Integrating Online *i*-vector into GMM-UBM for Text-dependent Speaker Verification

Xiaowei Jiang, Shuai Wang, Xu Xiang, Yanmin Qian
Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China
E-mail: {niujincidian, feixiang121976, chinoiserie, yanminqian} @sjtu.edu.cn

***Abstract*—GMM-UBM is widely used for the text-dependent task for its simplicity and effectiveness, while *i*-vector provides a compact representation for speaker information. Thus it is promising to fuse these two frameworks. In this paper, a variation of traditional *i*-vector extracted at frame level is appended with MFCC as tandem features. Incorporating this feature into GMM-UBM system achieves 26% and 41% performance gain compared with DNN *i*-vector baseline on the RSR2015 and RedDots evaluation set, respectively. Moreover, the performance of the proposed system that trained on 86 hours data is on par with that of the DNN *i*-vector baseline trained on a much larger dataset (5000 hours).**

I. Introduction

Speaker verification is the task of identifying whether the target speaker speaks in a test utterance. According to the text contents of the test utterances, speaker verification can be classified into two categories, text-dependent and text-independent. For the text-dependent task, the contents of the test and target utterances are restricted to be identical, whereas the text-independent task does not have such constraint.

Over the last few decades, a variety of frameworks were proposed for the speaker verification task. In [1], Gaussian Mixture Model-Universal Background Model (GMM-UBM) with Maximum a Posteriori framework was proposed. In this framework, the UBM represents a speaker independent model and the speaker-specific GMM is adapted from it. Based on GMM-UBM framework, Joint Factor Analysis (JFA) was then applied to model the supervectors adapted from UBM in independent speaker and channel subspace [2]. However, work in [3] showed that the channel factors in JFA also contains speaker information. *I*-vector was then proposed, where the speaker and channel variability are jointly modeled by a single total variability subspace [4].

Recently, motivated by the success of Deep Neural Network (DNN) in automatic speech recognition (ASR) [5], [6], [7], the use of DNN in speaker verification is intensively investigated [8], [9], [10], [11], [12]. Other than GMM, DNN provides an alternative way to calculate the Sufficient Statistics (SS) for estimating the *i*-vector. It is observed that DNN posterior based *i*-vector system achieves significant improvement over GMM posterior based *i*-vector system [10].

In spite of the superiority of *i*-vector based framework in text-independent speaker verification, GMM-UBM framework is reported to achieve better performance for the text-dependent task [13]. Since *i*-vectors carry elaborate speaker information, it is expected that incorporating *i*-vector based features into GMM-UBM framework can make further improvement. To be compatible with GMM-UBM framework, a frame level *i*-vector called online *i*-vector is used in this work.

In this paper, we propose to concatenate online *i*-vector and Mel-Frequency Cepstral Coefficients (MFCC) in the tandem manner and use it as features for a GMM-UBM text-dependent speaker verification system. The performance of the proposed system is evaluated on RSR2015 and RedDots datasets.

The rest of the paper is organized as follows. Sec. II shows the conventional speaker verification systems. After the brief introduction of baseline systems, Sec. III describes the online *i*-vector extraction procedure. In Sec. IV, we will present the details of our proposed GMM-UBM systems with online *i*-vector features. The experiment setups and results are shown and discussed in Sec. V and Sec.VI. The conclusion is presented in Sec. VII.

II. Baseline Systems

A. GMM-UBM System

GMM-UBM [1] framework is a classical approach used in speaker verification systems. Building a GMM-UBM system has several phases:

- Feature Extraction. The baseline system adopts MFCC as features.
- Training a speaker-independent background model using huge amount of data from different speakers.
- Obtain the speaker-specific GMM by adapting the trained UBM parameters via MAP algorithm.
- Compute the log likelihood ratio of the test utterances against the declared speaker GMM and UBM. The score s depending on both the target model (λ_{target}) and background model (λ_{UBM}) is defined as follows,

$$s = \frac{1}{L}\sum_{t=1}^{L}\{\log p(\mathbf{x}_t|\lambda_{target}) - \log p(\mathbf{x}_t|\lambda_{UBM})\} \quad (1)$$

which measures the difference of the target and background models in generating the observations $\mathbf{x}_1, \ldots, \mathbf{x}_L$

B. GMM Posterior Based *i*-vector System

In the *i*-vector framework, the speaker- and session-dependent supervector $\mathbf{M}$ is modeled as

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{2}$$

where $\mathbf{m}$ is the CF-dimensional mean supervector of UBM , C is the number of Gaussian components and F represents the feature dimension. $\mathbf{T}$ is a rectangular low-rank matrix which captures speaker and session variability. $\mathbf{w}$ is a realization of a latent variable $\mathbf{W}$ having a standard normal prior distribution. For each supervector adapted from an utterance, the speaker information is assumed to be contained in $\mathbf{w}$. Suppose the input utterance consists of L frames, the acoustic features are represented as a set of F-dimensional vectors: $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_L}\}$. The *i*-vector for the utterance is defined as the point estimation of the conditional distribution of $\mathbf{W}$ given the utterance. The *i*-vector of the utterance can be calculated as follows:

$$\mathbf{\Phi} = (\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}(\mathcal{X})\mathbf{T})^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{F}}(\mathcal{X}) \tag{3}$$

where $\mathbf{\Sigma}$ is a diagonal covariance matrix of shape $(CF \times CF)$ that describes the residual variability not captured by $\mathbf{T}$ matrix. $\mathbf{N}(\mathcal{X})$ is a diagonal matrix whose diagonal blocks are $N_c\mathbf{I}(c = 1, 2, ..., C)$ and $\tilde{\mathbf{F}}(\mathcal{X})$ is the supervector obtained by stacking $\tilde{\mathbf{F}}_c$. The sufficient statistics are calculated as follows:

$$\gamma_c(\mathbf{x_t}) = p(c|\mathbf{x_t}, \lambda_{UBM}) \tag{4}$$

$$N_c = \sum_{t=1}^{L} \gamma_c(\mathbf{x_t}) \tag{5}$$

$$\tilde{\mathbf{F}}_c = \sum_{t=1}^{L} \gamma_c(\mathbf{x_t})(\mathbf{x_t} - \mathbf{m_c}) \tag{6}$$

where $\gamma_c(\mathbf{x_t})$ and $\mathbf{m_c}$ are the occupation probability and mean vector of c-th Gaussian component, respectively.

C. DNN Posterior Based *i*-vector Systems

In conventional *i*-vector systems described in the previous subsection, posteriors $\gamma_c(\mathbf{x_t})$ used for the calculation of sufficient statistics are derived from the UBM. However, it is shown in [10] that with the posteriors obtained from a phonetically-aware DNN, the *i*-vector system can achieve significant performance gain. In this framework, the use of DNN "senone" (context-dependent triphones) posterior is proposed to compute the alignments $\gamma_c(\mathbf{x_t})$, where c denotes the c-th senone used in the phonetically-aware DNN. In this paper, time-delay deep neural network (TDNN) [11] is used.

III. Online *i*-vector Extraction

Online *i*-vectors are *i*-vectors extracted from short segments of speech utterances, which makes it possible to represent short duration speaker characteristics of speech utterances. The online *i*-vector has been investigated in ASR [14], speaker diarization [15] and speaker verification system [13]. Different from traditional *i*-vectors which are extracted at utterance level, online *i*-vectors are extracted every $2L+1$ (context size $L = 10$ in our proposed systems) frames with a shift step of 1 frame. The sufficient statistics of an online *i*-vector are computed with posteriors either from a GMM-UBM or from a phonetically-aware DNN. As online *i*-vectors are extracted at frame level, it can be used like other frame-wise features, such as MFCC, to model speaker specific traits better. In this paper, we propose to use online *i*-vectors as features to construct a series of GMM-UBM systems for text-dependent speaker verification.

Conventionally, the training of the $\mathbf{T}$ matrix accumulates sufficient statistics at utterance level. In this paper, the data for $\mathbf{T}$ matrix training is drawn from NIST SRE and Switchboard datasets with an average duration per utterance of 5 to 6 minutes. However, the extraction of each online *i*-vector in this paper is performed on a short segment with a duration of only 21 frames. Considering the consistency between training process and extraction process, the training utterances are cut into small segments. The impact of such preprocessing step on the system performance can be found in the experiment part.

IV. Online *i*-vector Based GMM-UBM Systems

GMM-UBM system demonstrates robust performance for text-dependent speaker verification system, while *i*-vector exhibits excellent performance in text-independent systems. Frame-level online *i*-vector is optimized to carry more "well-organized" speaker identity information, hence it can be adopted as features in the traditional GMM-UBM system. In this paper, we investigated two paradigms of integrating online *i*-vector features into GMM-UBM system, using online *i*-vector only or concatenated with MFCC in a tandem manner. Experiments show that the new tandem feature achieves promising performance improvement compared with the baseline system. The diagram of the proposed system is shown in Fig. 1.

V. Experimental Setups

A. Training and Evaluation Datasets

All experiments in this paper are performed on 8 kHz speech files. Switchboard dataset ($\sim$ 300 hours) [16] is used to train the phonetically-aware DNN [10]. *I*-vector extractor is trained on a 86-hour subset of NIST SRE 2004-2008, Switchboard Cellular 1&2 and Switchboard Phase 2&3 datasets. RSR2015 part1 background data ($\sim$ 24 hours, down sampled to 8 kHz [1]) is taken as the development data for the training of PLDA and the training of UBM in GMM-UBM systems. RSR2015 part1 and RedDots part1 (down-sampled to 8 kHz) are chosen as the evaluation datasets. Both of them are designed for short duration text-dependent speaker verification. In text-dependent speaker verification, three test conditions are defined according to three impostor types: (1) the content does not match (2) the speaker does not match (3) neither the speaker nor the content match.

[1]Compared with using 16 kHz data, there is a reasonable deterioration in performance using 8 kHz data.

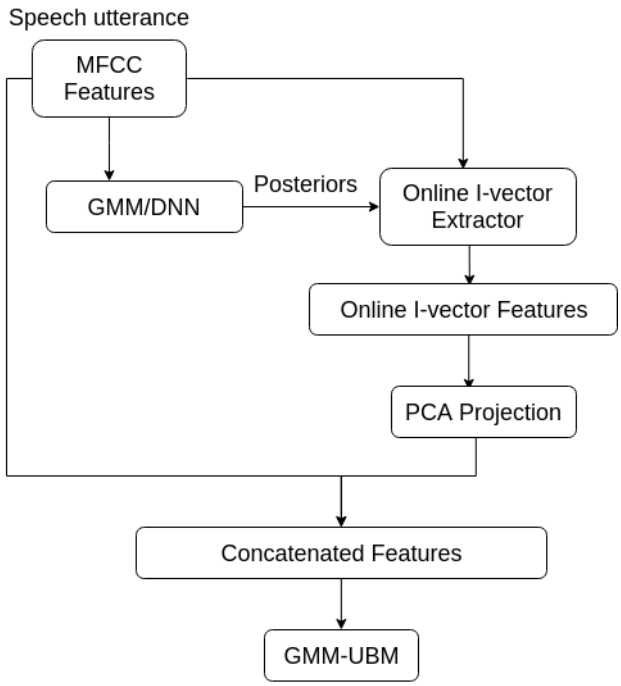

Fig. 1. The diagram of proposed online *i*-vector based GMM-UBM system

- RSR2015 part1: a close set text-dependent speaker verification evaluation dataset in English language. This dataset aims at providing a database for the study on lexical variability in text-dependent verification. The detailed description for RSR2015 evaluation dataset can be found in [17]. In this paper, following the settings of [17] the trials of condition 3 in RSR2015 part1 are excluded as these are easy trials.
- RedDots part1: an open set text-dependent speaker verification evaluation dataset in English language. The speech utterances are collected from 62 speakers through mobile crowd-sourcing over a one year period. Compared with RSR2015 part1, the RedDots part1 corpus shows a high degree of intra-speaker variations due to the long recording period and various recording conditions. The detailed description for RedDots project can be found in [17].

B. Baseline Systems

The acoustic features used in baseline systems are 20-dimensional MFCC features extracted from 25ms duration frames with a frame shift of 10ms, appended with delta and acceleration parameters. All features are processed using cepstral mean normalization. In GMM-UBM baseline system, these features are taken as the input features for UBM training and scoring. In the *i*-vector systems, the MFCC features are used for sufficient statistics calculation with a UBM model or a DNN model. All the UBMs in this paper have 1024 Gaussian mixture components. The dimension of *i*-vectors is set to 600. The DNN for posterior calculation is trained with 5419 output units and it takes 40-dimensional MFCC features appended with delta and acceleration parameters as input. A time delay deep neural network (TDNN) is used instead of the traditional feed-forward deep neural network. The same configuration as in [11] is adopted (The standard recipe provided in Kaldi egs/sre10/v2). The descriptions of three baseline systems are listed below:

- **MAP (MFCC)**: GMM-UBM system with 60-dimensional MFCC features only.
- ***i*-vector**: GMM posterior based *i*-vector system with 600-dimensional *i*-vectors, scoring with a PLDA backend.
- **DNN-*i*-vector**: DNN posterior based *i*-vector system, 600-dimensional *i*-vectors, scoring with a PLDA backend.

C. Online i-vector Based GMM-UBM Systems

The online *i*-vector based systems are built on top of GMM-UBM framework. The $\mathbf{T}$ matrix for online *i*-vector extraction is trained on short segments with a length of 21 frames. Those short segments are directly cut from the original training utterances. Considering the computation limitation, the online *i*-vectors are further projected using Principle Component Analysis (PCA) into 60-dimensional features. We proposed to use the concatenation of projected online *i*-vectors with the original 60-dimensional MFCC features, as the input to a GMM-UBM system. The detailed description of the experiments are listed below:

- **MAP (online)**: GMM-UBM system with 60-dimensional PCA projected online *i*-vector features
- **MAP (concat)**: GMM-UBM system with concatenated 60-dimensional PCA projected online *i*-vector features and 60-dimensional MFCC features
- **MAP (DNN-online)**: GMM-UBM system with 60-dimensional PCA projected online *i*-vector features extracted using DNN posteriors
- **MAP (DNN-concat)**: GMM-UBM system with concatenated 60-dimensional PCA projected online *i*-vector features extracted using DNN posteriors and 60-dimensional MFCC features

VI. Experiment Results

A. Comparison of proposed system and baseline

In this section, the experiment results are shown in Equal Error Rate (EER) performance metric. As shown in Tab. I, II, the concatenated tandem feature based system outperforms the systems based on MFCC or online *i*-vector, which shows that MFCC features and online *i*-vector features are complementary to each other.

The best baseline system is the "**DNN-*i*-vector**" system. It can be observed that the proposed "**MAP (DNN-concat)**" system obtained 41% EER relative reduction over the best baseline system on RedDots evaluation dataset. On RSR2015 part1, the EER is reduced by 26% with the proposed system. A larger performance improvement on RedDots evaluation set is achieved, which exhibits the robustness of the proposed system in complicated evaluation condition.

TABLE I
PERFORMANCE OF PROPOSED SYSTEMS ON REDDOTS

	system	cond-1	cond-2	cond-3	cond-all
baseline systems	**MAP(MFCC)**	8.75	6.40	2.14	3.12
	***i*-vector**	14.91	9.00	5.19	5.99
	DNN-*i*-vector	**6.86**	**5.52**	**2.32**	**3.12**
proposed systems	**MAP (online)**	11.12	7.92	4.54	5.29
	MAP (concat)	4.39	4.39	1.21	1.96
	MAP (DNN-online)	7.48	6.37	2.68	3.61
	MAP (DNN-concat)	**3.82**	**4.46**	**0.90**	**1.83**

TABLE II
PERFORMANCE OF PROPOSED SYSTEMS ON RSR2015

	system	cond-1	cond-2	cond-all
baseline systems	**MAP(MFCC)**	1.19	2.44	2.08
	***i*-vector**	2.40	3.00	2.78
	DNN-*i*-vector	**0.50**	**1.44**	**1.22**
proposed systems	**MAP (online)**	1.59	2.24	2.02
	MAP (concat)	0.29	1.33	1.10
	MAP (DNN-online)	0.66	1.27	1.17
	MAP (DNN-concat)	**0.14**	**1.11**	**0.90**

B. Comparison of proposed system and baselines trained on 5000-hour data

Another three baseline systems are built on a larger training dataset (about 5000 hours), including NIST SRE 2004-2008, Switchboard Cellular 1&2 and Switchboard Phase 2&3. As shown in Tab. III and Tab. IV, on the RedDots evaluation dataset, the proposed system trained on 86 hours subset still outperforms slightly the baseline systems trained on 5000 hours data. On RSR2015 evaluation dataset, the proposed system also achieves comparable performance compared with the baseline systems. Moreover, this observation verifies the robustness of proposed method under complicated evaluation condition once again.

TABLE III
PROPOSED SYSTEM V.S. BASELINE SYSTEMS (5000 HOURS) ON REDDOTS

hours	system	cond-1	cond-2	cond-3	cond-all
5000	**MAP(MFCC)**	9.16	6.91	2.32	3.30
	***i*-vector**	8.00	6.45	2.30	3.23
	DNN-*i*-vector	**5.44**	**4.13**	**1.50**	**2.17**
86	**MAP (DNN-concat)**	**3.82**	**4.46**	**0.90**	**1.83**

TABLE IV
PROPOSED SYSTEM V.S. BASELINE SYSTEMS (5000 HOURS) ON RSR2015

hours	system	cond-1	cond-2	cond-all
5000	**MAP(MFCC)**	1.19	2.44	2.08
	***i*-vector**	0.94	1.39	1.27
	DNN-*i*-vector	**0.43**	**1.00**	**0.86**
86	**MAP (DNN-concat)**	**0.14**	**1.11**	**0.90**

C. Effectiveness of short segment training

As described in Sec. III, the length of training utterances for $\mathbf{T}$ matrix training should be consistent with that of the short segments for online *i*-vector extraction. To verify the effectiveness of the proposed $\mathbf{T}$ matrix training method, we conducted another set of experiments with the $\mathbf{T}$ matrix trained on original utterances of full length. As indicated in Tab. V and Tab. VI, short segment training can achieve consistent performance improvement. The experiment results reflect the effectiveness of the proposed training method of $\mathbf{T}$ matrix for online *i*-vector extraction.

TABLE V
COMPARISON OF TWO $\mathbf{T}$ MATRIX TRAINING METHODS IN GMM-UBM FRAMEWORK EVALUATED ON REDDOTS

	system	cond-1	cond-2	cond-3	cond-all
full[a] segment training	**MAP(online)**	20.18	9.88	6.63	7.48
	MAP(concat)	5.96	4.93	1.60	2.55
	MAP(DNN-online)	14.50	8.38	4.98	5.78
	MAP(DNN-concat)	**5.26**	**4.72**	**1.47**	**2.22**
short[a] segment training	**MAP (online)**	11.12	7.92	4.54	5.29
	MAP (concat)	4.39	4.39	1.21	1.96
	MAP (DNN-online)	7.48	6.37	2.68	3.61
	MAP (DNN-concat)	**3.82**	**4.46**	**0.90**	**1.83**

[a] full/short segment training indicates the training of $\mathbf{T}$ matrix is performed on original length utterances and short segments.

TABLE VI
COMPARISON OF TWO $\mathbf{T}$ MATRIX TRAINING METHODS IN GMM-UBM FRAMEWORK EVALUATED ON RSR2015

	system	cond-1	cond-2	cond-all
full segment training	**MAP (online)**	3.87	3.27	3.49
	MAP (concat)	0.41	1.42	1.17
	MAP (DNN-online)	1.65	1.96	1.84
	MAP (DNN-concat)	**0.29**	**1.20**	**0.99**
short segment training	**MAP (online)**	1.59	2.24	2.02
	MAP (concat)	0.29	1.33	1.10
	MAP (DNN-online)	0.66	1.27	1.17
	MAP (DNN-concat)	**0.14**	**1.11**	**0.90**

VII. CONCLUSIONS

In this paper, we have presented the application of online *i*-vectors based on GMM-UBM framework for the text-dependent speaker verification task. The proposed "**MAP (DNN-concat)**" system achieves 26% and 41% performance gain compared with DNN *i*-vector baseline on the RSR2015 and RedDots evaluation set, respectively. Moreover, this performance is comparable with the DNN *i*-vector baseline trained on a much larger dataset (86 hours v.s. 5000 hours). Experiments also exhibit the robustness of the proposed method in complicated evaluation condition.

ACKNOWLEDGEMENTS

This work was supported by the Shanghai Sailing Program No. 16YF1405300, the China NSFC projects (No. 61573241 and No. 61603252) and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
[3] N. Dehak, "Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification," Ph.D. dissertation, École de technologie supérieure, 2009.
[4] 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, 2011.
[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.
[7] Y. Qian and P. C. Woodland, "Very deep convolutional neural networks for robust speech recognition," *arXiv preprint arXiv:1610.00277*, 2016.
[8] T. Fu, Y. Qian, Y. Liu, and K. Yu, "Tandem deep features for text-dependent speaker verification." in *INTERSPEECH*, 2014, pp. 1327–1331.
[9] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.
[10] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.
[11] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 92–97.
[12] O. Novotnỳ, P. Matějka, O. Glembek, O. Plchot, F. Grézl, L. Burget *et al.*, "Analysis of the dnn-based sre systems in multi-language conditions," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 199–204.
[13] S. Dey, P. Motlicek, S. Madikeri, and M. Ferras, "Template-matching for text-dependent speaker verification," *Speech Communication*, vol. 88, pp. 96–105, 2017.
[14] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks." in *INTERSPEECH*, 2015, pp. 2440–2444.
[15] S. Madikeri, I. Himawan, P. Motlicek, and M. Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," Idiap, Tech. Rep., 2015.
[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.
[17] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.