Speaker Classification via Supervised Hierarchical Clustering Using ICA Mixture Model

Muhammad Azam[1(✉)] and Nizar Bouguila[2]

[1] Department of Electrical and Computer Engineering,
Concordia University, Montreal, QC, Canada
mu_azam@encs.concordia.ca

[2] Concordia Institute for Information Systems Engineering,
Concordia University, Montreal, QC, Canada
nizar.bouguila@concordia.ca

Abstract. In this paper, speaker classification using supervised hierarchical clustering is provided. Bounded generalized Gaussian mixture model with ICA is adapted for statistical learning in the clustering framework. In the presented framework ICA mixture model is learned through training data and the posterior probability is used to split the training data into clusters. The class label of the training data is further selected to mark each cluster into a specific class. The cluster-class information from the training process is taken as reference for the classification of test data into different speaker classes. This framework is employed for the gender and 10 speakers classification and TIMIT and TSP speech corpora are selected to validate and test the classification framework. This classification framework also validate the statistical learning of our recently proposed ICA mixture model. In order to examine the performance of the ICA mixture model, the classification results are compared with same framework using Gaussian mixture model. It is observed that: (i) presented clustering framework performs well for the speaker classification, (ii) ICA mixture model outperforms Gaussian mixture model in the statistical learning based on the classification accuracy for gender and multi-class scenarios.

Keywords: Bounded Generalized Gaussian Mixture Model (BGGMM) · Independent Component Analysis (ICA) · Speaker classification · Supervised hierarchical clustering · ICA mixture model

1 Introduction

Speaker classification is a fundamental component of speaker recognition systems which performs two alternative tasks: speaker identification and verification. The goal of speaker identification is to label an unknown speech file with a speaker identity. The task of speaker verification is to validate and confirm the claim of a speaker about its identity [1,2]. Speaker classification has been used in

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 193–202, 2016.
DOI: 10.1007/978-3-319-33618-3_20

human-machine dialog systems, forensics, medical and many other applications. One interesting application of speaker classification is in the speech recognition and keyword spotting as preprocessing to reach the speaker of interest which is further useful in many security applications. Mixture models have been widely adopted to address the speaker classification task [3]. Recently Mixture model have been employed to address the object recognition and classification tasks through clustering in [4,5]. A two level hierarchical clustering framework based on inverted Dirichlet mixture model is presented in [6] which is selected for object clustering and recognition. In this work, the same hierarchical clustering framework is adapted using bounded generalized Gaussian mixture model (BGGMM) with ICA and employed for speaker classification. In this paper, gender and 10 speakers classification is performed through the hierarchical clustering framework using ICA mixture model. Bounded generalized Gaussian mixture model with ICA presented in [7] is applied for the statistical learning of the clustering framework. Speaker classification based on supervised hierarchical clustering also serves the purpose to validate the effectiveness of ICA mixture model in speaker recognition and statistical learning. The gender speaker classification is performed on TIMIT and TSP speech databases and 10 speakers classification is conducted on TSP speech database. Both classification frameworks are also implemented using Gaussian mixture model in order to compare the performance of ICA mixture model in statistical learning. It is observed that classification framework based on hierarchical clustering performs well for both classification scenarios and ICA mixture model outperforms the GMM in model learning based on the classification rate. It is also observed that conventional problem of female speaker recognition is improved by employing multi-cluster model instead of classical model during the learning.

2 Supervised Hierarchical Clustering via ICA Mixture Model

In this section, supervised hierarchical clustering framework based on ICA mixture model is presented, which is applied to the speaker classification. The ICA mixture model is trained using training data and the posterior probability is employed to compute the specific cluster membership for each observation of the training data. The class label of the training data is selected to decode the clusters into particular class. The posterior probability is computed for the testing data and cluster-class information from the training is employed to find the particular class for each observation of the testing data. Since the class label of the training data is used to decode the clusters in the particular class and ICA mixture model is adapted for the statistical learning, therefore this framework is called the supervised hierarchical clustering framework based on ICA mixture model. Let us consider the training data represented as $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ where each observation is D-dimensional random variable $\boldsymbol{X}_i = (X_1, \ldots, X_D)$. The random variable $\boldsymbol{X}$ follows a K components mixture distribution if its probability distribution is written in the following form:

$$p(\boldsymbol{X}_i|\Theta) = \sum_{j=1}^{K} p(\boldsymbol{X_i}|\theta_j)p_j \tag{1}$$

provided $p_j \geq 0$ and $\sum_{j=1}^{K} p_j = 1$. In Eq. (1), $\Theta = \{p_1, \ldots, p_K, \theta_1, \ldots, \theta_K\}$ where θ_j is the set of parameters of the jth component and p_j represents the mixing proportion for the jth component of the mixture model. For the training data $\mathcal{X}$ having N independent and identically distributed vectors, the mixture model with K components can be expressed as follows:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{K} p(\boldsymbol{X_i}|\theta_j)p_j \tag{2}$$

For each random variable $\boldsymbol{X}_i$, let Z_i be a K dimensional vector representing the missing group indicator which suggests to which component $\boldsymbol{X}_i$ belongs, such that Z_{ij} will be equal to 1 if $\boldsymbol{X}_i$ belongs to class j and 0 otherwise. The complete data likelihood is then:

$$p(\mathcal{X}, Z|\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{K} (p(\boldsymbol{X_i}|\theta_j)p_j)^{Z_{ij}} \tag{3}$$

The complete data log-likelihood can be written as:

$$L(\Theta, Z, \mathcal{X}) = \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \log\left(p(\boldsymbol{X_i}|\theta_j)p_j\right) \tag{4}$$

By replacing each Z_{ij} by its expectation, defined as posterior probability that the ith observation belongs to jth component of the mixture model as follows:

$$Z_{ij} = p(j|\boldsymbol{X}_i) = \frac{p(\boldsymbol{X}_i|\theta_j)p_j}{\sum_{j=1}^{K} p(\boldsymbol{X}_i|\theta_j)p_j} \tag{5}$$

The membership of $\boldsymbol{X_i}$ computed from the posterior probability can be selected to mark the clusters into a particular class. This information will further help for decoding the clusters into particular class for testing data using the membership function of the posterior probability for the observations of test data. If testing data is represented as $\mathcal{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L)$, the posterior probability for $\boldsymbol{Y}_l$ can be computed using the trained mixture model and is represented as follows:

$$p(j|\boldsymbol{Y}_l) = \frac{p(\boldsymbol{Y}_l|\theta_j)p_j}{\sum_{j=1}^{K} p(\boldsymbol{Y}_l|\theta_j)p_j} \tag{6}$$

The supervised hierarchical framework for gender speaker classification is shown in Fig. 1. The speech data contains the MFCC features for male and female speakers and the class label is also provided. The ICA mixture model is trained in unsupervised fashion and the posterior probability for each observation of the

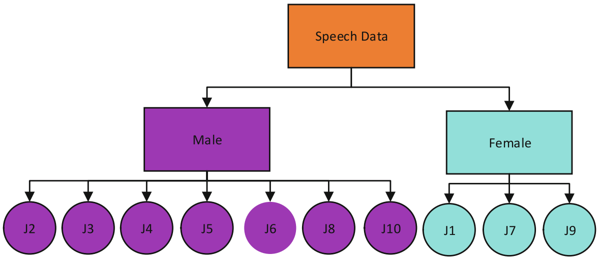


Fig. 1. Gender speaker classification using clustering

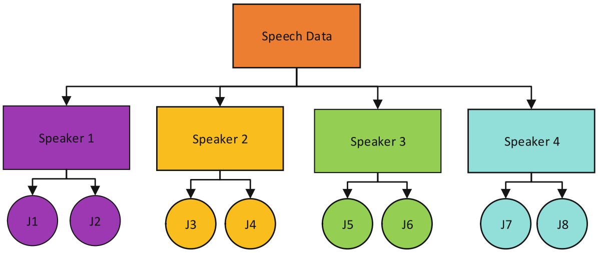


Fig. 2. Multi-speakers classification using clustering

training data is computed. The posterior probability marks each observation to a specific cluster and the class information of the training data can be selected to mark each cluster to a specific class to whom it belongs. For instance, if $\boldsymbol{X}_i$ belongs to the male class and it lies in the cluster 2, then cluster 2 is marked as male cluster. All the clusters can be marked as male or female from the training information and class label. In Fig. 1, it is assumed that the ICA mixture model is learned with 10 mixture densities and we have the class label for each observation. From posterior probability it is inferred that female observations from the speech data belongs to cluster J1, J7 and J9, so these clusters can be further labeled as female class and rest of the clusters were inferred as male class in the same way. It is worth mentioning that training of the ICA mixture model is unsupervised because the speech data is adopted without any class label during the training. However, the clustering framework is supervised because class label is employed after the training to mark the clusters into specific class. In the 10 speakers classification, the same binary classification framework is extended for 10 classes (see Fig. 2) and clusters obtained from the posterior probability are decoded into particular classes based on class label of the training data. In the classification using clustering, one important aspect is to accurately mark the number of classes representing data. In the classical approach, data is modeled by a fixed number of components of the mixture model which is equal to the number of classes. There are two problems associated with classical approach: (i) one single density component for each class does not necessarily fit the class data (ii) there is an overlap between the classes when using a single distribution to

Algorithm 1. Model Learning with BGGMM using ICA

```
Input:Dataset $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, $t_{min}$.
Output: $\Theta$.
{Initialization}: K-Means Algorithm. Set $\boldsymbol{\lambda} = 2$.
while relative change in log-likelihood $\geq t_{min}$ do
  {[E Step]}:
    for all $1 \leq j \leq K$ do
      Compute $p(\boldsymbol{X}_i|\theta_j)$ for $i = 1, \ldots, N$.
      Compute $p(j|\boldsymbol{X}_i)$ for $i = 1, \ldots, N$.
    end for
  {[M step]}:
    for all $1 \leq j \leq K$ do
      start ICA Algorithm
        Update the basis functions $\mathrm{A}_j$.
        Update the bias vector $\mathrm{b}_j$.
        Update the shape parameter $\boldsymbol{\lambda}_j$.
      end ICA
      Update the mixing parameter $p_j$.
      Update the mean $\boldsymbol{\mu}_j$.
      Update standard deviation $\boldsymbol{\sigma}_j$.
    end for
end while
```

model each class [6]. In speaker recognition, while modeling several speakers in one class or even a single speaker in one class may have the above problems. This is because the several speakers in a single class always have some distinct features and even same speaker will have dissimilar behavior while pronouncing the same words or utterances on different times. Due to the problems associated with classical model, we have adopted multi-cluster model which improve the learning of classification framework. There is another problem with the learning of female speakers and it is reported that speaker recognition performance of female speakers is almost worse as compare to the male speakers [8,9]. It is observed that in the multi-cluster modeling, the performance of female speakers is improved during learning for their particular class. Bounded generalized Gaussian mixture model with ICA proposed in [7] is employed as statistical model for learning which uses the maximization of log-likelihood and ICA model for the estimation of its parameters. In an ICA mixture model, it is assumed that observed data comes from a mixture model and it can be categorized into mutually exclusive classes which means that each class of the data is modeled as an ICA [10–12]. The mixture model represented in Eq. (2) is composed of bounded generalized Gaussian distributions (BGGDs) which has mean μ, standard deviation σ and shape parameter λ as its parameters. The idea of bounded support mixture models and bounded generalized Gaussian mixture model was proposed in [13] and [14] respectively. For the ICA mixture model, each D-dimensional data vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iD})$ can be represented as: $\boldsymbol{X}_i = \mathrm{A}_j \mathrm{s}_{j,i} + \mathrm{b}_j$ where A_j is $L \times D$ basis functions, $\mathrm{s}_{j,i}$ is D-dimensional source vector and b_j is an L-dimensional bias vector for a particular mixture j [10–12,15]. For the simplicity, number of linear combinations (L) is considered to be equal to the number of sources (D) for each observation of the dataset. In an ICA mixture model, we need to estimate the basis functions A_j and bias vector b_j along with the parameters of the mixture model. The parameters mean, standard deviation and prior probability are estimated using the maximization of the log-likelihood. The shape parameters, basis functions and bias vector are estimated using the standard ICA model and gradient ascent. The parameter estimation for BGGMM with ICA is provided in [7] and complete learning procedure is given in Algorithm 1.

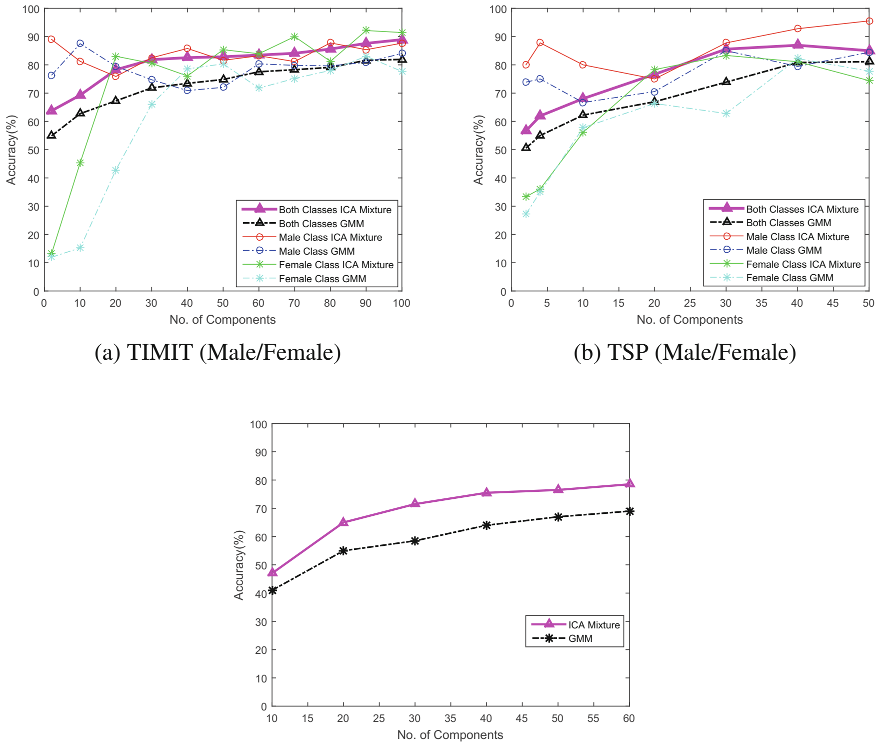
3 Experiments and Results

3.1 Design of Experiments

In this section, experimental framework for male/female and 10 speakers classification based on supervised hierarchical clustering is presented, which uses ICA mixture model for the statistical learning as described in section II. In the pre-processing stage, voice activity detection (VAD) is employed to distinguish between speech and non-speech parts of the speech sequences. By introducing the VAD in the pre-processing it is assured that the training of ICA mixture model is not inferred with the non-speech segments of the data set. The next stage is feature extraction and Mel Frequency Cepstral Coefficients (MFCCs) are selected as features. MFCCs have demonstrated their effectiveness in speech recognition and speaker classification and we have computed 13 dimensional features same as standard hidden Markov model toolkit (HTK). The ICA mixture model is trained using training part of the speech databases and the posterior probability is employed to determine the membership of an observation to a particular cluster. The class label for the training data is adopted to decode the clusters into particular class. The posterior probability is computed for the testing data and clustering information from the training is selected to find the particular class for each observation of the testing data. This classification framework is called the supervised hierarchical clustering based on ICA mixture model and presented in a detail in section II. This framework is also implemented using Gaussian mixture model in order to compare and examine the validity of the statistical learning of ICA mixture model in speaker classification.

3.2 Experimental Framework and Results

The speaker classification based on supervised hierarchical clustering is evaluated on TIMIT and TSP speech databases [16,17]. The TIMIT speech corpus consists of 6300 speech utterances which contains 4620 speech utterances for training and 1680 speech utterances for testing. The TSP speech database consists of 1378 speech utterances spoken by 23 speakers (11 male, 12 female). For gender speaker classification, 6 speakers are selected for testing from the TSP and rest of the data is dedicated for training. For 10 speakers classification, 10 speakers (5 male, 5 female) having 60 speech utterances for each speaker are selected from the TSP with 40 speech utterances for training and 20 utterances for testing. The TIMIT speech corpus is employed for gender speaker classification whereas TSP database is selected for both classification scenarios. In the clustering framework for both scenarios, each speech utterance is segmented into frames of 25 ms with a window shifting of 10 ms, where each frame is represented by 13 MFCCs. The VAD is applied before feature extraction in order to have only speech frames in the training and testing data. The k-means is employed to initialize the parameters of ICA mixture model, with shape parameter set to 2 for each component of the mixture model. For the gender speaker classification, ICA mixture model is trained using the training sets of both speech databases separately. From the posterior probability, speech utterances are divided into clusters by the membership of particular component of the mixture model. The class label for each utterances is provided for the training data which further leads to label the clusters into particular class. Once the clusters are labeled into the particular classes, the cluster-class information can be selected to decode the testing data into male/female speakers. The classification framework is evaluated using classification accuracy computed from the confusion matrices. For the TIMIT speech corpus, the classification accuracy is computed for different number of component of mixture model between 2–100 and plotted in Fig. (3a). In the classification accuracy curve for both classes, it is observed that by increasing the number of components of the mixture model, the classification rate is increased. However after 30 components of the mixture model, the increase in classification accuracy is slow. The classification framework having ICA mixture model is compared with the same framework having GMM on the basis of classification rate. The overall classification rate for ICA mixture model in the setting of 100 mixture components is 88.92 % whereas in same setting for GMM, the classification rate is 81.87 %. It is also noted that for smaller number of mixture components, the recognition of female speakers is very poor which is improved for higher number of mixture components. It is also observed that multi-cluster model has improved the model learning for both classes as compared to the classic model. In the classic model, the female speakers have poor performance while fitting the data in one class. In comparison with GMM, ICA mixture model has performed well which validates the effectiveness of ICA mixture model for speaker classification and statistical learning. For the TSP speech database, the speech utterances from 17 speakers (8 male, 9 female) are adopted to train the ICA mixture model whereas 6 speakers (half male, half female) are employed for the testing with each speaker having 60 speech utterances. The classification accuracy for different number of components of ICA mixture model and GMM in gender speaker classification framework is computed and plotted in Fig. (3b). The highest value for overall classification accuracy is observed at 40 mixture components (86.94 %) for ICA mixture model and at 50 mixture components (81.11 %) for GMM. For the 10 class speaker classification TSP speech database is employed for tuning the speaker classification framework. In this scenario, 10 speakers are chosen and 40 speech utterances for each speaker are selected for training and 20 speech utterances for each speaker are adopted for testing. The classification results are computed for different number of mixture components and the resulting confusion matrices for classic and multi-cluster models are shown in Table (1a), (1b) and (1c). In order to have a comparison of ICA mixture model with GMM for 10 speakers classification, the same framework is



(a) TIMIT (Male/Female)

(b) TSP (Male/Female)

(c) TSP (10 Speakers)

Fig. 3. Classification accuracy for male/female and 10 speakers using ICA mixture and GMM (Colour figure online)

implemented with GMM and overall classification rate is plotted for both models in Fig. (3c). The highest classification rate is observed at 60 mixture components for both scenarios of 10 speakers classification (78.50 % for ICA mixture & 69 % for GMM) which demonstrates the effectiveness of ICA mixture model in this setting.

Table 1. 10 speakers classification confusion matrix using TSP database.

(a) ICA Mixture, M=10

	MH	MI	MJ	MK	ML	FH	FI	FJ	FK	FL
MH	12	1	2	1	3	0	1	0	0	0
MI	2	9	1	4	1	1	0	1	0	1
MJ	1	3	11	1	2	0	1	0	0	1
MK	2	1	5	9	1	1	0	1	0	0
ML	1	1	2	1	10	1	1	1	2	0
FH	1	0	1	1	0	8	1	2	4	2
FI	0	1	0	2	1	5	7	1	1	2
FJ	0	0	1	1	0	0	1	12	2	3
FK	1	1	0	1	0	2	1	3	9	2
FL	1	1	0	1	0	1	2	5	2	7

(b) ICA Mixture, M=40

	MH	MI	MJ	MK	ML	FH	FI	FJ	FK	FL
MH	15	1	1	1	1	0	0	1	0	0
MI	0	13	2	2	1	1	0	0	0	1
MJ	1	1	17	1	0	0	0	0	0	0
MK	1	1	1	16	1	0	0	0	0	0
ML	0	1	0	1	18	0	0	0	0	0
FH	1	0	0	1	0	13	1	2	1	1
FI	0	1	0	0	0	1	15	1	1	1
FJ	0	0	1	0	0	1	1	14	1	2
FK	1	0	0	0	0	1	1	2	14	1
FL	0	0	0	0	1	1	0	1	1	16

(c) ICA Mixture, M=60

	MH	MI	MJ	MK	ML	FH	FI	FJ	FK	FL
MH	17	1	0	1	1	0	0	0	0	0
MI	1	16	1	0	1	1	0	0	0	0
MJ	0	1	18	0	0	1	0	0	0	0
MK	2	0	1	14	1	1	0	1	0	0
ML	0	1	2	1	13	1	1	0	0	1
FH	0	0	0	0	0	15	1	1	2	1
FI	0	0	0	0	0	1	17	1	0	1
FJ	0	1	0	0	0	1	0	16	1	1
FK	1	0	1	0	0	1	3	1	13	0
FL	0	0	0	0	0	1	1	0	0	18

4 Conclusion

In this paper supervised hierarchical clustering framework is presented which is adopted for speaker classification. The first stage of the clustering is performed by the ICA mixture model and in the second stage, clusters received from the posterior probability are further classified using the class label of the training data. The cluster-class label information from training process is used for the classification of testing data. The classification framework is validated on TIMIT and TSP speech corpora. This framework also validates the statistical learning of ICA mixture model proposed in [7]. In order to examine the performance of the ICA mixture model, the classification framework is also implemented with GMM and the classification accuracy in different modes is compared. The proposed framework having ICA mixture model is employed for gender and 10 speakers classification. It is concluded that supervised hierarchical clustering framework has performed considerably well for the speaker classification and ICA mixture model surpass the GMM in the classification rate and model learning. It is also concluded that multi-cluster model has improved the problem of female speakers to fit the class data as compared to classic model.

Acknowledgment. The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Hansen, J., Hasan, T.: Speaker recognition by machines and humans: a tutorial review. Sig. Process. Mag. IEEE **32**, 74–99 (2015)
2. Markowitz, J.: The many roles of speaker classification in speaker verification and identification. In: Mller, C. (ed.) Speaker Classification I. LNCS, vol. 4343, pp. 218–225. Springer, Berlin (2007)
3. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Sig. Process. **10**(1), 19–41 (2000)
4. Bourouis, S., Mashrgy, M.A., Bouguila, N.: Bayesian learning of finite generalized inverted Dirichlet mixtures: application to object classification and forgery detection. Expert Syst. Appl. **41**, 2329–2336 (2014)
5. Bdiri, T., Bouguila, N., Ziou, D.: Visual scenes categorization using a flexible hierarchical mixture model supporting users ontology. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, pp. 262–267, Herndon, VA, USA, 4–6 Nov 2013
6. Bdiri, T., Bouguila, N., Ziou, D.: Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling. Expert Syst. Appl. **41**, 1218–1235 (2014)
7. Azam, M., Bouguila, N.: Unsupervised keyword spotting using bounded generalized Gaussian mixture model with ICA. In: 2015 IEEE Global Conference on Signal and Information Processing (General Symposium), Orlando, USA (2015)
8. Nguyen, P., Le, T., Tran, D., Huang, X., Sharma, D.: Fuzzy support vector machines for age and gender classification. In: INTERSPEECH, pp. 2806–2809 (2010)

9. Vergin, R., Farhat, A., O'Shaughnessy, D.: Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In: Fourth International Conference on Spoken Language, ICSLP 1996, Proceedings, vol. 2, pp. 1081–1084 (1996)
10. Salazar, A.: ICA and ICAMM methods. In: On Statistical Pattern Recognition in Independent Component Analysis Mixture Modelling. Springer Theses, vol. 4, pp. 29–55. Springer, Berlin (2013)
11. Lee, T.-W., Lewicki, M.S.: The generalized Gaussian mixture model using ICA. In: International Workshop on Independent Component Analysis, ICA 2000, pp. 239–244 (2000)
12. Lee, T.-W., Lewicki, M.S., Sejnowski, T.J.: ICA mixture models for unsupervised classification with non-Gaussian sources and automatic context switching in blind signal separation. In: IEEE Transactions on Pattern Recognition and Machine Learning (2000)
13. Lindblom, J., Samuelsson, J.: Bounded support Gaussian mixture modeling of speech spectra. IEEE Trans. Speech Audio Process. **11**, 88–99 (2003)
14. Nguyen, T.M., Wu, Q.J., Zhang, H.: Bounded generalized Gaussian mixture model. Pattern Recogn. **47**, 3132–3142 (2014)
15. Lee, T.-W., Lewicki, M.S.: Unsupervised image classification, segmentation, and enhancement using ICA mixture models. IEEE Trans. Image Process. **11**(3), 270–279 (2002)
16. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993). http://www.ldc.upenn.edu/Catalog/LDC93S1.html
17. Kabal, P.: TSP Speech Database. Technical report, Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada (2002)