2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media)

A Hybrid Approach for Mobile Phone Clustering With Speech Recordings

Mou Wang, Xiao-Lei Zhang, Susanto Rahardja
CIAIC and School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, China
wangmou21@mail.nwpu.edu.cn,{xiaolei.zhang, susanto}@nwpu.edu.cn

***Abstract*—Acquisition device clustering based on speech recordings is a critical problem in the field of speech forensic, especially for mobile phone clustering (MPC). Previous studies on mobile phone recognition or clustering can be categorized mainly to two approaches. One approach utilizes handcraft features such as Mel-frequency cepstral coefficients (MFCCs), while the other uses learned features from neural networks. In this paper, we propose a hybrid system for MPC. Specifically, we first extract supervectors from MFCCs by a Gaussian mixture model and obtain the deep bottleneck features by a deep auto-encoder network. Then, we feed the two features to spectral clustering respectively, which outputs two low-dimensional vectors by the Laplacian eigen-decomposition of the spectral clustering. Finally, we fuse the two vectors and conduct clustering on the fused feature by *k*-means. The performance of the proposed method is evaluated on a public corpus—MOBIPHONE. The results show that the proposed method is effective, and moreover, the supervectors and deep bottleneck features provide complementary information of the intrinsic characteristics of the speech recordings recorded by the mobile phones.**

***Index Terms*—Acquisition device recognition, auto-encoder network, spectral clustering, Gaussian mixture model.**

I. Introduction

The information acquired from portable acquisition devices, e.g. mobile phone, has a huge potential in many applications, such as forensic evidence [1], [2], information security, robots and etc. Because the acquisition devices do not have the same frequency characteristics due to their electronic components and structures [3], each acquisition device provides its unique intrinsic characteristics in the acquired speech recordings, which can be used to identify itself [4]. This paper takes mobile phones as the representative acquisition devices, and their identification problem is addressed by mobile phone clustering (MPC) [5], [6].

Most previous studies on mobile phone recognition are supervised. Specifically, various audio features such as Mel-frequency cepstral coefficients (MFCCs) are first extracted [3], [7]–[10], and then, a classifier such as support vector machine [3], [4], [7] is trained for each acquisition device to identify other speech recordings. They all assume that the categories and numbers of mobile phone were known as a priori, which faces the following problems. First, the categories and numbers of mobile phones are not always available in practice. In addition, the categories of mobile phones increase rapidly, which makes it difficult to identify new mobile phones that is not in the categories of the training data. Moreover, in some real-world applications such as the information forensic, only the speech recordings are required to be identified and there is no need to recognize the specific identities of the acquisition devices [5]. In such cases, the mobile phone recognition problem becomes a clustering problem that does not need the prior information of the mobile phones and pre-trained classifiers.

Due to the above problems, Li *et. al* [5] conducted the first study on the MPC problem. They first extracted frame-level bottleneck features from a deep neural network (DNN), and then trained a Gaussian mixture model (GMM) on the bottleneck features for segment-level *deep Gaussian supervectors*. Finally, spectral clustering was applied to the supervectors for MPC. The work was not exactly an unsupervised method since label information was utilized when training the DNN. To remedy this problem, in [6], they further applied auto-encoder network to replace DNN, which caused the system to be completely unsupervised.

Most previous work on the acquisition devices clustering problem focused on extracting a good acoustic feature or developing a powerful clustering algorithm. Inspired by the work on acoustic scene analysis [11] where the handcraft features and deep representations learned by neural networks possessed complementary information, in this paper, we propose a hybrid system for MPC. We first extracted acoustic features from each recordings, and then trained a deep auto-encoder networks (DAE) and a Gaussian mixture model-universal background model (GMM-UBM) to extract deep representations and Gaussian supervectors respectively. Those two features are complementary in representing the intrinsic characteristics left behind by mobile phones in speech recordings. Finally, we combined the two features by spectral clustering, and further clustered the output features of the spectral clustering by *k*-means for MPC. The main contributions of this paper include:

- the exploration of an effective unsupervised method to fuse different types of feature;
- the evaluation of its effectiveness on MPC;
- the performance evaluation of the proposed method on a public corpus of speech recordings acquired by mobile phones.

The rest of this paper is organized as follows. Section 2

978-1-7281-2820-7/19/$31.00 ©2019 IEEE
DOI 10.1109/Ubi-Media.2019.00047

describes our method. Section 3 presents the experiments in detail. Finally, we conclude our work in Section 4.

II. Methods

The block diagram of the proposed method for MPC is shown in Fig. 1, where N_p is the number of mobile phones for clustering. The system includes three modules: auto-encoder network, GMM-UBM and spectral clustering. In the figure, the inputs are speech recordings and the first step is to extract the MFCC features from each speech recording, and then feeds the MFCC features into the auto-encoder and GMM-UBM respectively to extract bottleneck features and Gaussian supervectors, which are further transformed into low-dimensional vectors by the Laplacian eigen-decomposition of the spectral clustering. Finally, we concatenate the two low-dimensional vectors and conduct clustering on the combined feature by *k*-means algorithm. We present the system in detail as follows.

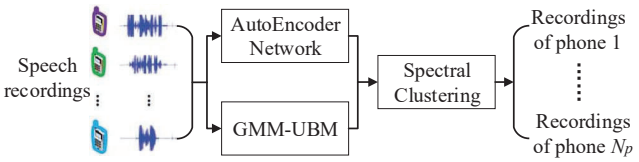

Fig. 1. Block diagram of the proposed method. N_p is the number of mobile phones for clustering

A. Auto-encoder network

As illustrated in Fig. 2, the deep auto-encoder network consists of two building blocks—an encoder and a decoder. The encoder f_E compresses the high-dimensional input ω into a low-dimensional representation x, i.e. $x = f_E(\omega)$. Then, the decoder f_D tries to reconstruct the original data ω from the encoded vector x as faithfully as possible, i.e. $\hat{\omega} = f_D(x)$ [12], where $\hat{\omega}$ is an estimate of ω. The network is trained to minimize the loss $L(\omega, \hat{\omega})$ between ω and $\hat{\omega}$, and therefore, does not need the label information during the training. The output x of the bottleneck layer, which is the narrowest hidden layer, of the deep auto-encoder network is used as a compact representation of the original high-dimensional inputs [13].

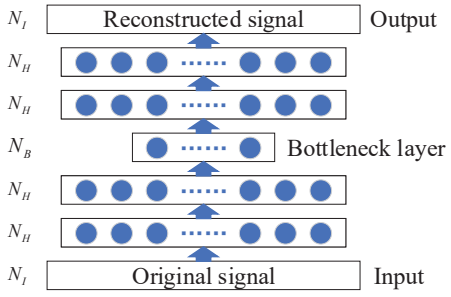

Fig. 2. Block diagram of the bottleneck feature extraction.

Our deep auto-encoder network adopts a similar structure and parameter setting from [6]. The difference is that we add a batch-normalization layer into our network, and do not pre-train the network with restricted Boltzmann machines, which leads to a simpler training process than the method in [6] and yields similar performance. To model the dynamic properties of the mobile phones and avoid overfitting, we expand each input frame with adjacent frames by a contextual window and reduce the dimensionality of the input with discrete cosine transform. We set the dimensions of the input and output layers of the auto-encoder, denoted both as N_I, to 624; we set the neuron number of the bottleneck layer N_b, i.e. the dimension of bottleneck feature, to 39 [6], and the neuron numbers of all other hidden layers to 500.

Finally, we average the frame-level bottleneck features of each speech recording in dimension for a segment-level feature of the recording. We denote the segment-level features as the bottleneck features of the speech recordings.

B. GMM-UBM

Motivated by [5], [14], we use the supervectors of the speech recordings extracted from GMM-UBM to represent the unique characteristics of the mobile phones [5]. Specifically, we first train a UBM from the MFCCs of all speech recordings. Suppose $\theta_{UBM} = \{w_m, \mu_m, \Sigma_m\}_{m=1}^{M}$ represents the parameters of a UBM with M Gaussian mixture components, where w_m, μ_m, and Σ_m represent the weight coefficient, mean vector and covariance matrix of the mth Gaussian mixture component, respectively. We first train the parameters by the expectation-maximization algorithm from all speech data. Then, we adapt a GMM $\theta_{GMM} = \{w'_m, \mu'_m, \Sigma'_m\}_{m=1}^{M}$ for each speech recording from the UBM by the maximum a posteriori (MAP) algorithm. Finally, we extract M mean vectors from each adapted GMM and concatenate the mean vectors successively as the supervector x of the corresponding speech recording. The length of the supervector is $M \times N_{mel}$, where N_{mel} is the dimensionality of the MFCC feature.

C. Spectral clustering

Spectral clustering conducts Laplacian eigen-decomposition to the affinity matrix of the input features to produce a new representation of the input features, and then does clustering on the new representation [15]. It has been proven to be effective in MPC [5], [6]. We present the usage of the spectral clustering in our system in detail as follows.

Suppose that x_n denotes a feature vector of the nth speech recording, which can be either a bottleneck feature or a Gaussian supervector, and X denotes the set of the feature vectors for clustering, i.e. $X = \{x_1, ..., x_N\}$, where N is the total number of the feature vectors. We first construct an affinity matrix A. It can be either a Gaussian kernel or a cosine kernel [16]. The cosine kernel is defined as

$$A_{kl} = \frac{\langle \mathbf{x}_k, \mathbf{x}_l \rangle}{|\mathbf{x}_k||\mathbf{x}_l|}, \quad 1 \le k, l \le L. \tag{1}$$

We also implement a Gaussian kernel [5] as

$$A_{kl} = \exp\left(-\frac{d(x_k, x_l)^2}{2\sigma_k\sigma_l}\right), \quad 1 \le k, l \le L \tag{2}$$

where $d(x_k, x_l)$ is the Euclidean distance between x_k and x_l, and σ_k (or σ_l) is a scaling factor for the feature vector x_k (or x_l). The scaling factor σ_k is defined in a nonparametric way by the nearest neighbor optimization:

$$\sigma_k = \Sigma_{x_l \in close(x_k)} d(x_k, x_l)/Q, \tag{3}$$

where $close(x_k)$ denotes the set containing the Q nearest neighbors of x_k. We set $Q = 5$ in this study. Then, we create a normalized Laplacian matrix L by

$$L = D^{-1/2} A D^{1/2}, \tag{4}$$

where D is a diagonal matrix whose element D_{ii} is the sum of all elements of the ith row of *A*. Decomposing L by eigenvalue decomposition produces the eigenvalues $\{\lambda_n\}_{n=1}^N$ and their corresponding eigenvectors $\{s_n\}_{n=1}^N$ of L. We choose the eigenvectors that corresponds to the largest N_c eigenvalues to form a matrix $S = [s_1, s_2, ..., s_{N_c}] \in R^{N \times N_c}$, where N_c is the number of clusters. Then, we generate a matrix Y by renormalizing each row of S,

$$Y_{ij} = \frac{S_{ij}}{(\Sigma_j S_{ij}^2)^{1/2}}, \quad 1 \le i \le N, 1 \le j \le N_c. \tag{5}$$

The jth row of Y is a new representation of the ith speech recording produced by spectral clustering, and the speech recordings is partitioned with the new representation into N_c clusters by *k*-means algorithm.

D. Fusion strategies

Motivated from a discussion on feature fusion strategies for supervised acoustic scene analysis [11], we first tried to concatenate the bottleneck features and Gaussian supervectors as the input of spectral clustering. However, we found that the system with the concatenated acoustic feature did not lead to better performance than the systems with the two features separately, which was quite different from its supervised counterpart. The reason led to this poor performance is that supervised learning focuses on finding classification hyperplanes among different categories, and hence would discard characteristic information of the samples. In contrast, unsupervised learning focuses on the measurement of the distances between the samples in a feature space. Therefore, some feature fusion approaches that are suitable to supervised learning may be unsuitable to unsupervised learning.

In this paper, an effective feature fusion strategy is proposed. We first feed the bottleneck features and Gaussian supervectors into the Laplacian eigen-decomposition module of two spectral clusterings respectively, which outputs two low-dimensional vectors. Then, we concatenate the two low-dimensional vectors as the input of *k*-means for MPC.

III. Experiments

A. Experimental settings

We evaluated the proposed method on a public corpus of speech recordings, MOBIPHONE, which is a popular corpus used in the previous studies [8]. It was acquired by 21 different mobile phones from 7 brands viz. HTC, LG, Nokia, Sony Ericsson, Apple, Samsung, and Vodafone. The sampling frequency of the audio data is 16 kHz. The dataset includes 24 speakers (12 males and 12 females) randomly chosen from the TIMIT database. Each speaker consists of 10 utterances. Each utterance is about 3 seconds long. The contents of the first two utterances are the same for all speakers, and the contents of the remaining utterances are different. To summarize, the dataset contains 5040 speech recordings with 240 speech recordings per mobile phone.

We first removed the silence segments of the speech recordings by an energy-based voice activity detection algorithm. Then, we extracted 13-dimensional MFCCs and their delta and double-delta coefficients from the non-silence segments, which amounted to 39-dimensional acoustic features, where a 30-millisecond Hamming window with a half overlap was applied to the feature extraction. The number of the mixture components of the GMM M was set to 256.

We measured the clustering quality between the produced clusters and the ground truth categories in terms of normalized mutual information (*NMI*) [17] and clustering accuracy (*ACC*). Let n_{ij} be the number of the speech recordings in cluster i acquired by mobile phone j, $n_{\cdot j}$ be the total number of speech recordings acquired by mobile phone j, and $n_{i\cdot}$ be the total number of speech recordings in cluster i, then we have:

$$N = \sum_{i=1}^{N_c}\sum_{j=1}^{N_p} n_{ij}, \quad n_{i\cdot} = \sum_{j=1}^{N_p} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{N_c} n_{ij}. \tag{6}$$

NMI was proposed to overcome the label indexing problem between the ground-truth labels and the predicted labels. It is one of the standard evaluation metrics of unsupervised learning. Note that *NMI* has a strong one-to-one correspondence with classification accuracy. The *NMI* score is defined as

$$NMI = \frac{\sum_{i=1}^{N_c}\sum_{j=1}^{N_s} n_{ij}\log(\frac{N \times n_{ij}}{n_{i\cdot} \times n_{\cdot j}})}{\sqrt{(\sum_i n_{i\cdot}\log\frac{n_{i\cdot}}{N})(\sum_j n_{\cdot j}\log\frac{n_{\cdot j}}{N})}}. \tag{7}$$

The *ACC* is defined as the maximal classification accuracy among all possible permutation mappings,

$$ACC = \left[\sum_{i=1}^{N}\delta(y_i, \mathrm{map}(c_i))\right]/N, \tag{8}$$

where y_i and c_i denote the true label and predicted cluster label of the ith speech recording respectively. $\delta(y, c)$ is a function that is equal to 1 if $y = c$, and 0 otherwise. $\mathrm{map}(\cdot)$ is a function that finds the optimal matching between the true labels and each permutation of the predicted cluster labels by the Hungarian algorithm. The higher the *NMI* and *ACC* scores are, the better the clustering quality is.

B. Main results

We first evaluated several common features adopted in the previous studies respectively, i.e. MFCCs [3], Gaussian Supervector (GS) [7], I-Vector (IV) [18], deep bottleneck features (DBF) from deep auto-encoder networks [6]. We also implemented deep representation (DR) [6], which is the state-of-the-art method for MPC. In [6], the authors first extracted a bottleneck feature for each recording by a deep auto-encoder. Then, they trained a GMM-UBM on the bottleneck features, and extracted a new high-dimensional feature, named deep representation, from the GMM-UBM for each recording. Finally, they conducted spectral clustering with the deep representation. In our system, we extracted features from the auto-encoder network and GMM-UBM respectively in parallel, and focus on developing an advanced fusion strategy of the features in an unsupervised way. For a fair comparison, we set the similar modules of our system and the system in [6], including the GMM-UBM, deep auto-encoder, spectral clustering, etc., with the same parameter settings, even though the parameter settings may not be optimal to our system. Here, we just present the result of fusing the Gaussian supervetors and deep bottleneck features so as to show the complementarity of the features.

TABLE I
PERFORMANCE OF THE COMPARISON METHODS. THE TERM "GS" IS SHORT FOR GAUSSIAN SUPERVECTOR. THE TERM "DBF" IS SHORT FOR DEEP BOTTLENECK FEATURES FROM DEEP AUTO-ENCODER NETWORKS.

	MFCC	I-vector	GS	DBF	DR	Fusion
NMI	82.08	90.80	90.25	91.5	93.96	**94.70**
ACC	72.81	88.69	88.69	87.8	93.73	**94.46**

As shown in Table I, the proposed method obtains the best performance in terms of both *NMI* and *ACC*. The result manifests that the handcraft feature and the deep representation from the deep auto-encoder provides complementary information, and our fusion method can boost the performance by utilizing the complementary information together.

Although the performance of the competing DR systems is quite close to our method, we find in our experiment that its performance is not robust since the deep auto-encoder and GMM-UBM is cascaded. Specifically, it achieves the best performance when the input of the GMM-UBM, which is the output of the deep auto-encoder, follows the Gaussian assumption of the GMM-UBM. However, the true hypothesis of the Gaussian assumption made on the output of the deep auto-encoder is not guaranteed. We find empirically that the output distribution of the deep auto-encoder is non-Gaussian in most types of hidden neurons, which limits the upper-bound performance and real-world applications of the comparison system. On the contrary, the auto-encoder and GMM-UBM of our system run in parallel without interaction. Thus, our method is simpler and more robust than the DR system.

C. The effects of different fusion strategies

To show the effectiveness of the proposed fusion strategy, we compare the following three fusion strategies, which are the strategies of (i) concatenating the Gaussian supervector and bottleneck feature for the input of the spectral clustering, (ii) averaging the two affinity matrices produced from the Gaussian supervector and bottleneck feature respectively, and (iii) the proposed one. We denote the three strategies as Fusion-1, Fusion-2 and Fusion-3.

Fig. 3 shows the performance of the competing fusion methods. From the figure, we find that Fusion-1 does not show performance improvement over the systems that adopt the individual feature only. This is mainly caused by the fact that the Gaussian supervector and bottleneck feature are not in the same density space, hence concatenating them forcefully breaks the continuity of their density spaces. We also see that the other two fusion methods are both effective, while the proposed Fusion-3 is slightly better than Fusion-2.

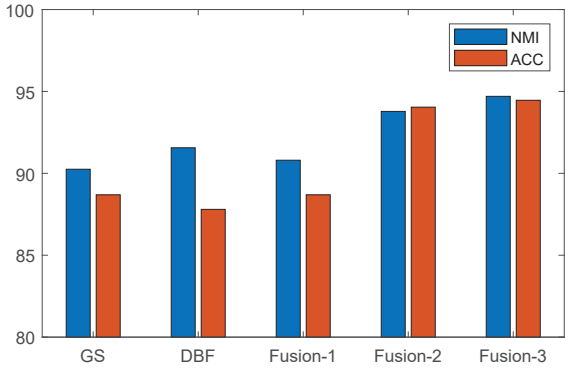

Fig. 3. *NMI* and *ACC* comparison (in percent) of different fusion systems.

IV. CONCLUSIONS

In this paper, we propose a hybrid MPC system for clustering the acquired speech recordings of mobile phones. For each speech recording, the system first extracts a Gaussian supervector from the GMM-UBM and a deep bottleneck feature from the deep auto-encoder networks, and then projects the two features into two low-dimensional representations by the Laplacian eigen-decomposition of the spectral clustering respectively. Finally, it concatenates the two low-dimensional representations as the final representation of the speech recording for clustering. The experimental results show the proposed hybrid method slightly outperforms the state-of-the-art baseline, and performs more stable than the latter. From this study, we also show that the handcraft features such as MFCCs and the deep representation from neural network are complementary in unsupervised learning. In addition, we explore different unsupervised feature fusion schemes on the MPC task, and find that conducting the feature fusion after the Laplacian eigen-decomposition achieves the best performance.

REFERENCES

[1] H. Zhao and H. Malik, "Audio recording location identification using acoustic environment signature," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1746–1759, November 2013.

[2] H. Malik and H. Zhao, "Recording environment identification using acoustic reverberation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 1833–1836.

[3] C. Hanilci, F. Ertas, T. Ertas, and O. Eskidere, "Recognition of brand and models of cell-phones from recorded speech signals," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 625–634, April 2012.

[4] C. L. Kotropoulos, "Source phone identification using sketches of features," *IET Biometrics*, vol. 3, no. 2, pp. 75–83, June 2014.

[5] Y. Li, X. Zhang, X. Li, X. Feng, J. Yang, A. Chen, and Q. He, "Mobile phone clustering from acquired speech recordings using deep gaussian supervector and spectral clustering," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2137–2141.

[6] Y. Li, X. Zhang, X. Li, Y. Zhang, J. Yang, and Q. He, "Mobile phone clustering from speech recordings using deep representation and spectral clustering," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 965–977, April 2018.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 1806–1809.

[8] C. Kotropoulos and S. Samaras, "Mobile phone identification using recorded speech signals," in *2014 19th International Conference on Digital Signal Processing*, August 2014, pp. 586–591.

[9] L. Zou, Q. He, and X. Feng, "Cell phone verification from speech recordings using sparse representation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 1787–1791.

[10] L. Zou, Q. He, J. Yang, and Y. Li, "Source cell phone matching from speech recordings by sparse representation and kiss metric," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2079–2083.

[11] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[13] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH-2011*, August 2011, pp. 237–240.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 1806–1809.

[15] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 849–856.

[16] V. Arora and L. Behera, "Musical source clustering and identification in polyphonic audio," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 22, no. 6, pp. 1003–1012, 2014.

[17] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, September 2003.

[18] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.