Vocal Tract Length Normalization Features for Audio Search

Maulik C. Madhavi[1(✉)], Shubham Sharma[2], and Hemant A. Patil[1]

[1] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
{maulik_madhavi,hemant_patil}@daiict.ac.in
http://www.daiict.ac.in
[2] Indian Institute of Science, Bangalore, India
shubham@mile.ee.iisc.ernet.in
http://www.iisc.ernet.in

Abstract. This paper presents speaker normalization approaches for audio search task. Conventional state-of-the-art feature set, *viz.*, Mel Frequency Cepstral Coefficients (MFCC) is known to contain speaker-specific and linguistic information implicitly. This might create problem for speaker-independent audio search task. In this paper, universal warping-based approach is used for vocal tract length normalization in audio search. In particular, features such as scale transform and warped linear prediction are used to compensate speaker variability in audio matching. The advantage of these features over conventional feature set is that they apply universal frequency warping for both the templates to be matched during audio search. The performance of Scale Transform Cepstral Coefficients (STCC) and Warped Linear Prediction Cepstral Coefficients (WLPCC) are about *3*% higher than the state-of-the-art MFCC feature sets on TIMIT database.

Keywords: Vocal tract length normalization · Audio search · Scale transform cepstral coefficients · Warped linear prediction coefficients

1 Introduction

Recently, speech-based search or retrieval technologies have gained keen attention. The simple reason could be the ease of their storage and retrieval capabilities. A speech signal is enriched with many attributes such as message conveyed from it, speaker information, emotion, age, gender, etc. Several studies have been involved in extracting such kind of information from the speech signal. National Institute Science and Technology (NIST) has started evaluation named Spoken Term Detection (STD) in *2006* which is involved in linguistic information extraction from spoken document [1]. The technology has a quite relevance with the Automatic Speech Recognition (ASR) task. However, ASR is slightly different technology where speech signal is decoded in terms of word transcription whereas STD is only interested to detect particular audio query within database [2].

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 387–395, 2015.
DOI: 10.1007/978-3-319-24033-6_44

Inspite of being different technology all together, speech researchers have attempted STD problem via ASR system [2],[3]. For an ASR experts, the task remains merely obtaining the transcripts and assigning them confidence measure. Word-based ASR has found to be effective on well resourced languages. However, this may introduce problem in out-of-vocabulary (OOV) query such as *named entities*. To deal with this issue, researchers have come up with an subword modeling approach [4].

For low-resourced languages, it is hard to find the labeled speech corpora. Hence, speech researchers have exploited the spoken query representation rather than textual representation. Considering the fact that query is taken from speech only, the STD technology is now termed as Query-by-Example Spoken Term Detection (QbE-STD) [5], [6]. In this paper, we refer QbE-STD as *audio search*, which basically involves matching and hence, highly dependent on the type of *representation* used in the matching task. In particular, representation used in audio search should be speaker-invariant so as to perform speaker-independent audio matching. In this paper, we explore universal frequency warping in two spectral features, *viz.*, Scale Transform Cepstral Coefficients (STCC) and Warped Linear Prediction Cepstral Coefficients (WLPCC) in order to develop speaker-invariant features for audio search task.

2 Relation to Prior Work

There have been many attempts made in designing the proper representation of QbESTD. Speaker-invariant representation has been primary need for speaker-independent audio search. There have been significant attempts made in posterior-based feature representation. They are mainly supervised phonetic posteriorgram and Gaussian posteriorgram.

Posteriogram representation is found to be more robust to speech information [6],[7]. They also used Gaussian Mixture Model (GMM)- based posteriorgram representation in their studies. The K-means clustering algorithm is combined with the GMM posteriorgrams front-end to obtain more discriminant features [8]. Parallel tokenizer-based approach for QbE-STD was used in [9] to combine the evidences from different systems. Each tokenizer uses both posteriorgram of query and utterance and combines the evidences from all [9]. In addition to the posteriorgram features, researchers have exploited speaker normalization technique, *viz.*, Vocal Tract Length Normalization (VTLN) [9], [10] for audio search task. The VTLN method discussed in those studies exploit different version of Mel warped filterbank. The optimum *warping factor* is estimated using grid-search under Maximum Likelihood (ML) criteria. This requires features to be computed for all different warped Mel filterbank and then estimation for proper warping factor using state-of-the-art Lee and Rose method [11]. In this paper, we have used universal warping-based approach of VTLN for audio search [12]. In addition, we have explored two different feature sets, *viz.*, STCC and WLPCC. These features avoid exhaustive grid search and hence, less cumbersome in terms of computation yet performing well in speaker mismatched condition. The presented approach does

not compare the absolute performance of audio search system presented earlier in the literature. The objective and focus of this work is to bring signal processing aspect at front end of audio search task.

3 Vocal Tract Length Normalization (VTLN)

In this paper, we have compared the performance of state-of-the-art mel frequency cepstral coefficients (MFCC) with vocal tract length (VTL) normalized features such as STCC [13] and WLPCC. For feature extraction, speech signal is divided into frames and each frame is Hamming windowed. In the following sub-Sections, we discuss and compare various feature extraction methods used in this study.

3.1 Scale Transform Cepstral Coefficients (STCC)

STCCs compensate for the differences in VTL using log-warping. Vocal tract is generally modelled as a uniform tube. For such a model, formant frequencies are inversely proportional to VTL. The spectrum of a particular speech sound of a speaker is the scaled version of the spectrum of another speaker uttering the same sound, i.e., $F_A(\omega) = F_B(\alpha_{AB}\omega)$. The scale α_{AB} is speaker-specific and is known as *warping factor*. Replacing ω by e^v, we get,

$$f_A(v) = F_A(\omega = e^v) = F_B(\alpha_{AB}e^v) = F_B(e^{(v+\ln\alpha_{AB})}) = f_B(v + \ln\alpha_{AB}). \quad (1)$$

This shows that in the log-warped domain, the spectra are shifted versions of each other with a translation factor of $\ln\alpha_{AB}$. Since Fourier transform is a shift-invariant transform and the shift factor appears in the phase part, the speaker-dependent warping factor is removed by taking the magnitude. These are used as VTL normalized features. Smoothed spectra are obtained by suppressing the pitch information by the method described in [12]. Here, mel-warping is used as it was observed that mel-warping provides better performance [14].

3.2 Warped Linear Prediction Cepstral Coefficients (WLPCC)

This feature set provides Bark scale-based frequency warping via warped linear prediction (WLP). WLP is obtained by replacing unit delays of classical LP filter by first-order all pass filters with transfer function given by [15],

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \quad (2)$$

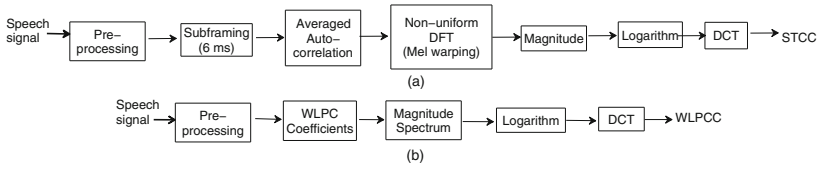

Fig. 1. Schematic block diagram for feature extraction of (a) STCC and (b) WLPCC. After, [12],[15].

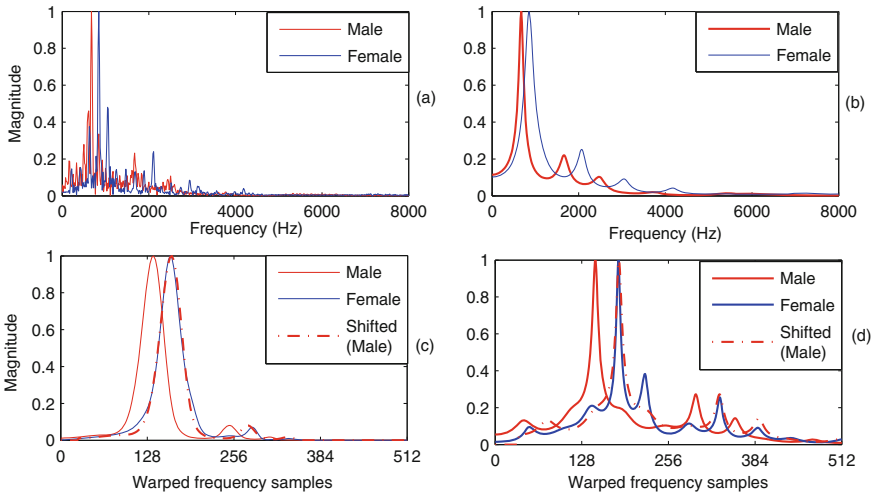

Fig. 2. Illustration of speaker normalization for vowel segment /ae/ of two different (male and female) speakers. (a) magnitude spectrum, (b) LP magnitude spectrum (c) Mel warped Fourier transform spectrum (STCC spectrum), and (d) Bark warped LP spectrum (WLPCC spectrum). In Fig. 2, (c) and (d), the spectra plotted with dotted line indicates that speaker differences are normalized.

and phase response [15],

$$\Psi(\omega) = \omega + 2\tan^{-1}\left(\frac{\lambda \sin\omega}{1-\cos\omega}\right), \tag{3}$$

where $-1 < \lambda < 1$ is the warping factor. For $0 < \lambda < 1$, lower frequencies are compressed and higher frequencies are expanded. The reverse warping happens for $0 > \lambda > -1$. An analytical expression provides the value of λ for warping similar to Bark scale [15] depending on the sampling frequency, i.e., f_s and given by,

$$\lambda_{fs} \approx 1.0674\left(\frac{2}{\pi}\arctan\left(0.6583\frac{fs}{1000}\right)\right)^{\frac{1}{2}} - 0.1916, \tag{4}$$

WLP coefficients (WLPC) are easily obtained by Levinson-Durbin algorithm using warped autocorrelation function. Bark scale-warped LP spectrum is obtained by the WLPCs. Cepstral features are obtained by taking DCT of log of the warped spectra [16]. Detailed procedure of STCC and WLPCC feature extraction is shown in Fig. 1. Here, for TIMIT dataset $fs = 16$ kHz, which corresponds to $\lambda_{fs} = 0.575$. Fig. 2 shows effectiveness of Mel and Bark warping for VTLN for vowel (/ae/) spoken by a male and a female subject. It can be observed that these spectra overlap and hence, speaker variability reduces. That can be useful evidence in the design of audio matching application.

Table 1. Statistics of query used in this work (# Train:#Test)

Query Index	Query	Query Index	Query
1	age (3:8)	2	artists (7:6)
3	children (18:10)	4	development (9:8)
5	money (19:9)	6	organizations (7:6)
7	problem (22:13)	8	surface (3:8)
9	warm (10:5)	10	year (11:5)

4 Experimental Results

4.1 Database Used

For audio search task, we have used TIMIT corpora [17]. *10* queries are taken from training database and the details are shown in Table 1. The selection of query word is as given in [7]. TIMIT dataset consists of *3,696* training utterances and *944* test utterances. Audio search queries are taken from training set and searching is performed on testing set.

4.2 Architecture of Audio Matching System

For audio search task, the state-of-the-art dynamic programming-based Dynamic Time Warping (DTW) [18] and their variant such as segmental DTW [19] have been used prominently for audio matching. For audio search task, two-pass strategy has been employed. In the first pass, segmental DTW is performed between query word and reference utterances in order to obtain hypothetical location of query within reference utterance. The segmental DTW is an extended version of DTW. In particular, in audio search task, reference data is utterance whereas query is merely a word or a phrase. The direct application of DTW between these two signals would not make much sense as both corresponds to different length information. However, we are not aware of the duration of the query present in the actual reference data. We need to perform segmental DTW using $R = 5$. We may call this task-1 as '*localization of query*'. The task-2, the remaining unwanted reference is chopped out and conventional template matching using DTW. This phase can be called as '*scoring*' as we need to rank the audio document based on the minimum DTW distance value.Conventional Euclidean distance is used to compute the local distance between two patterns via DTW algorithm.

4.3 Results and Discussion

To evaluate the performance of audio search following evaluation measures are considered, namely, 1) $p@N$ precision at N, (i.e., number of hits in top N retrieved documents, where N corresponds to the number of actual document present in database) 2) % EER: Equal Error Rate.

Table 2. Experimental Results

Feature	p@N	%EER	Feature	p@N
MFCC	24.65	25.30	MFCC-fused	40.17
STCC	27.98	23.73	STCC-fused	44.68
WLPCC	27.13	23.25	WLPCC-fused	41.29

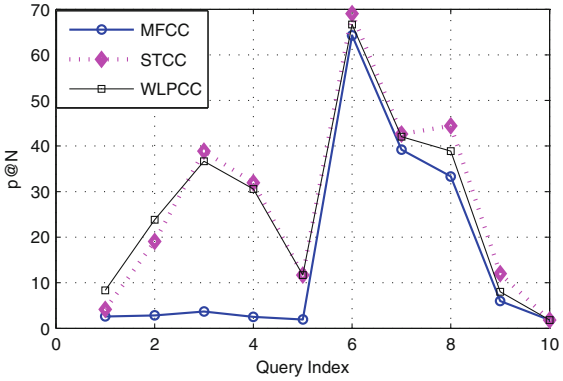

Fig. 3. Performance of audio search for each individual query.

Overall performance of the audio search system is shown in Table 2. It can be observed that the VTLN-based feature sets, namely, STCC and WLPCC performs better than MFCC alone. About *3* % absolute improvement can be observed using STCC and WLPCC features. In Table 2, the performance of each isolated query of each feature sets are mentioned as MFCC, STCC and WLPCC, respectively. In addition, distortion score from the same query are fused. This will improve the statistical confidence about the query detection task. The fused features are called as MFCC-fused, STCC-fused and WLPCC-fused, respectively. From Table 2, it can be observed that the fusing of multiple evidences indeed improve the audio search performance for all *3* feature sets. Simple averaging of distortion score is used as fused score. 5^{th} column of Table 2 shows the performance using fused score improves. In order to investigate the performance *w. r. t.* every query, mean $p@N$ is computed for individual query. From Fig. 3, it can be observed that for most of the query word, STCC and WLPCC performs better than MFCC. The query associated with particular query index is listed in Table 1.

4.4 Evaluation of Class Separability

In order to investigate the effectiveness of these feature sets, we conducted class separability test. For different vowel class J-measure is computed. Separability property of MFCC, STCC and WLPCC is compared for four vowels /iy/, /ih/, /ae/ and /ow/ taken from *13* female and *22* male speakers of *dr7* of TIMIT database. Middle *32* ms speech signal of *150* examples of each vowel is considered. Separability is measured as suggested in [20].

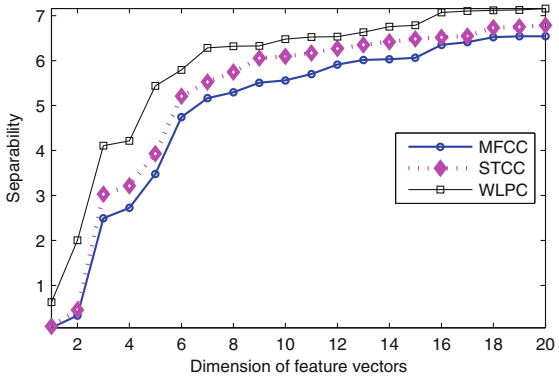

Fig. 4. Class separability of vowels for different feature sets, *viz.*, MFCC, STCC, WLPCC.

$$J = tr(S_W^{-1} S_B), \tag{5}$$

where S_B is the between groups mean square and S_W is the within groups mean square. S_B and S_W are obtained as

$$S_B = \frac{1}{I}\sum_{i=1}^{I}(M_i - M_o)(M_i - M_o)^T,\ S_W = \frac{1}{I}\sum_{i=1}^{I} R_i,$$

where M_i and R_i denote the mean feature vector and covariance matrix, respectively and $M_0 = \frac{1}{I}\sum_{i=1}^{I} M_i$, where I is the total number of phoneme classes being compared. Fig. 4 shows the separability of different feature extraction methods *w. r. t.* dimension of feature vectors. It can be observed that the separability of STCC is higher than that of MFCC and still higher for WLPCC. This along with normalization of speaker-specific spectra (as shown in Fig. 2 (c) and Fig. 2 (d) may be the reason for better performance of STCC and WLPCC feature than MFCC for audio search task.

5 Summary and Conclusions

This paper presented audio search system using VTLN-based approach. There have been several ways to exploit VTLN in order to suppress the speaker variation for speaker-independent audio search task. This work involved universal warping-based feature extraction, namely, STCC and WLPCC. Performance of audio search system is found to be improved when these representation are used instead of conventional MFCC. Our future plan is to incorporate other VTLN aspects such as estimation of frequency warping relation for speaker pair to improve the audio search performance. In addition, we would like to explore telephone recorded speech for audio search task.

Acknowledgments. The authors would like to thank Department of Electronics and Information Technology (DeitY), Government of India for sponsoring the project and the authorities of DA-IICT for their support to carry out this research work. This work is partially supported by the project "Indian Digital Heritage (IDH) - Hampi" sponsored by Department of Science and Technology (DST), Govt. of India (Grant No: NRDMS/11/1586/2009/Phase-II).

References

1. The Spoken Term Detection (STD) 2006 Evaluation Plan (2006). http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf (last accessed on March 25, 2015)
2. Vergyri, D., Shafran, I., Stolcke, A., Gadde, V.R.R., Akbacak, M., Roark, B., Wang, W.: The SRI/OGI 2006 spoken term detection system. In: INTERSPEECH 2007, Belgium, pp. 2393–2396 (2007)
3. Parlak, S., Saraclar, M.: Spoken term detection for turkish broadcast news. In: Proc. IEEE Int. Conf. on Acous. Speech, and Signal Process. ICASSP 2008, Las Vegas, USA, pp. 5244–5247 (2008)
4. Wallace, R., Vogt, R., Sridharan, S.: A phonetic search approach to the 2006 NIST spoken term detection evaluation. In: INTERSPEECH 2007, Belgium, pp. 2385–2388 (2007)
5. Metze, F., Anguera, X., Barnard, E., Davel, M.H., Gravier, G.: Language independent search in mediaeval's spoken web search task. Computer Speech & Language **28**(5), 1066–1082 (2014)
6. Hazen, T.J., Shen, W., White, C.M.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU, 2009, Merano/Meran, Italy, pp. 421–426 (2009)
7. Zhang, Y., Glass, J.R.: Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU, 2009, Merano/Meran, Italy, pp. 398–403 (2009)
8. Anguera, X.: Speaker independent discriminant feature extraction for acoustic pattern-matching. In: IEEE Int. Conf. on Acoust. Speech and Signal Process., ICASSP 2012, Kyoto, Japan, pp. 485–488 (2012)
9. Wang, H., Lee, T., Leung, C., Ma, B., Li, H.: Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection. In: IEEE Int. Conf. on Acoust. Speech and Signal Process., ICASSP 2013, Vancouver, BC, Canada, pp. 8545–8549 (2013)
10. Tejedor, J., Szöke, I., Fapso, M.: Novel methods for query selection and query combination in query-by-example spoken term detection. In: Proc. of 2010 Int. Workshop on Searching Spontaneous Conversational Speech. SSCS 2010, New York, NY, USA, pp. 15–20. ACM (2010)
11. Lee, L., Rose, R.C.: Speaker normalization using efficient frequency warping procedures. In: IEEE Int. Conf. on Acoust. Speech and Signal Process., ICASSP 1996, Atlanta, Georgia, USA, pp. 353–356 (1996)
12. Umesh, S., Cohen, L., Marinovic, N., Nelson, D.J.: Scale transform in speech analysis. IEEE Transactions on Speech and Audio Processing **7**(1), 40–45 (1999)
13. Umesh, S., Sanand, D.R., Praveen, G.: Speaker-invariant features for automatic speech recognition. In: IJCAI 2007, Proc. 20th Int. Joint Conf. on Artificial Intelligence, Hyderabad, India, pp. 1738–1743 (2007)

14. Sinha, R., Umesh, S.: Non-uniform scaling based speaker normalization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, May 13–17 2002, Orlando, Florida, USA, pp. 589–592 (2002)
15. Iii, J.O.S., Abel, J.S.: Bark and ERB bilinear transforms. IEEE Transactions on Speech and Audio Processing **7**(6), 697–708 (1999)
16. Kim, Y., Smith, J.O.: A speech feature based on bark frequency warping-the non-uniform linear prediction (nlp) cepstrum. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1999, pp. 131–134. IEEE (1999)
17. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM (1993)
18. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics Speech and Signal Processing **26**(1), 43–49 (1978)
19. Park, A.S., Glass, J.R.: Unsupervised pattern discovery in speech. IEEE Transactions on Audio, Speech & Language Processing **16**(1), 186–197 (2008)
20. Nicholson, S., Milner, B.P., Cox, S.J.: Evaluating feature set performance using the f-ratio and j-measures. In: Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece (1997)