TYPE-2 FUZZY GMM FOR TEXT-INDEPENDENT SPEAKER VERIFICATION UNDER UNSEEN NOISE CONDITIONS

Héctor N. B. Pinheiro, Sérgio R. F. Vieira, Tsang Ing Ren, George D. C. Cavalcanti, Paulo S. G. de Mattos Neto

Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, PE, Brazil

{hnbp,srfv,tir,gdcc,psgmn}@cin.ufpe.br

ABSTRACT

This paper describes a novel GMM-UBM based system that deals with the session noise variability problem. The system uses the Type-2 Fuzzy GMM framework by considering the speaker GMM parameters to be uncertain in an interval. The parameters intervals are estimated using a multicondition model training on noisy speeches that are synthesized from the speaker's utterances. Experiments were conducted using the MIT Device Speaker Verification Corpus with utterances having the lowest noise level as training data. The result shows an improvement in the EER of 24.11% for the proposed method compared to the GMM-UBM when evaluated over the noisiest utterances. This shows that the method reduces the effects of the session variability.

***Index Terms*—** Text-independent speaker verification, session variability, type-2 fuzzy GMM, multicondition model training.

1. INTRODUCTION

The Gaussian Mixture Model-Universal Background Model (GMM-UBM) [1] is a framework for text-independent speaker verification applications [2–4]. The GMM capability of modeling the different phonetic variations from the speaker's utterances associated with its insensitivity to temporal aspects demonstrates the effectiveness of the method [4]. The i-vector approach, which is considered the state-of-the-art, is also influenced by the GMM-UBM model [5]. This approach is based on the extraction of statistics of a UBM with a big number of Gaussian components.

In real applications, GMM-UBM has to model training and testing utterances recorded in different sessions with different background noise. This mismatched environmental condition is known as session variability [6] and it leads to intra-speaker variability. We argue that the GMM parameters are corrupted since noise is present in a speech signal. Here, we propose a text-independent speaker verification system that handles GMMs with uncertain parameters.

This work was partially supported by Brazilian agencies CNPq, CAPES, FACEPE

Zeng *et al.* [7] introduced the Type-2 Fuzzy GMM (T2F-GMM) framework to describe the GMMs uncertain parameters and provides intervals for the likelihood of an observation. We applied the T2F-GMM model in the speaker verification problem and obtained better results than the traditional GMM-UBM [8–10]. In [8] and [9], we assumed a fixed uncertainty to train the models and in [10], the uncertainty of the GMM parameters was estimated using different noisy speeches. Therefore, it was necessary to collect training utterances from different environments. These models, however, are not considered robust to the session noise variability. Here, we applied the multicondition model training approach [11] that allows the estimation of the uncertainty with no prior knowledge on the conditions of the environment.

In the remainder of this paper, we describe the T2F-GMM in Section 2. In Section 3, the proposed method is introduced. Section 4 presents the experiments and the comparative results. Finally, conclusions are presented in Section 5.

2. TYPE-2 FUZZY GMM FRAMEWORK

The GMM likelihood of a D-dimensional observation $\boldsymbol{x}$, considering M mixtures, is defined as:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} \omega_i N(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{1}$$

in which $N(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ is the multivariate Gaussian. The parameter ω_i is defined as the mixture weight, having the following property $\sum_{i=1}^{M} \omega_i = 1$. The parameters $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$ are the D-dimensional mean vector and $D \times D$-dimensional covariance matrix, respectively. The model $\lambda = \{\omega_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}$, for $i = 1, 2, \ldots, M$ is estimated from the training data by the Expectation-Maximization (EM) algorithm [12].

The likelihood $p(\boldsymbol{x}|\lambda)$ may be corrupted as the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ parameters may have uncertain values due to noise or insufficient data. Zeng *et al.* [7] proposed the Type-2 Fuzzy GMM (T2F-GMM) framework to handle GMMs with uncertain parameters by using the theory of Type-2 Fuzzy Sets [13]. The framework assumes that the values μ and σ for each component of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, are uniformly distributed in the intervals $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ whose boundaries are defined as:

$$\underline{\mu} = \mu - k_m\sigma, \quad \overline{\mu} = \mu + k_m\sigma; \tag{2}$$

$$\underline{\sigma} = k_v\sigma, \quad \overline{\sigma} = \frac{\sigma}{k_v}. \tag{3}$$

The uncertainty parameters are $k_m \in [0, 3]$ and $k_v \in [0.3, 1]$, since the one-dimensional Gaussian has 99.7% of its probability concentrated in the range $[\mu - 3\sigma, \mu + 3\sigma]$.

The uncertain normal density function is defined considering the uncertain mean vector $\tilde{\boldsymbol{\mu}}$ and the uncertain covariance matrix $\tilde{\boldsymbol{\Sigma}}$:

$$N(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \prod_{i=1}^{D} e^{\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]}, \mu_i \in [\underline{\mu_i}, \overline{\mu_i}]; \tag{4}$$

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \prod_{i=1}^{D} e^{\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]}, \sigma_i \in [\underline{\sigma_i}, \overline{\sigma_i}]. \tag{5}$$

Note that the diagonal covariance matrix $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_D^2)$ is used in Eqs. (4) and (5).

Each uncertain exponential factor in Eqs. (4) and (5) is the primary Member Function (MF) of the Gaussian with uncertain mean or standard deviation, and is denoted as:

$$f(x; \mu, \sigma) = e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)\right]}. \tag{6}$$

Type-2 fuzziness is defined as the MF of a primary MF that is called second MF. The upper second MF with uncertain mean is defined as:

$$\overline{h}(x) = \begin{cases} f(x; \underline{\mu}, \sigma), & x < \underline{\mu}, \\ 1, & \underline{\mu} \leq x \leq \overline{\mu}, \\ f(x; \overline{\mu}, \sigma), & x > \overline{\mu}. \end{cases} \tag{7}$$

and the lower second MF is:

$$\underline{h}(x) = \begin{cases} f(x; \overline{\mu}, \sigma), & x \leq \frac{\underline{\mu}+\overline{\mu}}{2}, \\ f(x; \underline{\mu}, \sigma), & x > \frac{\underline{\mu}+\overline{\mu}}{2}. \end{cases} \tag{8}$$

The upper second MF with uncertain standard deviation is:

$$\overline{h}(x) = f(x; \mu, \overline{\sigma}) \tag{9}$$

and the lower second MF is:

$$\underline{h}(x) = f(x; \mu, \underline{\sigma}). \tag{10}$$

3. PROPOSED METHOD

In the proposed method, the noise compensation is performed through the T2F-GMM framework. The basic idea is to estimate the intervals of the hypothesized speaker model's parameters (Equations 2 and 3) and perform the verification task using the log-likelihoods intervals provided by the framework. The estimation process was designed in order to model the parameters' distortion resulted from possible unknown background noise. A multiconditional training approach is used for this purpose.

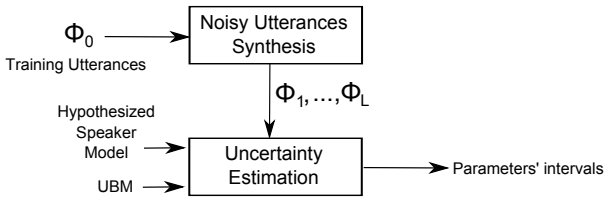

Fig. 1. Architecture of the proposed method. The system estimates the uncertainty factors of the parameters (Km and Kv) as an input to the T2F-GMM framework.

The proposed architecture is shown in Figure 1. From the original speeches of the training set from speaker, Φ_0, different noisy utterances $\Phi_1, \ldots, \Phi_L$ are synthesized by introducing degradation of different characteristics. The multicondition noisy speeches are then used to estimate the distortion of the speaker model parameters. The uncertainty produced by the distortions is modeled through the uncertainty factors.

Multiconditional training has been used in speech [14, 15] and speaker [11] recognition in order to increase robustness to noise conditions that are different from the original training. The main idea of the proposed method is to use the multicondition noisy speeches to estimate the uncertainty of the speaker model parameters. Both UBM and the speaker models are estimated in the same way as it was proposed by the standard GMM-UBM [1]. The UBM model is estimated using the EM algorithm in speeches from a large number of speakers. The speaker-specific model is estimated by adapting the UBM using the original training set Φ_0 through a *maximum a posteriori* (MAP) estimation.

The multicondition noisy speeches were produced by adding a White Gaussian Noise (WGN) at various Signal-to-Noise Ratios (SNRs), similarly to the procedure used by Ming *et al.* [11]. By decreasing the SNR at each step in the synthesis, the distortion presented in Φ_l is greater than the distortion presented in Φ_{l-1}, for $1 \leq l \leq L$. The goal of the proposed uncertainty estimation is to obtain an interval for each parameter of the speaker model that is able to cover the maximum range of distortion without losing its speaker-specificity. The idea then is to track the parameter distortion caused by the increase of the noise in the training speeches and compare the distorted parameters to the parameters presented in the mixtures of the UBM.

In order to observe the parameters' distortion caused by the increase of the noise, the GMMs were estimated in cascade using the synthesized speeches. The speaker model is estimatedby the MAP adaptation of the UBM means using Φ_0, similarly to the standard GMM-UBM method. In any

further stage of the cascade, the GMM_l is estimated by the MAP adaptation of the parameter of GMM_{l-1} using Φ_l, for $1 \leq l \leq L$. Since GMM_{l-1} is estimated using Φ_{l-1} and Φ_l is noisier than Φ_{l-1}, it is possible to observe the corruption of the parameter caused by the increasing in noise. Figure 2 illustrates this method for the corruption parameter P.

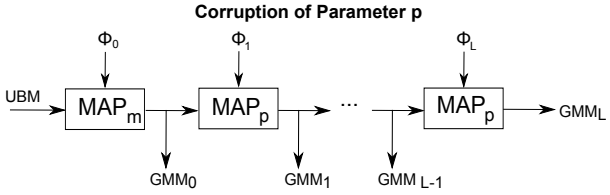

Fig. 2. Schematic of the multiconditional training. A set of GMMs with corrupted parameters $\{\text{GMM}_1, \ldots, \text{GMM}_9\}$ are estimated in cascade by the MAP adaptation of the parameter of interest ($p \in \{m, v\}$) using the noisy speeches sets $\{\Phi_1, \ldots, \Phi_L\}$.

The goal of the uncertainty estimation is to create the intervals for each parameter of the speaker model, GMM_0. By defining the parameters' intervals, Equations 7 to 10 is used to compute the likelihoods intervals for the posterior verification task. Since the covariance matrices are diagonal, the means and standard deviations of each component of the model can be analyzed independently. Consider μ_{ij}^l and σ_{ij}^l the mean and the standard deviation of component i and dimension j from GMM_l, for $1 \leq i \leq M$, $1 \leq j \leq D$ and $1 \leq l \leq L$, where M is the number of components of the models and D is the number of features extracted from the speeches. Similarly, consider $\tilde{\mu}_{ij}$ and $\tilde{\sigma}_{ij}$ the correspondent values of the UBM.

The parameters' intervals are defined by the upper and lower boundaries. The boundaries is set to maximize the parameters distortions without losing the speaker-specificity of the model. Therefore, the parameters of the UBM are used to restrict the boundaries.

To define the interval of the mean μ_{ij} from GMM_0, consider the sets:

$$\Psi = \{\psi | \psi = \tilde{\mu}_{ij} - \tilde{\sigma}_{ij} \textit{ and } \psi > \mu_{ij}\} \tag{11}$$

and

$$\Gamma = \{\gamma | \gamma = \tilde{\mu}_{ij} + \tilde{\sigma}_{ij} \textit{ and } \gamma < \mu_{ij}\}. \tag{12}$$

The boundaries intervals of μ_{ij} are limited by the interval (γ^*, ψ^*), where

$$\psi^* = min(\Psi) \tag{13}$$

and

$$\gamma^* = max(\Gamma). \tag{14}$$

All the components of the UBM are used to define the interval that restricts the boundaries of the parameters.

In order to maximize the coverage of the corrupted parameters, the upper boundary of μ_{ij} is defined by the greatest value of μ_{ij}^l that is lower than ψ^*, for $0 \leq l \leq L$:

$$\overline{\mu}_{ij} = max(\{\mu_{ij}^l | \mu_{ij}^l < \psi^*\}). \tag{15}$$

Similarly, the lower boundary is defined by the lowest value of μ_{ij}^l that is greater than γ^*:

$$\underline{\mu}_{ij} = min(\{\mu_{ij}^l | \mu_{ij}^l > \gamma^*\}). \tag{16}$$

On the other hand, the boundaries of σ_{ij} must be defined so that $\mu_{ij} \pm \overline{\sigma}_{ij}$ and $\mu_{ij} \pm \underline{\sigma}_{ij}$ are limited by the interval (γ^*, ψ^*). For this reason, the upper and lower boundaries are defined by the maximum and minimum values of σ_{ij}^l, for $1 \leq l \leq L$, respectively.

3.1. Verification Task

Given a set of feature vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ extracted from a testing utterance, the log-likelihood correspondent to model λ is defined as $v(\boldsymbol{X}|\lambda) = \frac{1}{T}\sum_{i=1}^{T} \log[p(\boldsymbol{x}_i|\lambda)]$.

The system then computes the intervals of the log-likelihood of $\boldsymbol{X}$ for uncertain means and uncertain standard deviations using the intervals $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ that were defined in the training phase. The likelihood $p(\boldsymbol{x}_i|\lambda)$ is computed using Equations 4 and 5. The exponential factors of these density functions are replaced by the primary MFs in Equations 7-10. The log-likelihood interval are obtained with respect to the uncertain means $[L_\lambda^\mu, U_\lambda^\mu]$ and the uncertain standard deviations $[L_\lambda^\sigma, U_\lambda^\sigma]$, where the subscript λ indicates the considered model (speaker S or UBM).

The final score computed for $\boldsymbol{X}$ is defined by the combination of the ratios between the upper boundaries and the intervals, computed for the hypothesized speaker S:

$$\Lambda(\boldsymbol{X}) = \frac{U_S^\mu - U_{UBM}^\mu}{U_S^\mu - L_S^\mu} + \frac{U_S^\sigma - U_{UBM}^\sigma}{U_S^\sigma - L_S^\sigma}. \tag{17}$$

Finally the verification task is performed by thresholding the computed score.

4. EXPERIMENTS

The main objective of the experiments is to compare the performances of the proposed method and the GMM-UBM at mismatched noise conditions. The systems were trained with utterances of low noise and evaluated over three disjoint test sets, each relative to a distinct noise level. For each test set, the Equal Error Rate (EER) performance metric was computed.

In order to observe the improvement of the proposed method without other influence, neither pre-processing algorithms [3], nor feature normalization techniques [2, 3] were considered. The Mel Frequency Cepstral Coefficients

(MFCCs) were computed in the feature extraction module. The speech signals were framed in a 20 ms Hamming window with an overlap of 10 ms. A total of 19 coefficients were extracted. The Δ_1 and Δ_2 were also computed and appended to the MFCCs. Then, a 57-dimensional features vectors was obtained.

The UBM was obtained by training two gender-dependent models and pooling them together. Each speaker GMM was estimated by MAP adaptation and only the top C best scoring mixture components were used in the computation of the log-likelihood ratio [1]. The experiments showed that the best results are achieved when only the means are adapted and by setting $C = 10$. The proposed method and the GMM-UBM were trained with a variant number of mixtures $M = 32, 64, 128, 256$. Furthermore, for the proposed method a total of $L = 9$ multiconditional noisy speeches sets were synthesized by varying the SNR from 20 dB (Φ_1) to 4 dB (Φ_9), in a step of 2 dB.

The experiments were conducted using speeches from the MIT Device Speaker Verification Corpus (MIT-DSVC) [16]. The speech data, collected on a hand-held device, 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 examples by session. For the enrolled users set, 48 individuals (22 female and 26 male) participated while for the impostors set 40 individuals (17 female and 23 male) were recorded.

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 for each speaker using her/his speeches and the impostors set, totaling 738 trials (18 true trials plus 18×40 false trials) per speaker. The EERs were computed considering the trials performed for all speakers in the set.

Table 1. *The EERs (in %) of the system for different location and mixtures.*

Location	GMM-UBM	M	T2F-GMM-UBM	M
Office	7.18	64	5.77	64
Lobby	18.86	64	15.97	128
Street	24.5	64	18.63	256

The effects of the session noise variability are analyzed. Table 1 shows the best performances for each location and number of mixtures. The proposed method yielded considerable improvements for all three different location. Especially in the busy street intersection, where a 24.11% gain in performance was achieved. In both methods, the performances decreases as the environmental noise increases. Nevertheless, the proposed method presents a better performance in all three situation compared to the GMM-UBM.

Table 2. *The overall EERs (in %) of the systems for different mixtures.*

N_{mix}	GMM-UBM	T2F-GMM-UBM
32	18.28	16.86
64	16.86	15.34
128	18.44	13.73
256	23.31	13.73

We also analyzed the overall performance of the methods, in applications that does not take into consideration the difference in the environment. Table 2 shows the average EERs obtained for different number of mixtures. The proposed method yielded better results for all numbers of mixtures. Comparing the worst case of the proposed method (32 mixtures) against the best case of the GMM-UBM (64 mixtures), the results shows that the performance are equal. Comparing the best performances of the systems, the T2F-GMM-UBM approach presented a general improvement of 18.56%. These results shows clearly that the proposed method has better performance than the standard GMM-UBM framework when considering mismatched noise conditions.

5. CONCLUSION

We propose a new scheme based on the multicondition model training for estimating the uncertainty of the GMM parameters in text-independent speaker verification tasks with session noise variability. The proposed T2F-GMM framework and the standard GMM-UBM were evaluated on the MIT-DSVC dataset using low noise speeches for training. The results shows that the proposed method reduce the effects of session noise variability when tested with the high noise utterances. A relative EER reduction of 24.11% was obtained when compared to the standard GMM-UBM model.

6. REFERENCES

[1] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[2] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification,"

EURASIP journal on applied signal processing, vol. 2004, pp. 430–451, 2004.

[3] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] Roberto Togneri and Daniel Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

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

[6] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Speaker and session variability in GMM-based speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1448–1460, 2007.

[7] Jia Zeng, Lei Xie, and Zhi-Qiang Liu, "Type-2 fuzzy Gaussian Mixture Models," *Pattern Recognition*, vol. 41, no. 12, pp. 3636–3643, 2008.

[8] Tsang Ing Ren, Dimas Gabriel, Hector NB Pinheiro, and George DC Cavalcanti, "Speaker verification using Type-2 Fuzzy Gaussian Mixture Models," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2336–2340.

[9] Hector NB Pinheiro, Tsang Ing Ren, George DC Cavalcanti, Tsang Ing Jyh, and Jan Sijbers, "Type-2 fuzzy GMM-UBM for text-independent speaker verification," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4328–4331.

[10] Hector NB Pinheiro, Tsang Ing Ren, George DC Cavalcanti, Tsang Ing Jyh, and Jan Sijbers, "Type-2 fuzzy GMMs for robust text-independent speaker verification in noisy environments," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 4531–4536.

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

[12] Todd K Moon, "The Expectation-Maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[13] Jerry M Mendel, "Type-2 fuzzy sets and systems: an overview," *Computational Intelligence Magazine, IEEE*, vol. 2, no. 1, pp. 20–29, 2007.

[14] Richard P Lippman, Edward Martin, Douglas B Paul, et al., "Multi-style training for robust isolated-word speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. IEEE, 1987, vol. 12, pp. 705–708.

[15] Li Deng, Alex Acero, Mike Plumpe, and Xuedong Huang, "Large-vocabulary speech recognition under adverse acoustic environments.," in *INTERSPEECH*, 2000, pp. 806–809.

[16] Ram H Woo, Alex Park, and Timothy J Hazen, "The mit mobile device speaker verification corpus: Data collection and preliminary experiments," in *In Proc. of Odyssey, The Speaker & Language Recognition Workshop*, 2006.