subclass of dental consonants but the actual place of articulation may vary from word to word. In the view of all these observations, LDAM2 configuration of articulatory classes was created by making following changes to LDAM1 configuration:

1) Vowel-like subclasses (semi-vowel /y/, semi-vowel /w/, trills (/r/ and /er/) and lateral /l/) were created in the Vowel class. The phonemes in these subclasses were shifted to *not cons.* (NM) subclass in Consonant Manner class.
2) A subclass for each of the phonemes /h/ and /v/ were also created in the Vowel class. These changes are also inspired by the finding that liquids, glides and /h/ are more closely associated with vocalic than with consonantal segments [11].
3) In the Consonant Manner class (table II), the nasal subclass was merged in the the voiced stop subclass (as the two subclasses are correlated in figure 1). A separate subclass was created for nasal phonemes in Vowel class instead of listing them with *not vowel* (NV) subclass.
4) The frames of the consonants /r/,/l/ were rejected during the training of the Consonant Position class as there is ambiguity about their place of articulation. The semivowels /y/ and /w/ were shifted to *not cons.* (NP) subclass in the Consonant Position class.
5) Plosives and fricatives under the palatal subclass of LDAM1 configuration was split into two subclasses viz. palatal fricatives and palatal plosives. Similarly, the dental subclass was split into two subclasses viz. dental fricatives and dental plosives.

The LDAM2 transcription of the word "average" is also shown in table III. The shortcoming of the LDAM1 transcription mentioned above is resolved in the LDAM2 transcription in the table III.

C. Cross-Validation accuracies of MLPs

TABLE IV: Cross-validation percentage accuracies of MLPs with respective output dimension

Phoneme	**LDAM1**		**LDAM2**	
Accuracy : Dimension	MLP	Accuracy : Dimension	MLP	Accuracy : Dimension
70.89 : 40	Vowel	80.72 : 12	Vowel	78.02 : 20
	Cons. Position	80.22 : 9	Cons. Position	83.36 : 11
	Cons. Manner	79.57 : 11	Cons. Manner	83.93 : 7

Table IV shows the cross-validation (CV) accuracies of the phonemes, LDAM1 and LDAM2 MLPs. The CV accuracy of an MLP is the proportion of patterns in the input of the MLP for which the net output unit with the greatest activation matches the designated target output [19]. The details of CV data and MLP settings is described in section IV. In table IV, it can be seen that the CV accuracy of the phonemes MLP (70.89%) is much less than the accuracy of the individual LDAM MLPs. This is probably because the individual LDAM MLPs have to classify among less number of elements than the phonemes MLP. It is interesting to note that even though the LDAM2 Consonant Position MLP has to classify among 11 constituent elements compared to the 9 elements of the LDAM1 Consonant Position MLP, its accuracy is better than the LDAM1 MLP by 3%.

To further investigate the Consonant Position mapping we compared two MLPs. The first one was the LDAM1 Consonant Position MLP (table II). The second one was the changed version of the first one. The vowel-like phonemes were shifted from their respective subclasses to *not cons.* (NP) subclass in second MLP. Both had the same output dimension i.e. 9. The cross-validation accuracy of the second one was better than the first one by about 3%. This strengthens the belief that vowel-like phonemes should be trained with vowels and not consonants.

D. Combining information from three LDAM Posteriorgrams

Three classes of articulatory features in the LDAM configuration result in a set of three posteriorgrams corresponding to each frame of speech. We call these posteriorgrams as LDAM posteriorgrams. Several functions viz. sum, product and weighted sum of KL distances were investigated for computing the distance between LDAM posteriorgrams of the query $\mathbf{q}_{ldam}$ and the test utterance $\mathbf{u}_{ldam}$. It was found that none of the other functions performed significantly better than the linear equally weighted sum of KL distances between the three posteriorgrams eq.(2) . In this case, the distance between the query and the test utterance posteriorgrams is computed as:

$$f(\mathbf{q}_{ldam}||\mathbf{u}_{ldam}) = kl_{vow} + kl_{cpos} + kl_{cman}, \qquad (2)$$

where f is the distance between the query and the test utterance LDAM posteriorgrams. Here, kl_{vow}, kl_{cpos} and kl_{cman} are the KL distances between the vowel, Consonant Position and Consonant Manner posteriorgrams of the query and the test utterance.

IV. Experiments

A. Database used

The WSJ0 [20] and WSJ1 [21] corpora have been used in all the experiments. Time aligned phonemic transcriptions of around 30000 sentences from WSJ0 corpus used in [5] were used in this work. About 25000 of these sentences were used for training, while the remaining 5000 sentences were reserved for cross-validation purposes.

Three thousand sentences (of duration around 5 hours) were selected from WSJ0 (different from neural network training data) to serve as the evaluation database of test utterances. Sentences containing 36 query terms (for evaluation) for Spoken Term Detection were selected from WSJ1. The number of occurrences of a query in the database ranged between 24 and 40. The development test utterances database also consisted of 3000 utterances and 36 query terms (different from evaluation query terms). The development data was selected from WSJ0 from non-overlapping sections. All the audio data was represented as 13 dimensional MFCCs and concatenated with their first and second order derivatives.

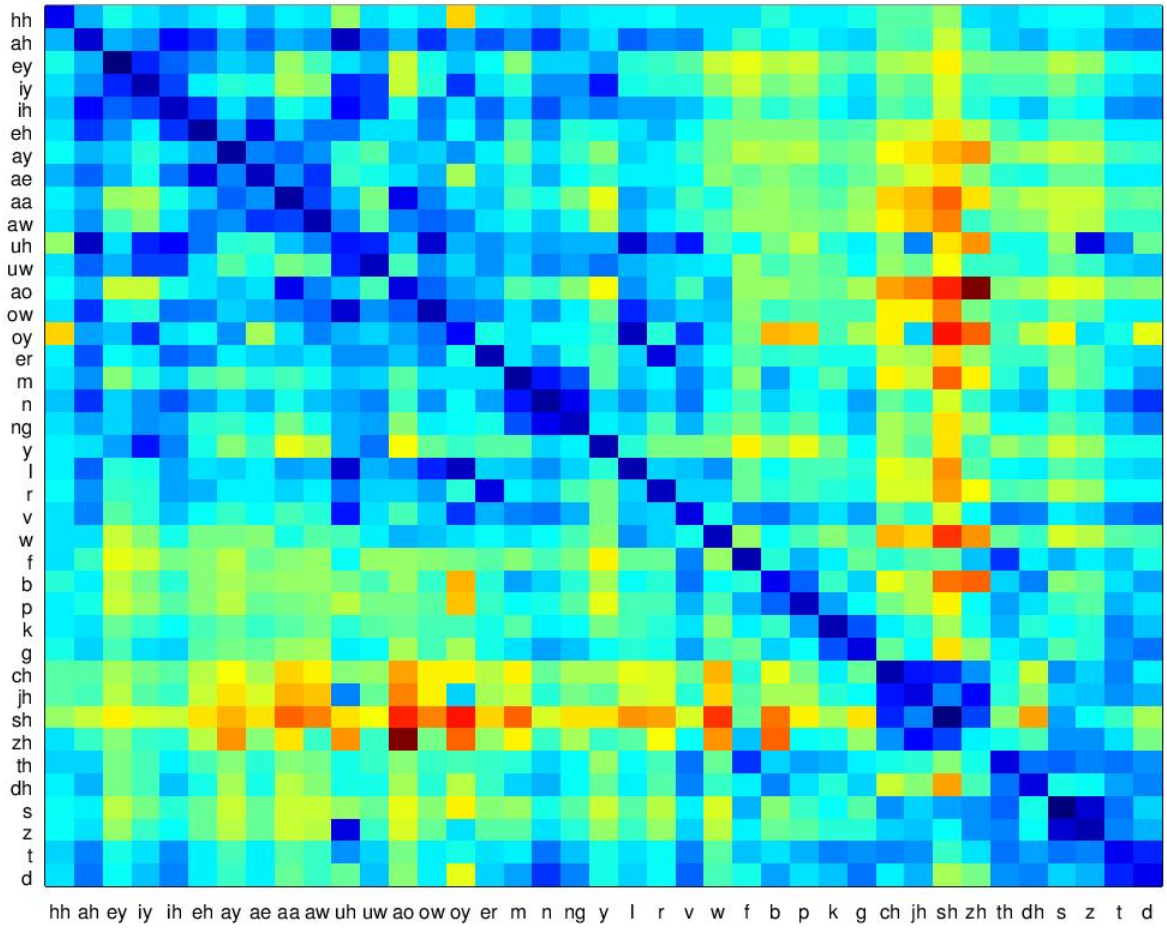

Fig. 1: Phonemes distance matrix; Blue indicates less value of distance (greater similarity) between phonemes while Red indicates higher value of distance (less similarity) between phonemes.

B. MLP settings

We used Quicknet [19] to train the MLPs and derive posteriorgrams. Seven MLPs were trained for the purpose of QbE-STD. One MLP was trained for phonemic posteriorgram and three each for LDAM1 posteriorgrams and LDAM2 posteriorgrams. For all these MLPs, a single hidden layer of 1000 neurons was used. The cross-validation accuracy of phoneme MLP saturates at 1000 neurons [5]. An input layer consisting of 39 units was used for all posteriorgrams. The Output layer consists of 12, 9 and 11 units for the Vowel, Consonant Position and Consonant Manner posteriorgrams respectively in the LDAM1 configuration. In LDAM2 configuration, output layer consists of 20, 11 and 7 units for Vowel, Consonant Position and Consonant Manner posteriorgrams respectively. The targets for training of the LDAM1 MLPs were derived from the mapping of phonemes to subclasses as given in table II. Changes were made in the LDAM1 configuration as given in subsection III-B to obtain LDAM2 configuration. The context window was set to 9 frames (four right and four left) for all the 7 MLPs.

C. Description of the QbE-STD experiment

The LDAM posteriorgrams and the phonemic posteriorgram of the database of 3000 test utterances and utterances containing queries were created using respective MLPs. A phonemic transcription of the utterances containing queries was created by selecting the most probable phoneme in the phonemic posteriorgram of each frame. The boundaries of the queries were determined manually looking at these phonemic transcriptions. The LDAM Posteriorgrams of the query terms were then extracted from the LDAM posteriorgrams of these sentences. The phonemic posteriorgrams of the queries were also created in the same way. DTW [5] was used for matching query posteriorgrams and posteriorgrams of test utterances. The Symmetric KL distance was used for DTW in phonemic posteriorgrams and equally weighted linear sum of the symmetric KL distances eq. (2) was used for LDAM posteriorgrams.

D. Experiment results and discussions

The results of QbE-STD in terms of average P@N over 36 queries are presented in table V for both phonemic and LDAM

posteriorgrams. Precision at N (P@N) is the the proportion of the relevant utterances in the top N utterances where 'N' is the total number of relevant documents in the set of test utterances. The LDAM2 configuration was first tried on the development database and then on evaluation database to make sure that the results are not data dependent. The results of

TABLE V: Average P@N with phonemic and LDAM posteriorgrams

Posterior-gram(s)	Total dimension	Average P@N(dev.)	Average P@N(eval.)
Phoneme	40	85.22	83.62
LDAM1	32	85.14	81.45
LDAM2	38	87.12	86.49

QbE-STD in table V show that the LDAM2 posteriorgrams outperform the phonemic posteriorgram by 2.87% on evaluation data. LDAM2 posteriorgrams perform significantly better than phonemic posteriorgrams ($p<.05$) both on development data and evaluation data using a 1-tailed Wilcoxon signed-ranks test. Since LDAM2 posteriorgrams have better performance than LDAM1 posteriorgrams, it can also be said that the vowel-like subclasses should be included in the Vowel class (like LDAM2) instead of Consonant Manner class (like LDAM1). One novel aspect of LDAM representation is that the vowels and consonants are represented separately and thus there is a possibility of using different kind of features for them. One potential limitation of this approach is that we have assumed the boundaries of the phonemes as boundaries of the articulatory features like place of articulation and manner of articulation.

V. Conclusions

A novel articulatory motivated mapping of phonemes (LDAM1) was proposed with minimal number of classes. This mapping was novel in two ways. Firstly, the coarse classification of phonemes i.e. vowels, consonant manner and consonant position were given the main emphasis rather than delving into many finer articulatory classes. Secondly, the phonemes which can not be categorized completely on the basis of traditional articulatory classes, eg. diphthongs and approximants, were allotted a unique subclass. Thorough experiments and analysis were performed to arrive at an enhanced configuration (LDAM2) in which *vowel-like* phonemes were included in the Vowel class. The efficacy of the enhanced configuration LDAM2 over LDAM1 configuration was confirmed from transcriptions, cross-validation accuracy of MLPs and QbE-STD results. LDAM2 posteriorgrams (total dimension 38) not only outperformed LDAM1 posteriorgrams but also significantly outperformed phonemic posteriorgrams (total dimension 40) which are considered to be one of the state of the art features for QbE-STD.

Acknowledgment

The first author wishes to thank his managers Mr. Shive Narayan, Mr. Biren Karmakar and Mr. Vipin Tyagi, Executive Director, Centre for Development of Telematics (CDOT), India for their encouragement and permission to carry out the research on a part-time basis.

References

[1] K. Thambiratnam and S. Sridharan, "Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting," in *Proc. of ICASSP*, vol. 1, March 2005, pp. 465–468.

[2] D. A. James and S. J. Young, "A fast lattice-based approach to vocabulary independent wordspotting," in *Proc. of ICASSP*, vol. i, Apr 1994, pp. I/377–I/380 vol.1.

[3] M. Saraclar, "Lattice-based search for spoken utterance retrieval," in *In Proceedings of HLT-NAACL 2004*, 2004, pp. 129–136.

[4] Y.-C. Pan and L. shan Lee, "Performance analysis for lattice-based speech indexing approaches using words and subword units," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1562–1574, Aug 2010.

[5] V. Gupta, J. Ajmera, A. Kumar, and A. Verma, "A language independent approach to audio search." in *Proc. of INTERSPEECH*. ISCA, 2011, pp. 1125–1128.

[6] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. of ASRU*, Nov 2009, pp. 421–426.

[7] C.-A. Chan and L.-S. Lee, "Integrating frame-based and segment-based dynamic time warping for unsupervised spoken term detection with spoken queries," in *Proc. of ICASSP*, 2011, pp. 5652–5655.

[8] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams," in *Proc of ASRU*, Nov 2009, pp. 398–403.

[9] X. Anguera, L. J. Rodriguez Fuentes, F. Metze, Szoke, Buzo, and M. Penagarikano, "Query-by-example spoken term detection on multilingual unconstrained speech," in *Interspeech 2014*, Singapore, 14-18 sep. 2014. [Online]. Available: http://gtts.ehu.es/gtts/NT/fulltext/AngueraInterspeech2014.pdf

[10] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, 2002.

[11] S. Chang, M. Wester, and S. Greenberg, "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," *Speech Communication*, vol. 47, no. 3, pp. 290–311, 2005.

[12] R. Prabhavalkar, K. Livescu, E. Fosler-Lussier, and J. Keshet, "Discriminative articulatory models for spoken term detection in low-resource conversational settings," in *Proc.of ICASSP*, May 2013, pp. 8287–8291.

[13] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and . etin, "Articulatory feature classifiers trained on 2000 hours of telephone speech." in *Proc. of INTERSPEECH*. ISCA, 2007, pp. 2485–2488.

[14] G. Mantena and K. Prahallad, "Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios," in *Proc. of ICASSP*, May 2014, pp. 7128–7132.

[15] "The CMU Pronouncing Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[16] E. Martnez-Celdrn, "Problems in the classification of approximants," *Journal of the International Phonetic Association*, vol. 34, pp. 201–210, 12 2004.

[17] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using kl-based acoustic models in a large vocabulary recognition task," IDIAP, Idiap-RR Idiap-RR-14-2008, 0 2008.

[18] "The International Phonetic Alphabet," http://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm.

[19] "Quicknet-ICSI," http://www1.icsi.berkeley.edu/ Speech/qn.html.

[20] "CSR-I (WSJ0) Complete," http://www.ldc.upenn.edu/Catalog/ CatalogEntry.jsp?catalogId=LDC93S6A.

[21] "CSR-II (WSJ1) Complete," http://www.ldc.upenn.edu/Catalog/ CatalogEntry.jsp?catalogId=LDC94S13A.