Broad Phonetic Classes for Speaker Verification with Noisy, Large-Scale Data

Howard Lei and Nikki Mirghafori

TR-14-002

August 2014

Abstract

While the incorporation of phonetic information has contributed to speaker verification improvements for lexically unconstrained speech in the past, improvements have not been widely observed using the state-of-the-art i-vector system, which typically performs best using a "bag-of-frames" approach. This work explores ways to incorporate Broad Phonetic Class (BPC) information for the i-vector system with noisy speech data that is not lexically constrained. Different approaches for combining the BPCs have been examined. Results suggest that, through parallelization and combination strategies, BPCs may contribute to roughly a 13% improvement over an i-vector baseline system. However, confounding factors such as increased parameter size, use of noise-generated speech data, and the advantage of combination strategies are potential caveats to attributing the improvement to the discriminating power of BPCs alone.

This work was funded by Air Force Research Laboratory (AFRL) award FA8750-12-1-0016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the AFRL.

1. INTRODUCTION

It has been hypothesized that for purposes of human communication, phonetic sounds have emerged to correspond to both human speech production (i.e., articulatory apparatus) as well as the human perception system (e.g., cochlear structure). In human speaker verification, phonetic classes provide a cognitive structure around which the acoustic information for verification is organized. In automatic speaker verification using conversational data that's not lexically constrained, however, the most common and popular use of the state-of-the-art i-vector system treats the acoustic information as a bag of frames, without regard to phonetic structure. While this "bag of frames" approach has led to state-of-the-art performances for i-vector systems, the systems can perhaps be improved by "enriching" their GMM models with phonetic sounds that correspond to human speech production and perception.

While incorporating phonetic information has not been commonly used with the state-of-the-art i-vector using data that's not lexically constrained, there has been a significant amount of work using past speaker verification approaches. The work of Omar and Pelecanos in 2010 investigated a GMM supervector system where the UBM mixture components are derived from mixture components of pre-trained acoustic models for automatic speech verification [1]. This approach performed better in the majority of the SRE08 core-core conditions compared to a maximum likelihood approach to UBM training, and provides a framework for the parallel initialization of UBMs [1]. The work of Gutman and Bistriz in 2002 investigated the use of phoneme-specific GMMs in a GMM-UBM system, where GMMs of each phoneme were either used in separate speaker verification systems, or as part of one system [2]. The work of Faltlhauser and Ruske in 2001 dealt with a parallel GMM speaker verification system approach, where each parallel component represented a system for a specific phoneme or group of phonemes. The work of Baker et al. in 2005 [3] investigated the use of broad phonetic categories (BPC) and syllabic events for speaker verification, using the GMM-UBM system. The work of Lei and Lopez-Gonzalo in 2009 investigated the use of simple BPCs [4] in a Factor Analysis-based GMM-UBM system. The work of Scheffer et al. [5] investigated phonetic combinations at the factor loading matrix-level using a Joint Factor Analysis (JFA) system. Lastly, Larcher et al. investigated the effect of incorporating phonetic information on i-vector system performance, but only on phonetically-constrained short utterance data [6]. This work investigated whether having target and test conversation sides with matched phonetic information improves speaker verification, and not whether the incorporation of phonetic information in speaker models is beneficial to speaker verification using lexically-unconstrained data.

In this work, we investigated the use of phonetic information for the i-vector system, in an attempt to discover scenarios in which use of phonetic information would be helpful with lexically-unconstrained data. In particular, we focused the modeling power of our systems on phonetic regions of speech by implementing standalone systems only on speech data regions containing certain phonetic content, such as vowels or nasals. Different ways of combining the standalone phonetic systems, were also investigated.

This paper is organized as follows: Section 2 describes the data used. Section 3 describes the phonetic units. Section 4 describes the speaker verification system and combination approaches. Section 5 describes the results, and Section 6 provides a summary and concerns that must be addressed in any future work.

2. DATA

All gender-dependent experiments were performed using SRE08 and SRE10 noiseless telephone data, either with or without added noise. We generated the noise-added data using the noiseless data. The noiseless data contains 706 male speakers and 1,059 females speakers, with 4,587 male and 7,037 female target speaker training conversation sides, and 1,041 male and 1,427 female test conversation sides. We note that the target speaker training conversation sides were also used as i-vector system development conversation sides, which conforms to the recent NIST Speaker Recognition Evaluation 2012 framework. To generate the noise-added conversation sides, samples of crowd noise and car exhaust noise – too common noises encountered in the environment – were first obtained from the publicly available audio data source *freesound.org* [7].

Each noise sample was repeated enough times such that their overall duration exceeds the conversation side durations. Next, four copies of each noiseless telephone conversation side were made, where each copy is mixed with one of the two noise samples at 10dB or 20dB SNR. The publicly available Filtering and Noise Adding Tool (FaNT) Toolkit [8] was used to mix the speech with the noise samples. The four combinations of noise and SNR – crowd at 10dB SNR, crowd at 20dB SNR, car exhaust at 10dB SNR, and car exhaust at 20dB SNR – were implemented for all conversation sides. The overall dataset consists of the four noisy copies of each conversation side, along with the original noiseless versions. There are a total of 119,133,175 trials with 243,500 true speaker trials for males, and 250,706,375 trials with 338,600 true speaker trials for females.

3. PHONETIC UNITS

The phonetic units include a set of 5 Broad Phonetic Classes (BPCs): Vowels, Nasals, Glides/Liquids, Fricatives, and Stops. Using a subset of the male data, we found that Vowels comprise of 51% of the data, Fricatives comprise 16%

of the data, Stops comprise 14% of the data, Nasals comprise 10% of the data, and Glides/Liquids comprise 9% of the data. While other phonetic units, such as the smaller Di-BPC (which are syllabic in nature) and phone units, can be used as well, but past experiments have shown that such units suffer from data sparsity. Note that phonetic labels for all conversation sides were obtained from SRI's DECIPHER phone recognizer [9] output, which uses forced-aligned phone decodings from automatic speech recognition word output.

4. SPEAKER VERIFICATION SYSTEMS

The i-vector speaker verification system was primarily used in our experiments. For each conversation side, 60-dimensional MFCC C0-C19 + Δ + $\Delta\Delta$ acoustic features are extracted, and Gaussian feature warping [10] is performed cross 3-second windows. A background GMM model, or Universal Background Model (UBM), is trained using the features from the development conversation sides. Zeroth, first, and second order Baum-Welch sufficient statistics are extracted for each conversation side using the UBM, and a low-rank Total Variability Matrix (T-Matrix) is trained using the development conversation sides statistics. This matrix captures the overall variability of the development conversation side statistics.

The T-matrix, first order statistics, and the UBM are used as follows to extract low-dimensional i-vectors for each conversation side:

$$M = m + T\omega \tag{1}$$

where T is the T-matrix, m is a vector of the UBM means, and ω are low-dimensional vectors, known as the identity vectors" or i-vectors [11]. The i-vectors ω for each conversation side were obtained via the approaches described in [12] and [13].

Once the i-vectors are extracted, a Probabilistic Linear Discriminant Analysis (pLDA) likelihood ratio [14] [15] is used to generate scores for each trial. The pLDA likelihood ratio is shown in Equation 2:

$$score(\omega_1, \omega_2) = logN\left(\begin{bmatrix}\omega_1\\ \omega_2\end{bmatrix}; \begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & \Sigma_{bc}\\ \Sigma_{bc} & \Sigma_{tot}\end{bmatrix}\right) - logN\left(\begin{bmatrix}\omega_1\\ \omega_2\end{bmatrix}; \begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{bmatrix}\right)$$

where ω_1 and ω_2 are the two i-vectors, $N(\cdot)$ is the normal Gaussian probability density function, and Σ_{tot} and Σ_{bc} are the total and between-class scatter matrices of the training i-vectors. Hence, given a pair of i-vectors, the approach computes a likelihood ratio of Gaussian distributions centered upon the i-vectors. Each distribution is parameterized by the the scatter matrices obtained from the training data. Note that prior to applying the likelihood ratio, the i-vectors are WCCN-normalized, and length-normalized [16] to be of unit length.

We note that for all experiments, the open-source ALIZE toolkit [17] is used for GMM model training. The Brno University of Technology's (BUT's) Joint Factor Analysis Matlab demo [18] is used to assist in i-vector extraction, and the HTK software [19] is used for acoustic feature extraction.

Three approaches were used to combine standalone BPC systems with one another as well as the baseline system. The first is the concatenation of i-vectors of each BPC unit to form longer i-vectors. Figure 1 illustrates this approach, which we refer to as "i-vector concatenation."

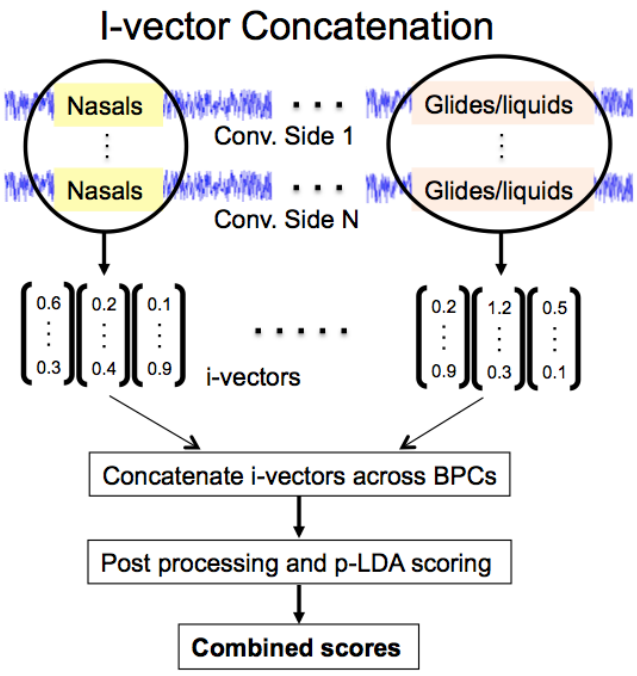

Fig. 1. *Generating BPC-combined scores for the i-vector system via i-vector concatenation.*

The second approach involves a simply averaging the scores for the standalone systems, and the third approach involves training logistic regression score combination using the Bosaris Toolkit [20]. The logistic regression function is training using a set of scores generated from all target speaker conversation side pairs. Note that we have also attempted a T-matrix-level combination technique inspired by the work of Scheffer et al. [5], but could not get performance improvements over the i-vector concatenation combination.

In addition to the i-vector system, the GMM-UBM system was used to verify the effectiveness of BPCs as demonstrated in prior GMM-UBM-based experiments, such as in [3]. The GMM-UBM approach involves training target speaker GMM models via MAP adaptations of the UBM for each target speaker conversation side, followed by the computation of the log-likelihood of the MFCC features of each test conver-

sation side for a set of target speaker GMM models [21].

5. EXPERIMENTS AND RESULTS

Initial experiments consist of comparing the performance of the i-vector-concatenation of standalone BPC systems with a baseline system without the use of phonetic information. The standalone BPC systems use 205 UBM mixtures and 80 i-vector dimensions, and combining the 5 BPCs gives 1,025 UBM mixtures and 400 i-vector dimensions. The standalone BPC system parameters are chosen such that when the 5 BPCs are combined, the total number of parameters roughly match the number of parameters for a baseline, non-BPC system, which has 1,024 UBM mixtures and 400 i-vector dimensions. Experiments are also performed to verify the effectiveness of BPCs using the GMM-UBM system, as demonstrated in prior work. Score-level-averaging is used to combine BPCs for the GMM-UBM system.

Note that these experiments were performed using only the clean telephone conversation sides in our SRE08 and SRE10 dataset. They consist of 10,161 true speaker and 4,966,080 impostor trials for males, and 13,993 true speaker and 10,373,383 impostor trials for females. Table 1 presents the i-vector and GMM-UBM results.

I-vector Results for Males		
BPC Units	UBM Mixtures	EER (%)
Vow	205	3.1
Nasals	205	6.7
Glides/Liquids	205	6.9
Fricatives	205	7.2
Stops	205	8.4
All 5 BPCs	1,025	1.3
Non-BPC Baseline	1,024	**0.6**
GMM-UBM Results for Males		
All 5 BPCs	1,025	**6.7**
Non-BPC Baseline	1,024	8.1
I-vector Results for Females		
BPC Units	UBM Mixtures	EER (%)
Vow	205	4.0
Nasals	205	7.0
Glides/Liquids	205	8.9
Fricatives	205	8.3
Stops	205	8.8
All 5 BPCs	1,025	2.0
No BPCs Baseline	1,024	**1.5**
GMM-UBM Results for Females		
All 5 BPCs	1,025	**9.0**
Non-BPC Baseline	1,024	9.5

Table 1. Results for BPC units using the i-vector system, with i-vector-level combinations, and 205 UBM mixtures and 80 dim i-vectors per BPC. One UBM is trained per BPC.

Results show that for the GMM-UBM system, combining the 5 BPCs gives performance improvements over the baseline for both males (6.7% vs. 8.1% EER) and females (9.0% vs. 9.5% EER), which confirms trends observed from prior experiments. However, the i-vector-level combinations of BPCs do not perform nearly as well as the non-BPC baseline system, with EERs of 0.6% for males and 1.5% for females. Among the standalone BPCs, the Vowel BPC performs the best using the i-vector system (3.1% EER and 4.0% EER for males and females), while nasals perform the second best (6.7% EER and 7.0% EER for males and females).

Results suggest that in contrast to the GMM-UBM system, the i-vector system is better able to handle the lexical variability contained in clean telephone conversational speech. In the above experimental framework, the focusing of modeling power on BPCs is not needed, and even hurts performance, as it limits the total amount of data used in each standalone i-vector system. Another issue is that of data sparsity. While BPCs are superior to smaller-sized units in terms of having more data per conversation side, the fact BPCs such as Glides/Liquids perform worst for females (8.9% EER) could still be due to data sparsity, since it comprises only 9% of our speech data (as discussed in Section 3). Our past experiments have shown that the over-training of i-vector system parameters is an issue for BPCs in the presence of insufficient i-vector system development data.

The next experiments consist of using BPC-based i-vector systems on the noisy extended dataset. This dataset provides two advantages for BPC-based experiments: First, the presence of noise increases intersession variability, while BPCs help reduce lexical variability. It could be that reducing lexical variability is more effective in the context of increased intersession variability. Second, the increased data assists in addressing the data sparsity issue for BPCs. In an additional attempt to address data sparsity, a 2-BPC system is implemented, with only the Vowel and Consonant BPCs, both comprising roughly 50% of speech data. The Consonant BPC combines all non-Vowel BPCs.

Note also that the system parameters are increased to 2,048 UBM mixtures and 600 i-vector dimensions for the new non-BPC baseline systems, which we refer to as *Baseline_1*. The parameters for the standalone BPCs in the 5-BPC and 2-BPC systems are chosen such that the number of parameters in the BPC-combined systems match the total number of parameters for the Baseline_1 system. Results obtained using the old baseline system (i.e. *Baseline_0*) using 1,024 UBM mixtures and 400 i-vector dimensions are also shown. Both the i-vector stacking (*i-vect*) and Logistic Regression (*LR*) combination approaches are used. Tables 2 shows male and female results using the noisy extended dataset.

Results suggest that in the presence of a large noisy dataset, phonetic-based systems perform much closer to the non-phonetic baseline systems. For males, the 5-BPC and 2-BPC systems give 1.8% and 1.7% EERs, compared to 1.7%

I-vector Results for Males				
System(s)	Comb. Approach	UBM Mix	i-vect dims	EER (%)
All 5 BPCs	–	2,050	600	1.8
All 2 BPCs	–	2,048	600	1.7
Baseline_1	–	2,048	600	1.6
Baseline_0	–	1,024	400	1.7
Cons+Vow	LR	1,024x2	300x2	1.8
Baseline_0+ Cons+Vow	LR	1,024x3 +	400+ 300x2	**1.4**
I-vector Results for Females				
System(s)	Comb. Approach	UBM Mix	i-vect dims	EER (%)
All 5 BPCs	–	2,050	600	2.8
All 2 BPCs	–	2,048	600	2.5
Baseline_1	–	2,048	600	**2.4**
Baseline_0	–	1,024	400	2.7
Cons+Vow	LR	1,024x2	300x2	2.6
Baseline_0+ Cons+Vow	LR	1,024x3	400 + 300x2	**2.4**

Table 2. Male noisy data results for BPC units using the i-vector system, with system combinations using i-vector concatenation and logistic regression via the Bosaris Toolkit. The EER is computed by averaging EERs across 12 and 18 splits of scores for male and female speakers.

and 1.6% EERs for the Baseline_1 and Baseline_0 systems. For females, the 5-BPC and 2-BPC systems give 2.8% and 2.5% EERs, compared to 2.4% and 2.7% EERs for the Baseline_1 and Baseline_0 systems. The 2-BPC system is also at least as good as the 5-BPC system. While the 5-BPC, 2-BPC, and the Baseline_1 systems all use similar numbers of parameters, the advantage of BPC-based systems is that parallelization can be used in their implementation. For instance, instead of training a single 2,048-mixture UBM or a rank-600 T-matrix, two 1,024-mixture UBMs and rank-300 T-matrices can be trained in parallel

The combination results suggest that combining BPC-based systems with non-BPC-based systems is helpful. Logistic regression combination of the Baseline_0, and standalone Vowel and Consonant BPC systems gives a minimum EER of 1.4% for males. This represents a 13% relative improvement over the best standalone male result (1.6% EER for the Baseline_1 system). Note that the number of parameters in the combination has been increased over the Baseline_1 system. However, a comparison of results for the Baseline_1 and Baseline_0 systems for males, suggests that having increased parameters does not necessarily imply significant EER improvements (1.7% vs. 1.6% EER where Baseline_1 has nearly twice the number of parameters as Baseline_0).

6. CONCLUSION AND CAVEATS

We have demonstrated that BPC-based systems are able to outperform the non-BPC baseline system using the GMM-UBM approach and telephone conversational speech without added noise. For males, the 5-BPC system using the GMM-UBM approach gives a 6.7% EER, compared to an 8.1% EER for the non-BPC baseline. For females, the 5-BPC system gives a 9.0% EER, compared to 9.5% for the non-BPC baseline. While the BPC-based systems do not outperform the baseline systems using the i-vector approach on the clean telephone conversational speech data, improvements using BPCs have been observed using the expanded dataset and two types of environmental noise – car and crowd – at 10dB and 20dB SNRs. The Vowel and Consonant BPCs, when combined with the a baseline system using Logistic regression, gives a 13% relative improvement over the standalone baseline system.

Nevertheless, there are potential caveats for the presented effort that ought to be addressed in any future work. The conditions for which the BPC-based i-vector systems improved over the non-BPC baseline are narrow and specific, and required manufacturing of the dataset, the noise conditions, and BPCs used. The noise conditions – conversational speech specifically in car and crowd noise at 10dB and 20dB SNRs – may not represent the general noise characteristics encountered in the real-world. Furthermore, it's possible that the observed improvements resulted simply from the fact that system combinations can lead to improvements in general, rather than the discriminative characteristics of the BPCs themselves. Lastly, using a combination of BPCs results in a system with significantly greater computational complexity, than using a single non-BPC system. The BPCs were also derived from an automatic speech recognizer, which itself is not perfect. Hence, systems that claim to use a specific BPC may also be incorporating data outside of the BPC, and leaving out certain regions of the BPC. An alternative solution is to use manually-transcribed speech, but that requires substantial human effort for a large speech corpora. The 13% improvement over the baseline i-vector system attributed to the BPCs, hence, needs to be evaluated given the above considerations and potential caveats.

7. ACKNOWLEDGEMENTS

This work was funded by AFRL award FA8750-12-1-0016.

8. REFERENCES

[1] M. Omar and J. Pelecanos, "Training Universal Background Models for Speaker Recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2010, pp. 52–57.

[2] D. Gutman and Y. Bistritz, "Speaker Verification Using Phoneme-Adapted Gaussian Mixture Models," in *Proceedings of EUSIPCO*, 2002.

[3] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, September 2005, pp. 2429–2432.

[4] H. Lei and E. Lopez-Gonzalo, "Mel, Linear, and Anti-Mel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition," in *Proceedings of Interspeech*, September 2009, pp. 2323–2326.

[5] J. Scheffer, R. Vogt, S. Kajarekar, and J. Pelecanos, "Combination Strategies for a Factor Analysis Phone-Conditioned Speaker Verification System," in *Proceedings of ICASSP*, 2009, pp. 4053–4056.

[6] A. Larcher, P.M. Bousquet, K.A. Lee, and D. Matrouf, "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *Proceedings of ICASSP*, March 2012.

[7] "Freesound - http://www.freesound.org," .

[8] "Fant - filtering and noise adding tool," in *http://dnt.kr.hsnr.de/download.html*.

[9] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, and R.R. Gadde, "Speaker Recognition Using Prosodic and Lexical Features," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003)*, November 2003, pp. 19–24.

[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Speaker Odyssey: The Speaker Recognition Workshop*, June 2001.

[11] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of INTERSPEECH*, 2009, pp. 1559–1562.

[12] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," in *IEEE Trans. Speech and Audio Proc.*, 2005, vol. 13(3), pp. 345–354.

[13] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," in *IEEE Transactiions on Audio, Speech, and Language Processing*, 2008, vol. 16(5).

[14] L. Burget, P. Oldřich, C. Sandro, Oldřej G., Pavel M., and N. Brümmer, "Discriminantly trained probabilistic linear discriminant analysis for speaker verification," in *Proceedings of ICASSP*, Brno, Czech Republic, 2011.

[15] S. Ioffe, "Probabilistic Linear Discriminant Analysis," in *Proceedings of ECCV*, 2006, pp. 531–542.

[16] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 249–252.

[17] J.F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Proceedings of ICASSP*, 2005, vol. 1, pp. 737–740.

[18] "Joint factor analysis matlab demo," http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo/.

[19] "Hmm toolkit (htk)," http://htk.eng.cam.ac.uk/.

[20] N. Brümmer, "Bosaris toolkit," in *https://sites.google.com/site/bosaristoolkit/*.

[21] D.A. Reynolds, T.F. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," in *Digital Signal Processing*, 2000, vol. 10(3), pp. 19–41.