OPTIMIZING SPEAKER-SPECIFIC FILTER BANKS FOR SPEAKER VERIFICATION

Hector N. B. Pinheiro[1] *Fernando M. P. Neto*[1] *Adriano L. I. Oliveira*[1]
Tsang Ing Ren[1] *George D. C. Cavalcanti*[1] *André G. Adami*[2]

[1]Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife, Brazil

[2] Centro de Ciências Exatas e da Tecnologia (CCET)
Universidade de Caxias do Sul (UCS)
Caxias do Sul, Brazil

ABSTRACT

In this work, we investigate speaker-specific filter banks for text-independent speaker verification. The proposed method performs an heuristic search for the best filter-bank configuration using the Artificial Bee Colony (ABC) algorithm and a proper fitness function for the standard i-vectors/PLDA-based speaker verification system. Furthermore, filter-bank decorrelated amplitudes are used instead of the cepstral coefficients produced by Discrete Cosine Transform (DCT). In the experiments, the proposed method is compared to standard Mel and linear scales in both cases where the decorrelation is performed using DCT and high-pass filtering. The comparison is performed on the MIT Mobile Device Speaker Verification Corpus in a gender-dependent trial scheme. The proposed method outperformed the baseline systems in almost all the test sets for both genders. Performance gains of 4.6% and 26.0% are achieved for male and female speakers, respectively.

Index Terms— Speaker verification, filter bank optimization, artificial bee colony algorithm.

1. INTRODUCTION

In recent years, the Speaker Recognition (SR) community has predominantly used the i-vector representation [1] of speech utterances for text-independent speaker verification. This technique maps any utterance to a fixed low dimensional representation preserving useful information about the speaker and it is based on the Factor Analysis Decomposition of the correspondent Gaussian Mixture Model (GMM) supervector. Although i-vectors can be effectively compared using the cosine distance metric, speaker and session variability compensation is also performed by post-processing techniques such as Gaussian and heavy-tailed Probabilistic Linear Discriminant Analysis (PLDA) [2, 3].

Despite the existence of several techniques proposed to represent, model or compensate speech signals, Mel-Frequency Cepstral Coefficients (MFCCs) have always been the most dominant feature extractor used in speaker recognition. They were first proposed for speech recognition and is based on the codification of the spectral information of the signal on the Mel scale in order to mimic human auditory perception. The codification of a short-term speech segment is performed by using a Mel-scale filter bank and applying the Discrete Cosine Transform (DCT) to the logarithm of the filter bank energies to compute a set of uncorrelated features. Since the spectral resolution of the Mel scale becomes lower as frequency increases, high frequency information is suppressed on the MFCCs extraction. However, the high frequency region of speech reflects important physiological characteristics of the speaker associated to the structure of the vocal tract, such as its length [4]. For this reason, some studies show that the use of linearly spaced filters on the cepstral coefficients extraction, Linear Frequency Cepstral Coefficients (LFCC) outperforms conventional MFCCs in some cases (e.g. on nasal and non-nasal consonants regions of speech [5] or in whispered speech [6]). Recently, MFCCs and LFCCs performances were compared in the NIST Speaker Recognition Evaluation (SRE) 2010 extended-core task [7] using the i-vectors representation with PLDA post-processing. LFCCs outperformed MFCCs in most conditions, mainly due to its better performance in the female trials[1].

Based on the uniqueness of the characteristics of the structure of the vocal tract of individuals and their reflection on the speech spectra, it is expected that distinct frequency warping functions may be employed to extract more discriminative features for each person. Following this assumption, several studies were conducted on finding an optimal frequency warping function not only for speaker recognition but also for speech recognition [8, 9], speaker diarization [10], and, more recently, signal classification [11]. In [12], optimal frequency warping functions and classifier parameters were estimated by a Generalized Probabilistic Decent (GPD) optimizing procedure based on Minimum Classification Error (MCE) criteria. Even though a common speaker-independent frequency

This research has been supported by the following Brazilian agencies: CNPq (446831/2014-0) and FACEPE (APQ-0192-1.03/14).

[1]The shorter vocal tract length of females results in higher formants frequencies in their speech signals.

warping function was used, better results than those obtained with MFCCs and LFCCs were achieved when speaker identification techniques based on GMMs and vector quantization (VQ) were used. More recently, Charbuillet *et. al* proposed a method that optimizes the fusion of two GMM-based speaker verification systems with different cepstral features defined by complementary linear filter banks [13]. The optimal configurations of the filter banks were estimated using an evolutionary algorithm and defining an optimization criterion based on the complementary of the log likelihoods produced by the GMM systems. Once again, the optimization process did not take into account the existence of particular optimal frequency warping functions, and fixed filter banks configurations were employed for different speakers.

In this work, we investigate the optimization of speaker-specific filter banks for text-independent speaker verification. Here, a filter bank is parameterized by the centers of triangular filters, whose bandwidths are defined by the centers of their neighbors. Similarly to [13], an heuristic search algorithm inspired in nature is used in the optimization process. The Artificial Bee Colony (ABC) algorithm [14] is individually employed to find the correspondent optimal filter banks. ABC algorithm is a relatively new technique proposed for numerical optimization with proven effectiveness when compared to other heuristic search algorithms [15]. Differently from previous works, the proposed optimization method is designed to minimize the verification errors of the standard i-vectors/PLDA-based system using a proper fitness function. This method is described in Section 2. Furthermore, instead of extracting uncorrelated features using DCT, our proposed method relies on the decorrelation of the log filter bank amplitudes by high-pass filtering, which produces features with specific sub-bands information. Decorrelated filter bank amplitudes/energies were successfully used in the past for robust speech [16, 17] and speaker [18] recognition, and it has been shown that they achieve equivalent or better performance than MFCCs [17]. Intuitively, the impact of the optimization of filter banks may be greater when features with specific sub-band information are used. In the experiments (Section 3), we compared the proposed method to conventional Mel and linear scaled filter banks using the MIT Mobile Device Speaker Verification Corpus (MIT-MDSVC) [19], which presents trials with high session variability from background noise. Gender-dependent comparative results are also presented for both scenarios when decorrelation is performed by high-pass filtering and DCT.

2. PROPOSED FILTER BANK OPTIMIZATION

The Artificial Bee Colony (ABC) algorithm is an optimizer proposed for multivariable continuous functions inspired by the intelligent foraging behaviour of a bee colony. In the ABC algorithm, each solution of the problem is represented by a food source, which is defined by a d-dimensional real-valued vector, and the fitness of the solution corresponds to the amount of nectar of the associated food source. There are three groups of foraging artificial bees, which are referenced as employed bees, onlookers and scouts. Employed bees are responsible for exploiting the current food sources and giving information about their quality (amount of nectar). Onlooker bees wait in the hive and decide on a food source to exploit based on the quality information shared by the employed bees. The random search of new food sources is performed by the scouts. The algorithm starts with a population of randomly generated solutions and three steps corresponding to the behaviour of the bees are repeated until a termination criterion is met. The parameters of the basic ABC algorithm are: the number of the food sources N_f, which is equal to the number of employed and onlooker bees (each one associated to one food source); the number of trials after which a food source is assumed to be abandoned, which is referred to as $limit$; and the termination criterion. Let $\mathbf{s}_i = \{s_{i1}, s_{i2}, ..., s_{id}\}$ represent the ith food source of the population and t_i represents its exploitation trials counter. The steps iteratively executed by the algorithm are defined as follows:

- Employed bees step: each employed bee generates a new solution ($\hat{\mathbf{s}}_i$) in the neighborhood of its current position:

$$\hat{s}_{ij} = s_{ij} + \phi_{ij}(s_{ij} - s_{kj}), \tag{1}$$

where j and k are randomly chosen dimension and food source indexes, respectively, with $k \neq i$. ϕ_{ij} is a uniformly distributed real random number in the range $[-1, 1]$. If the altered dimension exceeds the predetermined boundaries of possible solutions, it is set to an acceptable value, as the upper/lower boundary value. Here, as the solutions represent filter bank configurations (centers of the filters) the dimensions are sorted. Furthermore, the dimensions are limited from zero to half of the sampling frequency. A greedy selection is then employed among them. If the fitness of $\hat{\mathbf{s}}_i$ is equal or better than that of $\mathbf{s}_i$, $\hat{\mathbf{s}}_i$ replaces $\mathbf{s}_i$ in the population and t_i is reset to zero. Otherwise $\mathbf{s}_i$ is retained and t_i is incremented.

- Onlooker bees step: Each onlooker bee evaluates the nectar information from all the employed bees and selects a food source to exploit. The selection consists of a roulette wheel scheme in which the size of each slice (probability of selection, p_i) is proportional to the fitness value of the correspondent solution: $p_i = fit_i / \sum_i fit_i$.

In each exploitation, positions and trials counters of the food sources are defined as the same as the previous step.

- Scout bees step: each scout bee selects a food source to be abandoned based on the trials counters (t_i) which were updated during search. Solutions that reached the $limit$ parameter are replaced by new randomly generated food sources with t_i reset to zero. In the basic ABC algorithm, at most one scout bee performs the replacing at each cycle.

In each iteration of the algorithm, the best solution found so far is memorized as the final solution. Termination criteria include: reaching the maximum number of iterations; reach-

ing the maximum fitness value; or reaching the maximum number of iterations without improvements (convergence criterion). Our goal here is to find the most discriminative configuration of the filter banks for a certain enrolled speaker, S. From the assumption that the best fitness function is the one that most closely resembles the final system, we designed a function that estimates the accuracy of the i-vectors/PLDA framework on the verification task.

Given the training speech samples from S and from impostor speakers different from S, the function computes a fitness value for a given solution representing the frequency centers of the triangular filters presented in the bank. Speech sentences are divided into frames of 20 ms at a frame period of 10 ms. A feature vector consisting of sub-band components derived from the decorrelated filter bank amplitudes [20, 16, 17] is extracted from each frame. As in [18], a filter bank with 21 triangular filters is used to obtain the log filter-bank amplitudes, $(a_1, a_2, ..., a_{21})$. They are decorrelated by applying a high-pass filter $H(z) = 1 - z^{-1}$ over a_i, resulting in 20 decorrelated amplitudes. The resulting feature vector is composed by these amplitudes with the addition of their first-order delta components. Furthermore sentence-level mean removal are also applied for channel compensation. From all the vectors extracted from the training utterances, gender-dependent Universal Background Models (UBMs) are estimated via Expectation-Maximization (EM) algorithm and combined to compose a single UBM, which is used in the i-vectors extraction from the speeches. For the computation of the fitness value, the i-vectors are equally divided into two distinct sets presenting different sentences from S and from the impostors. While the first set is used to train an i-vectors/PLDA system for speaker S, the second is used to produce the true/false positive scores. In order to avoid overfitting, the division of the speech samples is performed several times and it is maintained fixed during all the search process. A single set of true/positive scores is produced by joining all the scores produced by different data divisions. A single Detection Error Tradeoff (DET) curve is produced and the final fitness value is defined by $1 - EER$, where EER refers to the achieved Equal Error Rate.

3. EXPERIMENTS

The goal of our experiments is to compare the use of the Mel, linear (Lin) and the proposed ABC-optimized scales in the feature extraction phase for the verification task using the i-vectors/PLDA system. The use of the scales is evaluated in both cases when the uncorrelated coefficients were produced by the DCT (as the conventional MFCCs and LPCCs extractions) and by the use of the high-pass filtering (as presented in Section 2). The evaluations were conducted using speeches from the MIT Mobile Device Speaker Verification Corpus (MIT-MDSVC) [19]. The speech data were collected using a handheld device with a sampling frequency of 16 KHz. The data collection consists of two unique sets of enrolled users and dedicated impostors. The set of enrolled users was collected during two different sessions and the impostor set was obtained in a single session. Both sets provide environmental noise variability since each session occurred in three different locations: a quiet office, a mildly noisy lobby and a busy street intersection. Each speaker recorded 18 utterances per location giving 54 speech sentences per session. The enrolled users set presents 48 individuals (22 female and 26 male) while the impostors set is composed by 40 individuals (17 female and 23 male). In this work, the training data consists of the utterances from the first session recorded in a quiet office. Three test sets were considered, each corresponding to a different location and including speeches from the second session and the impostors set. The systems were tested following a gender-dependent scheme where no cross-gender trials were performed (since they are less challenging as discussed in Section 1). Consequently, 18 true trials were performed for each enrolled speaker while 414 (23×18) and 306 (17×18) false trials were performed for male and female speakers, respectively.

3.1. Experimental setup

In the evaluation of all systems, the training speech sentences were used to estimate the UBM produced by the combination of two gender-dependent UBMs with half number of components. The training samples were then used to produce the total variability matrix in the i-vectors extraction phase. From the extracted 100-dimensional i-vectors a two-class PLDA model was estimated for each speaker by labelling all the impostors' training i-vectors to the same class. In our experiments, the number of components of the UBM were varied in powers of 2 from 8 to 1024 for all the systems and the best results were considered for comparison. All the systems were tested in both cases where the decorrelation was performed by DCT (as in conventional extraction of cepstral coefficients) and by a $H(z) = 1 - z^{-1}$ high-pass filter. When the DCT was used, the filter banks were composed by 29 filters[2] and the feature vectors comprised the cepstral coefficients and the additional first- and second-order delta coefficients. When high-pass filtering was used to produce sub-band decorrelated amplitudes, banks with 21 filters were used and only the first-order delta components were considered. This configuration is the same used in [18]. In both cases sequence-level mean removal was performed to the vectors. The proposed optimization method was executed for each enrolled speaker using UBMs with 32 components. In the fitness function evaluation a gender-independent trial scheme with a total of 100 data divisions was used. The ABC algorithm was executed using a population with 50 food sources and the limit parameter was set to 20. The searches were set to terminate when 20 consecutive iterations without gain of fitness were reached.

[2]Following the commonly used expression $n = \lfloor 3 \times \log Fs \rfloor$, where n and Fs refers to the number of filters and the sampling frequency, respectively.

Table 1. Equal Error Rates (EERs) and minimum Decision Cost Functions (minDCFs) achieved by the systems. They are defined by the frequency scale used in the filter banks: Mel, linear (Lin), and the proposed optimized (ABC) scales; and by decorrelation post-processing: DCT (MFCC and LFCC) and high-pass filtering (HPF). The best results are presented in bold. The performance gains were computed by comparing the proposed method to the systems with best performance on the male (Mel/HPF) and female (Lin/HPF) conditions, presented in italics.

System	EER (in %) | minDCF $\times 10^4$															
	Male								Female							
	Office		Lobby		Intersection		Avg.		Office		Lobby		Intersection		Avg.	
MFCC	4.30	107.84	7.91	205.89	11.54	215.63	7.92	176.45	5.05	168.08	7.63	205.54	14.39	264.71	9.02	212.77
LFCC	4.27	126.83	7.46	200.97	9.40	211.92	7.04	179.91	5.81	159.45	8.59	227.47	10.86	229.77	8.42	205.56
Mel/HPF	*4.06*	*101.46*	*7.71*	*207.73*	***7.69***	*206.19*	*6.49*	*171.79*	4.80	171.18	9.60	199.24	16.16	313.49	10.19	227.97
Lin/HPF	5.56	145.17	8.12	221.44	8.11	**188.44**	7.26	185.02	*5.30*	*123.29*	*6.82*	*183.45*	*12.12*	*244.03*	*8.08*	*183.59*
ABC/HPF	**2.78**	**95.67**	**6.62**	**161.34**	9.16	227.53	6.19	161.51	**3.54**	**93.18**	**4.04**	**138.61**	**10.35**	**206.40**	**5.98**	**146.06**
Gain (%)	31.5	5.7	14.1	22.3	-16.1	-9.4	4.6	6.0	33.2	24.4	40.8	24.4	14.6	15.4	26.0	20.4

3.2. Results

For the comparison of the systems, we computed the Equal Error Rates (EERs) and the minimum Decision Cost Function (minDCF) with cost of miss, cost of false alarm, and prior probability of a target trial set to 10, 1 and 0.01, respectively, as defined in [21]. Table 1 shows the performances achieved by the systems. We found that all the systems achieved the best performance at the case where UBMs with 256 components were used, except for the one using Mel-scaled features with decorrelation via DCT, which achieved the best results using 128 components. Independently on the decorrelation post-processing performed by the systems, better results were achieved by the use of the linear scale for the female speakers in comparison to the use of the Mel scale. This result was expected since the linear scale has a better resolution in the high-frequency region of spectrum. By comparing the baseline systems, one can see that the best average performances comes from the use of decorrelated filter-bank amplitudes for both male and female speakers. This is due to the gain of performance on the Intersection (males) and Lobby (females) testing conditions.

The proposed method mostly outperformed all the systems for both male and female speakers. In terms of EER, the average gain in performance compared to the best baseline system for males (Mel/HPF) was 4.6%. On the other hand a 26.0% of average performance gain was achieved when compared to the best female system (Lin/HPF). In terms of minDCF, the average gain was 6.0% for males and 20.4% for females. Since the proposed optimization has the capability of finding filter-bank configurations with better resolution for higher frequencies, greater gains are also expected for females. Since only the speech samples recorded in office were used in the training phase, the proposed optimization method searched for the filter-bank configuration that achieves the best verification rates using clean data. For this reason better results were achieved in the Office and Lobby testing environments and the system performed worse with the increasing of background noise. Indeed, the proposed method did not outperform the baseline systems in the Intersection testing set for male speakers. However, artificial noise can be incorporated to the samples used in the trials during the fitness computation of the solutions. This multi-conditional training approach were successfully used in the past in order to increase the robustness of speaker verification systems [22, 18] and can be easily incorporated in the proposed method. Furthermore, gender-dependent trials can also be used in order to increase the discriminant power of the filter-bank configurations.

4. CONCLUSIONS

In this work, we investigated the optimization of speaker-specific filter banks for text-independent speaker verification. The proposed method performs a heuristic search for the best filter-bank configuration defined by the centers of the triangular filters present on the bank. The proposed optimization relies on the use of the Artificial Bee Colony algorithm and a proper fitness function that estimates the accuracy of the standard i-vectors/PLDA-based speaker verification system. Furthermore, filter-bank decorrelated amplitudes are used as features instead of the conventional cepstral coefficients produced by DCT. The amplitudes resulted from the filters are decorrelated through a high-pass filter.

The proposed method was compared to the use of the standard Mel and linear scales in both cases where the decorrelation is performed using DCT and high-pass filtering. The comparison was performed on the MIT Mobile Device Speaker Verification Corpus in a gender-dependent trial scheme. The proposed method outperformed the baseline systems in almost all the test sets for both male and female speakers. In terms of EER, the average gain of performance compared to the best baseline system for males was 4.62%, while a 25.99% of average performance gain was achieved when compared to the best female baseline system. In terms of minDCF, the average gain was 5.98% for males and 20.44% for females. Further noise-robustness improvements may be achieved by following a multi-conditional approach in the proposed optimization procedure.

5. REFERENCES

[1] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, 2010, p. 14.

[3] Simon J. D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[4] Kenneth N. Stevens, *Acoustic phonetics*, vol. 30, MIT Press, 2000.

[5] Howard Lei and Eduardo López Gonzalo, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition.," in *INTERSPEECH*, 2009, pp. 2323–2326.

[6] Xing Fan and John H. L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4553–4556.

[7] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma, "Linear versus Mel frequency cepstral coefficients for speaker recognition," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2011, pp. 559–564.

[8] Mohamed Chetouani, Marcos Faundez-Zanuy, Bruno Gas, and Jean-Luc Zarader, "Non-linear speech feature extraction for phoneme classification and speaker recognition," in *Nonlinear Speech Modeling and Applications*, pp. 344–350. Springer, 2005.

[9] Leandro D. Vignolo, Hugo L. Rufiner, Diego H. Milone, and John C. Goddard, "Evolutionary cepstral coefficients," *Applied Soft Computing*, vol. 11, no. 4, pp. 3419–3428, 2011.

[10] Christophe Charbuillet, Bruno Gas, Mohamed Chetouani, and Jean-Luc Zarader, "Filter bank design for speaker diarization based on genetic algorithms," in *International Conference on Acoustics Speech and Signal Processing*. IEEE, 2006, vol. 1.

[11] Maxime Sangnier, Jérôme Gauthier, and Alain Rakotomamonjy, "Filter bank learning for signal classification," *Signal Processing*, vol. 113, pp. 124–137, 2015.

[12] Chiyomi Miyajima, Hideyuki Watanabe, Keiichi Tokuda, Tadashi Kitamura, and Shigeru Katagiri, "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction," *Speech Communication*, vol. 35, no. 3, pp. 203–218, 2001.

[13] Christophe Charbuillet, Bruno Gas, Mohamed Chetouani, and Jean-Luc Zarader, "Optimizing feature complementarity by evolution strategy: Application to automatic speaker verification," *Speech Communication*, vol. 51, no. 9, pp. 724–731, 2009.

[14] Dervis Karaboga and Bahriye Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, vol. 39, no. 3, pp. 459–471, 2007.

[15] Dervis Karaboga and Bahriye Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Applied soft computing*, vol. 8, no. 1, pp. 687–697, 2008.

[16] Kuldip K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition.," in *Eurospeech*, 1999, vol. 99, pp. 85–88.

[17] Climent Nadeu, Dušan Macho, and Javier Hernando, "Time and frequency filtering of filter-bank energies for robust hmm speech recognition," *Speech Communication*, vol. 34, no. 1, pp. 93–114, 2001.

[18] Ji Ming, Timothy J. Hazen, James R. Glass, and Douglas A. Reynolds, "Robust speaker recognition in noisy conditions," *Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[19] Ram H. Woo, Alex Park, and Timothy J. Hazen, "The MIT mobile device speaker verification corpus: data collection and preliminary experiments," in *Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–6.

[20] Climent Nadeu, Javier Hernando, and Monica Gorricho, "On the decorrelation of filter-bank energies in speech recognition.," in *Eurospeech*. Citeseer, 1995, vol. 95, pp. 1381–1384.

[21] "2010 NIST speaker recognition evaluation," `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`.

[22] Hector N. B. Pinheiro, Sergio R. F. Vieira, Tsang I. Ren, George D. C. Cavalcanti, and Paulo S. G. Mattos Neto, "Type-2 fuzzy GMM for text-independent speaker verification under unseen noise conditions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.