VTLN-Warped Gaussian Posteriorgram for QbE-STD

Maulik C. Madhavi and Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India,
Email: {maulik_madhavi, hemant_patil}@daiict.ac.in

***Abstract*—Vocal Tract Length Normalization (VTLN) is a very important speaker normalization technique for speech recognition tasks. In this paper, we propose the use of Gaussian posteriorgram of VTLN-warped spectral features for a Query-by-Example Spoken Term Detection (QbE-STD). This paper presents the use of a Gaussian Mixture Model (GMM) framework for estimation of VTLN warping factor. This GMM framework does not require phoneme-level transcription and hence, it can be useful for unsupervised tasks. We propose the iterative approach for VTLN warping factor estimation with two GMM training approaches, namely, Expectation-Maximization (EM) and Deterministic Annealing-Expectation Maximization (DAEM). The VTLN-warped Gaussian posteriorgram gave the better QbE-STD performance. The performance of TIMIT QbE-STD was investigated with different evaluation factors, such as a number of Gaussian components in GMM, various local constraints, and a number of iterations in VTLN warping factor estimation. VTLN-warped Gaussian posteriorgram reduces the speaker-specific variation in Gaussian posteriorgram and hence, it is expected to give better performance than Gaussian posteriorgram.**

I. Introduction

The problem of retrieving the audio documents and detect the presence of query with the help of its spoken example is known as Query-by-Example Spoken Term Detection (QbE-STD) [1]. QbE-STD directly exploits the acoustic-level information for matching between spoken documents and a spoken query without transcribing them into phonemes or words. QbE-STD is important for low-resourced languages and under non-mainstream conditions and hence, it was also called an unsupervised STD [1], [2]. As a part of the MediaEval campaign, the Spoken Web Search (SWS) was started in 2011 [3]. This task involves a language-independent audio search for low-resource languages, which has been held almost every year in MediaEval campaign [1].

QbE-STD task involves the matching between the template representation of spoken query and audio document. The spoken realization of the same word uttered from different speakers may have different duration because the different physiological factors are associated with speech production mechanism. The Segmental Dynamic Time Warping (SDTW) was used to perform feature sequence matching between the spoken query and the test utterance [4]. In SDTW, test feature sequence is separated into overlapping segments that are having the same length as a query. The SDTW needs to be executed multiple times to detect the presence of spoken query, which increases computational requirements. To overcome this computational requirement, subsequence Dynamic Time Warping (subDTW) [5] or non-segmental version of DTW were proposed in [6], [7].

For QbE-STD tasks, a spoken data is converted into the posteriorgram representation that resembles linguistic information. The feature vectors represent the acoustic properties such as formants of the vocal tract. For the better QbE-STD system, feature vectors should represent the linguistic (phoneme) information rather than the speaker-specific information. Unsupervised Gaussian posteriorgram and supervised phonetic posteriorgrams are extensively used to represent audio data [4],[8]. Due to distribution learning capability of Gaussian-Bernoulli Restricted Boltzmann Machines (GBRBM), RBM-based posteriorgrams were found to be comparable to Gaussian posteriorgrams. Restricted Boltzmann Machines (RBM) and Deep Belief Network (DBN) were used for QbE-STD tasks as an alternative to Gaussian posteriorgrams [9], [10].

This paper presents the use of Vocal Tract Length Normalization (VTLN) warping factor estimation for its application to QbE-STD. The conventional method (such as Lee-Rose method [11], [12]) for VTLN warping factor estimation requires a phoneme-level transcription whereas the proposed Gaussian mixture model (GMM) framework does not require a phoneme-level transcription. In this paper, we refer to the Lee-Rose method of VTLN warping factor estimation as Hidden Markov Model (HMM)-based VTLN warping factor estimation (which is a supervised approach as it requires manual phonetic transcription). In addition, the proposed approach uses GMM that can also be *exploited* for Gaussian posteriorgram computation. Hence, the novelty of presented work is to exploit trained GMM for VTLN warping factor estimation and then use VTLN-warped features to re-train the GMM and compute the Gaussian posteriorgram.

II. Vocal Tract Length Normalization

It has been studied in the speech processing literature that for a uniform vocal tract model, the formants of the vocal tract are *inversely* related to the length of the vocal tract [13]. The formant frequencies of vocal tract system are given by,

$$F_n = \frac{(2n-1)v}{4L}, \quad n = 1, 2, \cdots, \tag{1}$$

where L = length of the vocal tract (which is typically *13* cm to *18* cm [12]) and v = velocity of the sound wave ($\approx$ *344* m/s, at sea-level and *70*^{o} F [13]). For instance, formant frequencies of two speakers, namely, *A* and *B* having

ISBN 978-0-9928626-7-1 © EURASIP 2017

average vocal tract length L_A and L_B, respectively, are given by $F_A \propto \frac{1}{L_A}$ and $F_B \propto \frac{1}{L_B}$. This results into $F_A = \alpha_{AB} F_B$, where α_{AB} represents VTLN warping factor associated with only two speakers, namely, *A* and *B*. In practice, the VTLN warping factor is estimated from each utterance w. r. t. to a general speaker model. Human vocal tract length can vary from nearly *13* cm for adult female up to *18* cm for adult male [12]. Due to this, formant frequencies can deviate by *25 %* among various speakers. To reflect this deviation, the VTLN warping factor is generally taken from a set of *13* distinct values (at equally spaced points between *0.88* and *1.12*) [12]. The introduction of the VTLN warping factor creates an adjustment in the frequency analysis to cope with such spectral scaling variations. In general, this is performed by considering different versions of the Mel filterbank (whose center frequencies are scaled linearly). In practice, warping factors are obtained via statistical modeling framework, i.e., Maximum Likelihood Estimation (MLE) [11]:

$$\hat{\alpha} = \underset{0.88 \leq \alpha \leq 1.12}{\arg\max} \; P(X^{\alpha} | \lambda_T, W). \tag{2}$$

Since a closed-form expression of eq. (2) is not available, MLE is computed for all the different warped feature vector $\mathbf{x}_t^{\alpha}$ against model λ_T for a given transcription, W.

A. GMM-based VTLN Warping Factor Estimation

A GMM-based framework differs from an HMM-based framework in terms of the objective function used to estimate the VTLN warping factor. The current work is focused on the linear warping factor estimation which is implemented in the frequency-domain. To that effect, we used Gaussian Mixture Model (GMM) likelihood scores to obtain the VTLN warping factor. The process of VTLN warping factor estimation and modified posteriorgram feature extraction is as follows.

1) *Feature Extraction*: Compute warped features, i.e., $X = [\mathbf{x}_1^{\alpha}, \mathbf{x}_2^{\alpha}, \ldots \mathbf{x}_T^{\alpha}]$ that carries information from different warping factors, namely, $\alpha = 0.88, 0.90, \cdots, 1.12$. Note that the number of distinct values of α is user defined and hence, can be empirically decided.
2) *Initial Training*: Train the GMM without warped features, i.e., $\mathbf{x}_t^{\alpha}$, where $\alpha = 1$, i.e., no VTLN warping. Let the initial GMM model be $\lambda_{init} \sim (\mu_{init}, \Sigma_{init}, w_{init})$. The GMM trained on large data comprises male and female speakers, is expected to have speaker invariant characteristics.
3) *VTLN warping factor estimation*: MLE is computed for all the different warped feature vectors $\mathbf{x}_t^{\alpha}$ against the initial model, λ_{init}, i.e.,
$$\hat{\alpha} = \underset{0.88 \leq \alpha \leq 1.12}{\arg\max} \; P(X^{\alpha} | \lambda_{init}). \tag{3}$$
4) *Retraining GMM*: GMM is re-trained on this optimal warped features, i.e., $\mathbf{x}^{\hat{\alpha}}$. This new model $\lambda_r \sim (\mu_r, \Sigma_r, w_r)$ is different from the earlier GMM model λ_{init}.
5) *Posteriorgram Computation*: Now, the VTLN warping factors of test and query features are estimated against the new GMM model λ_r. Gaussian posteriorgrams are computed based on the estimated VTLN warping factors.

The proposed idea of GMM-based VTLN warping factor estimation can be explained as follows: Let VTLN-warped features be X^{α}, $(0.88 \leq \alpha \leq 1.12)$. Initially, the GMM is trained on unwarped features, i.e., $\alpha = 1$ and hence, $(X^{\alpha} \equiv X^1)$, i.e.,

$$\lambda_{init} = \underset{\lambda}{\text{argmax}} \; P(X^1 | \lambda). \tag{4}$$

Now, VTLN warping factor estimation is performed based on the maximum likelihood estimates (MLE), i.e.,

$$\hat{\alpha} = \underset{0.88 \leq \alpha \leq 1.12}{\text{argmax}} \; P(X^{\alpha} | \lambda_{init}). \tag{5}$$

In the next iteration, we consider VTLN-warped features to build GMM and new model parameters are given by:

$$\lambda^{(1)} = \underset{\lambda}{\text{argmax}} \; P(X^{\hat{\alpha}} | \lambda). \tag{6}$$

This implies $P(X^{\hat{\alpha}} | \lambda^{(1)}) \geq P(X^1 | \lambda_{init})$. Thus, maximization in likelihood results into better Gaussian posteriorgram representation in the following iterations.

Next, we will investigate the relation between two VTLN warping factor estimates on MFCC feature sets. To investigate the effectiveness of the GMM-based VTLN warping factor estimation, the VTLN warping factors are estimated using the GMM and HMM-based approaches for *3696* training utterances of TIMIT database. We employed linear frequency scaling to implement VTLN, i.e., $\alpha = 0.88, 0.90, \cdots, 1.12$. Fig. 1 displays the mapping between these two VTLN warping factor estimates using a supervised Lee-Rose method [12] and the proposed unsupervised method. The diagonal band in Fig. 1 indicates that most of the warping factors obtained through these two techniques are nicely correlated with each other. Moreover, it was observed that around *30-40 %* utterances have the same VTLN warping factor for HMM-based estimation and GMM-based estimation. This analysis shows the potential of proposed GMM-based VTLN warping factor estimation under the absence of transcription.

III. Experimental Setup

A. Experimental Dataset

We used TIMIT training and testing dataset without /SA/ sentences to train the GMM. TIMIT test dataset having (1344) utterances (8 utterances per speaker) were used as audio documents (168 speakers). We have used 84 queries that contain 7 to 20 occurrences in the testing dataset and having at least 6 letters. Spoken queries are taken from the training dataset. All the queries are distributed across all the speakers such that at least one speaker contains at least one query. The performance of QbE-STD is measured in terms of precision@N (p@N) and Mean Average Precision (MAP) [2]. The value of N varies according to the query (from 7 to 20). Besides TIMIT dataset, we have also used MediaEval SWS 2013 dataset. MediaEval SWS 2013 contains two sets, i.e., Dev set (505 queries) and

ISBN 978-0-9928626-7-1 © EURASIP 2017

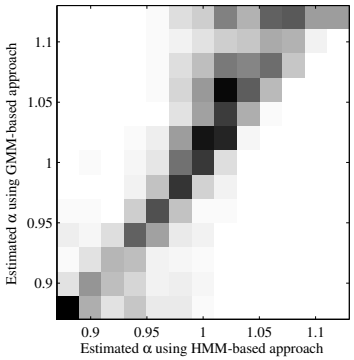

Fig. 1: Estimated values of VTLN warping factor using two different methods, namely, HMM (supervised) and GMM (unsupervised) on TIMIT training database.

Eval set (503 queries). The performance of SWS QbE-STD is evaluated in terms of Maximum Term Weighted Value (MTWV) [14].

B. Feature Extraction

In this paper, MFCC [15] and PLP [16] features are used. The features are extracted on *25* ms window duration, *10* ms frame shift and *26* Mel subband filters and *13* coefficients along with their delta (Δ) and delta-delta (Δ^2) features are considered. MFCC and PLP features are extracted using the Hidden Markov Model Toolkit (HTK) [17].

C. Speech Activity Detection

The phone posteriors were computed by applying the open source Brno University's phoneme recognizer [18]. Czech (CZ), Hungarian (HU), and Russian (RU) phonetic recognizer systems were trained on the CZ, HU and RU SpeechDat-E databases. We performed speech activity detection (SAD) using all the phone posteriorgrams (i.e., CZ, HU, and RU). We considered the average of the posterior probability of non-speech units from CZ, HU and RU to perform SAD.

D. Searching System

The searching subsystem consists of subDTW as searching algorithm [5]. The local distance between two posterior vectors is computed using symmetric Kullback-Leibler (KL) divergence [9]. In addition, pseudo relevance feedback is employed onto first *25 %* retrieved documents (i.e., $0.25 \times 1344 = 336$ in this paper) and we considered top-5 hits as pseudo-relevant example [19].

IV. Experimental Results

A. Effect of Local Constraints (LC)

We analyze the performance of QbE-STD for various local constraints of DTW. Fig. 2 shows three different local constraints for DTW-based searching. The relative temporal mismatch between query and reference due to different speaking rates by various speakers may require additional treatments

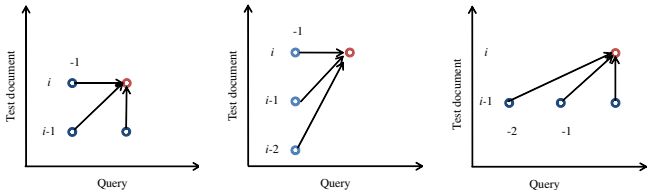

Fig. 2: Types of local constraints used in this study: (a) LC_1, (b) LC_2, and (c) LC_3.

Table 1: Effect of local constraints (LC) on TIMIT QbE-STD systems (p@N performance)

Local constraints	VTLN	p@N		MAP	
		MFCC	PLP	MFCC	PLP
LC_1	×	29.69 (32.98)	31.57 (35.66)	30.38 (35.98)	32.25 (37.89)
	✓	34.47 (38.75)	36.92 (46.90)	35.67 (42.41)	39.28 (49.20)
LC_2	×	33.70 (41.39)	36.36 (41.68)	34.84 (44.06)	37.12 (43.66)
	✓	**37.95** (**46.20**)	**41.11** (**50.69**)	**40.40** (**50.18**)	**42.95** (**52.58**)
LC_3	×	26.35 (29.66)	28.87 (31.45)	27.75 (32.51)	28.86 (33.11)
	✓	31.59 (35.17)	33.35 (41.38)	32.28 (37.52)	34.42 (42.76)

(The number in the brackets indicates the performance after pseudo relevance feedback. × = No VTLN, ✓ = VTLN)

in the search algorithm. In particular, locality consideration while computation of accumulated distance matrix. The feature alignment is performed by similarity matching of consecutive features by considering different local constraints.

Table 1 shows the performance of QbE-STD systems for different local constraints, namely, LC_1, LC_2 and LC_3. It can be observed from Table 1 that LC_2 performs better than other local constraints, probably due to its ability to capture a wide range of features along test utterances. For each local constraint, it can be also observed that VTLN-warped Gaussian posteriorgrams improve QbE-STD performance over Gaussian posteriorgrams.

B. Number of Iterations

In the proposed approach for VTLN warping factor estimation, we initially build a GMM on unwarped (i.e., $\alpha = 1$) features and estimate the appropriate VTLN warping factor using MLE. Now, new VTLN-warped features are used to build a GMM and VTLN warping factor estimation. This process can be executed in iterative manner till VTLN warping factor estimates or few finite times. In this paper, we examined the effect on QbE-STD performance till *5* iterative optimization.

Table 2 shows the performance with various number of iterations used in VTLN warping factor estimation. It can be observed that performance improves as the number of iteration increases. After a certain number of iterations, performance

ISBN 978-0-9928626-7-1 © EURASIP 2017

saturates that might be due to possible overfitting to training dataset.

Table 2: Performance (p@N) of TIMIT QbE-STD systems with various number of iterations

# iter	p@N		MAP	
	MFCC	PLP	MFCC	PLP
×	33.69	36.36	34.83	37.11
1	37.94	41.11	40.39	42.95
2	39.71	42.85	41.99	44.71
3	40.14	43.29	42.92	45.31
4	**41.67**	43.36	43.57	**45.49**
5	41.63	**43.60**	**43.70**	45.35

(× indicates No VTLN, and 1-5=number of iterations used to etimate VTLN warping factor)

C. Number of Gaussian

The number of Gaussian components in Gaussian posteriorgram plays an important role in QbE-STD tasks [4], [6]. In this Section, we investigate the effect of the number of mixture components used in VTLN warping factor estimation on QbE-STD tasks. In particular, we considered *64* and *128* mixture components for GMM training and VTLN warping factor estimation. It can be analyzed from Table 3 that an increasing number of mixture components improves the performance of a QbE-STD system. In addition, performance using the proposed approach is better than the Gaussian posteriorgram. This finding matches a previous study reported in [6]. This might be because of the increasing number of clusters better represents the speech signal at the frame-level. However, increasing number of Gaussians demands additional processing and storage cost and hence, we restrict our experiments till *128* number of clusters. Performance is further improved with the use of pseudo relevance feedback.

Table 3: Effect of the number of Gaussians on TIMIT QbE-STD systems on performance (p@N)

NG	VTLN	p@N		MAP	
		MFCC	PLP	MFCC	PLP
64	×	31.54 (39.96)	32.84 (41.77)	32.9 (42.95)	33.98 (43.03)
	✓	37.27 (39.07)	35.63 (43.49)	43.3 (46.93)	37.94 (46.01)
128	×	33.7 (41.39)	36.36 (37.12)	34.84 (44.06)	37.12 (43.66)
	✓	**37.95** (**46.20**)	**41.11** (**42.95**)	**40.40** (**50.18**)	**42.95** (**52.58**)

(The number in the brackets indicates the performance after pseudo relevance feedback. NG=Number of Gaussians, × = No VTLN, and ✓ = VTLN)

D. Deterministic Annealing Expectation Maximization (DAEM)

Deterministic Annealing Expectation Maximization (DAEM) is an alternative to Expectation Maximization problem where maximization of likelihood problem is

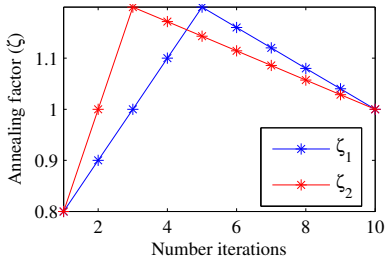

Fig. 3: Values of ζ at every iterations.

posed as minimizing free energy [20]–[22]. This results into modified posterior function that takes into account annealing factor ζ that is inversely proportional to the temperature. The parameters of GMMs in EM framework, i.e., $\theta := \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ can be estimated using EM algorithm. The class assignments for each observation vector $\mathbf{o}_t$ can be made based on the posterior probabilities which is given by [22]:

$$\gamma_t^k = E_{\theta_0}[Z_t^k] = \frac{\tilde{\pi}_k \mathcal{N}(\mathbf{o}_t; \tilde{\mu}_k \tilde{\Sigma}_k)}{\sum_{k=1}^{K} \tilde{\pi}_k \mathcal{N}(\mathbf{o}_t; \tilde{\mu}_k \tilde{\Sigma}_k)}, \tag{7}$$

where $\theta_0 := \{\tilde{\pi}_k, \tilde{\mu}_k, \tilde{\Sigma}_k\}_{k=1}^{K}$ is old parameter values. For DAEM, eq. (7) is modified by annealing parameter ζ as [22]:

$$\gamma_t^k = E_{\theta_0}[Z_t^k] = \frac{(\tilde{\pi}_k \mathcal{N}(\mathbf{o}_t; \tilde{\mu}_k \tilde{\Sigma}_k))^{\zeta}}{\sum_{k=1}^{K} (\tilde{\pi}_k \mathcal{N}(\mathbf{o}_t; \tilde{\mu}_k \tilde{\Sigma}_k))^{\zeta}}. \tag{8}$$

The values of ζ as in Fig. 3. Here, we perform anti-annealing and annealing in DAEM algorithm. As discussed in sub-Section IV-B, we used DAEM against EM for GMM training (number of Gaussians = 128) and performance of QbE-STD is shown in Table 4. It can be seen that performance of DAEM is comparable to EM. This might be due to initial parameters that are set from vector quantization (which is the common for all two DAEM approaches). Again, it can be seen that VTLN-warped Gaussian posteriorgram improves the performance as number of iterations increases. SWS 2013 QbE-STD task, the performance of EM and DAEM (ζ_1) is shown in Table 5. It can be observed that, VTLN-warped Gaussian posteriorgram gave better performance than the Gaussian posteriorgram.

V. Summary and Conclusions

In this study, GMM framework for VTLN warping factor estimation and Gaussian posteriorgram computation is presented for the QbE-STD task. In GMM framework of VTLN warping factor is essentially a grid search w.r.t. the likelihood. This approach does not require transcription to estimate VTLN warping factor. We estimate VTLN warping factor via the Lee-Rose method that uses transcription in the HMM-likelihood framework. We found a high correlation between the estimated VTLN warping factors using both these methods. We exploited GMM-based VTLN warping factor estimation technique for QbE-STD tasks. GMM-based framework can also be exploited

ISBN 978-0-9928626-7-1 © EURASIP 2017

Table 4: Performance of DAEM on TIMIT QbE-STD

VTLN	EM				DAEM (ζ_1)				DAEM (ζ_2)			
	p@N		MAP		p@N		MAP		p@N		MAP	
	MFCC	PLP	MFCC	PLP	MFCC	PLP	MFCC	PLP	MFCC	PLP	MFCC	PLP
×	33.70 (41.39)	36.36 (41.68)	34.84 (44.06)	37.12 (43.66)	33.72 (41.73)	35.95 (42.00)	35.04 (44.55)	37.06 (43.76)	33.72 (41.73)	36.08 (41.84)	35.07 (44.57)	37.01 (43.70)
✓	37.95 (46.20)	41.11 (50.69)	40.40 (50.18)	**42.95** (**52.58**)	37.77 (46.16)	40.91 (50.25)	40.41 (50.12)	**42.96** (**51.92**)	37.84 (46.31)	40.96 (50.40)	40.39 (50.14)	**43.00** (**52.11**)

(× = No VTLN, 1-5=number of iterations used to estimate VTLN warping factor)

Table 5: Performance of DAEM on SWS 2013 QbE-STD (in terms of MTWV)

VTLN	Dev Set				Eval Set			
	EM		DAEM		EM		DAEM	
	MFCC	PLP	MFCC	PLP	MFCC	PLP	MFCC	PLP
×	0.188	0.195	0.188	0.200	0.138	0.145	0.139	0.146
✓	0.209	**0.222**	0.211	**0.222**	0.159	**0.160**	0.159	**0.160**

to extract Gaussian posteriorgrams. Our future work is to explore the possible unsupervised approaches for VTLN warping factor estimation such as Dynamic Frequency Warping (DFW) [23] and elastic registration [24].

REFERENCES

[1] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "Language independent search in MediaEval's Spoken Web Search task," *Computer, Speech and Language*, vol. 28, no. 5, pp. 1066–1082, 2014.

[2] Lin-Shan Lee, James R. Glass, Hung-yi Lee, and Chun-an Chan, "Spoken content retrieval-beyond cascading speech recognition with text retrieval," *IEEE/ACM Trans. on Audio, Speech, & Language Processing,*, vol. 23, no. 9, pp. 1389–1420, 2015.

[3] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. van Heerden, G.V. Mantena, A. Muscariello, K. Prahallad, I. Szoke, and J. Tejedor, "The spoken web search task at MediaEval 2011," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 5165–5168.

[4] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Merano, Italy, 2009, pp. 398–403.

[5] Meinard Müller, *Information Retrieval for Music and Motion*, vol. 2, Springer, 2007.

[6] G. Mantena, S. Achanta, and K. Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *IEEE/ACM Trans. on Audio, Speech & Language Process.,TASLP*, vol. 22, no. 5, pp. 946–955, 2014.

[7] Luis Javier Rodríguez-Fuentes, Amparo Varona, Mikel Peñagarikano, Germán Bordel, and Mireia Díez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Florence, Italy, 4-9 May 2014, pp. 7819–7823.

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

[9] P. R. Reddy, S. Nayak, and K. S. R. Murty, "Unsupervised spoken word retrieval using Gaussian-Bernoulli restricted Boltzmann machines," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 1737–1741.

[10] Y. Zhang, R. Salakhutdinov, H. A. Chang, and J. Glass, "Resource configurable spoken query detection using deep Boltzmann machines," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 5161–5164.

[11] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Atlanta, Georgia, USA, 1996, pp. 353–356.

[12] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech & Audio Process.*, vol. 6, no. 1, pp. 49–60, 1998.

[13] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education, 2006.

[14] Luis Javier Rodríguez-Fuentes and Mikel Peñagarikano, "MediaEval 2013 spoken web search task: System performance measures," Tech. Rep. TR-2013-1, Department of Electricity and Electronics, University of the Basque Country, 2013.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, & Signal Proces.*, vol. 28, no. 4, pp. 357–366, 1980.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoust. Society of America (JASA)*, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[18] Petr Schwarz, Pavel Matejka, Lukas Burget, and Ondrej Glembek, "Phoneme recognizer based on long temporal context," *Speech Processing Group, Faculty of Information Technology, Brno University of Technology.[Online]. Available: http://speech.fit.vutbr.cz/en/software*, 2003, (Last Accessed on 4^{th} March, 2016).

[19] C. Chan and L. Lee, "Integrating frame-based and segment-based dynamic time warping for unsupervised spoken term detection with spoken queries," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 5652–5655.

[20] Naonori Ueda and Ryohei Nakano, "Deterministic annealing EM algorithm," *Neural Networks*, vol. 11, no. 2, pp. 271–282, 1998.

[21] N. J. Shah, H. A. Patil, M. C. Madhavi, H. B. Sailor, and T. B. Patel, "Deterministic annealing EM algorithm for developing TTS system in Gujarati," in *9th Int. Symp. on Chinese Spoken Language Process. (ISCSLP), 2014*, Singapore, 2014, pp. 526–530.

[22] Iftekhar Naim and Daniel Gildea, "Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients," in *Proc. of the* 29^{th} *Int. Conf. on Machine Learning, ICML 2012*, Edinburgh, Scotland, UK, June 26 - July 1 2012.

[23] Edward P Neuburg, "Frequency warping by dynamic programming," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, New York, USA, 1988, pp. 573–575.

[24] Florian Müller and Alfred Mertins, "Enhancing Vocal Tract Length Normalization with Elastic Registration for Automatic Speech Recognition," in *Proc. INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 1364–1367.

ISBN 978-0-9928626-7-1 © EURASIP 2017