INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Dialect Recognition Based on Unsupervised Bottleneck Features

Qian Zhang, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering
University of Texas at Dallas, Richardson, Texas, U.S.A.
{qian.zhang, john.hansen}@utdallas.edu

Abstract

Recently, bottleneck features (BNF) with an i-Vector strategy has been used for state-of-the-art language/dialect identification. However, traditional bottleneck extraction requires an additional transcribed corpus which is used for acoustic modeling. Alternatively, an unsupervised BNF extraction diagram is proposed in our study, which is derived from the traditional structure but trained with an estimated phonetic label. The proposed method is evaluated on a 4-way Chinese dialect dataset and a 5-way closely spaced Pan-Arabic corpus. Compared to a baseline i-Vector system based on acoustic features MFCCs, the proposed unsupervised BNF consistently achieves better performance across two corpora. Specifically, the EER and overall performance $C_{avg} * 100$ are improved by a relative +48% and +52%, respectively. Even under the condition with limited training data, the proposed feature still achieves up to 24% relative improvement compared to baseline, all without the need of a secondary transcribed corpus.

Index Terms: Language/Dialect recognition, unsupervised learning, bottleneck feature, phonetic label estimation.

1. Introduction

In recent decades, a number of novel techniques have been proposed with success for language/dialect identification (LID/DID) [1, 2] or speaker identification (SID) [3, 4, 5]. In particular, i-Vector is the state-of-the-art latent feature extraction method which can be applied on other speech identification tasks [6]. Along with deep neural networks (DNN) introduced into automatic speech recognition, a phonetic-aware DNN has been proposed for LID [7, 8]. Instead of Gaussian mixture model (GMM) posterior probabilities, the output posteriors of DNN senones are used for i-Vector extraction. At the front-end level, a variety of robust feature extraction strategies have been explored [9], addressing different types of channel mismatch or background noise. Subsequently, bottleneck features (BNF) [10] are proposed for LID/SID as an alternative feature which contains both acoustic and phonetic information. Specifically, BNFs with i-Vector strategy are start-of-the-art which have also been demonstrated on changeling data (i.e., NIST Language recognition evaluation (LRE) 2015) by the top 10 participants [11, 12]. Recently, stacked BNF [10] and multi-lingual BNF [13] are proposed for better acoustic modeling.

However, an additional transcribed corpus is required for traditional BNF extraction, since it is based on a well trained DNN based senone recognition system [14] with a bottleneck layer. The transcribed corpora contain limited phonetic information, since usually only English corpus are utilized for acoustic modeling. Additionally, there might be some discrepancies between the transcribed corpora and original LID corpus, such as background noise, channel mismatch, speech format mismatch, etc. Actually, phonetic-aware DNN based i-Vector as well as traditional BNF extraction, which benefits system performance dramatically, utilizes the additional transcribed corpus for phonetic alignment and acoustic modeling. Without additional transcribed data, could acoustic modeling with deep learning still benefit on LID/DID?

In our study, an unsupervised BNF is proposed for LID/DID without additional transcribed corpus. Similar to traditional BNFs, a DNN based phonetic level recognition system is trained with concatenated acoustic input features. Instead of performing an English senone alignment, a universal phonetic alignment is estimated based on the UBM posterior probabilities. Therefore, the proposed BNFs (i.e., unsupervised BNFs) are extracted without any additional resources. To demonstrate the effectiveness of the proposed method, two corpora are adopted for evaluation in our study, which includes a Chinese dialects dataset and the Pan-Arabic corpus.

This paper is organized as follows: Sec. 2 describes the traditional BNF extraction approach and the proposed BNF based on unsupervised phonetic label estimation. More information about system set-up and a brief description on the evaluation corpora are in Sec. 3. Sec. 4 illustrates the effectiveness of the proposed method through a performance analysis. Finally, conclusions are summarized in Sec. 5.

2. Bottleneck feature

2.1. Traditional bottleneck feature

Traditional BNFs are derived from DNN based ASR acoustic modeling with a bottleneck layer. In state-of-the-art ASR systems, each utterance is represented by a sequence of senones (i.e., the tied-triphone states) which are introduced for acoustic modeling. Since only word level transcription is usually provided, obtaining senone alignment is the first and fundamental step. Specifically, a Hidden Markov Model (HMM)/GMM ASR model are employed for forced alignment before subsequent DNN training. A monophone model is trained for acoustic modeling that does not include any contextual information about the preceding or following phone. Based on that, triphone models are created that represent a phoneme variant in the context of two other (left and right) phonemes. Since not all triphones occur in the training data with sufficient statistics, a phonetic decision tree with a maximum likelihood algorithm is used for generating feasible senone set[15]. In addition, a decision tree is also produced for construction of systems which have unseen triphones. Therefore, senone model training includes additional arguments for the number of HMM states on the decision tree and the number of Gaussians based on heuristics. Once the senone set is defined, a Viterbi decoding is em-

This project was funded by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

http://dx.doi.org/10.21437/Interspeech.2017-576

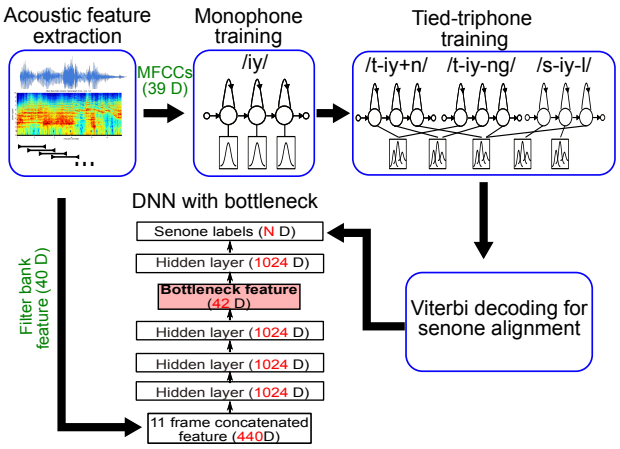

Figure 1: *Traditional bottleneck extraction diagram.(N=8973 for switchboard corpus)*

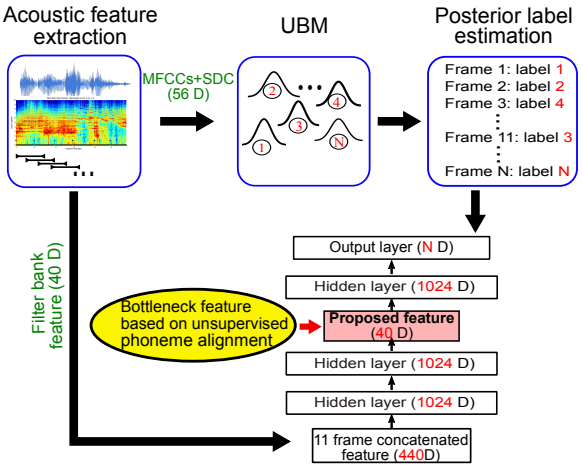

Figure 2: *Unsupervised phonetic label based bottleneck diagram.(N=1024 in our study)*

ployed to align audio data into the corresponding senones. Usually, the parameters of the acoustic model are estimated in the acoustic training steps which can be better optimized by cycling through training and alignment phases.

In recent systems, a DNN is used to estimate the senone posteriors based on acoustic features. Instead of MFCCs, the input acoustic feature is a 40 dimensional filter bank feature. The performance of acoustic modeling is still comparable if replacing one of the hidden layers with a bottleneck layer. Instead of the output layer, traditional BNFs are extracted from the middle bottleneck layer. Figure 1 represents a flow diagram for training a DNN for extracting bottleneck. There are 5 hidden layers (1024-1024-1024-42-1024) between the input layer (e.g., 11 frame concatenated acoustic features) and output layers (e.g., aligned senones label). The bottleneck layer is set to be the second last layer according to previous work [16], which shows that it contains more discriminative language information compared with other layers.

2.2. Bottleneck feature based on unsupervised phonetic label

Traditional BNFs have become popular as an alternative to MFCCs for speech tasks such as ASR, SID and LID, since they contain both acoustic and phonetic information. However, there are two potential negative impacts. Firstly, limited (e.g., English) phonetic information is considered for forced alignment during the training phase. Since there are more than one language in a regular LID task, unitary phonetic alignment is less accurate. Secondly, the additional corpus used for DNN training may cause more challenges, because of mismatch acoustics characteristics (i.e., the channel information, background noise, speech format, etc.). Our previous study shows that for the LID task on large-scale challenge corpus LRE15, bottleneck based on Switchboard improved overall system performance by a relative 10%. However, if AMI meeting corpus is used for acoustic model training and bottleneck extraction, the performance is even worse than classic MFCCs, conforming the sensitivity of this solution to acoustic mismatch.

Therefore, an unsupervised phonetic label based BNF is proposed in our study. The basic concept is similar to traditional BNF, but without extra an transcribed English corpus. Instead of forced aligned senone labels, the GMM mixture number is assumed to represent phonetic information. The diagram is shown in Figure 2. Firstly, a universal background model (UBM) is trained with all enrollment data based on MFCCs with Shifted Delta Cepstral (SDC) features. Specifically, the universal phonetic space is modelled with N Gaussian mixture components (i.e., N=1024 in our study). Subsequently, frame level phonetic label is estimated according to posterior probability. There are only 4 hidden layers (1024-1024-40-1024) between the input and output layers, because the size of LID corpus in our study is smaller than Switchboard (e.g., 28 hours vs. 250 hours)

Table 1: *Chinese dialect*

	CMN	HSN	WU	YU	Total
Train (Hrs)	6.3	8.9	5.1	7.7	28
Test (Hrs)	2.2	2.9	1.7	2.6	9.4
Avg. Dur.	10 sec				

3. System set-up and corpora

This section focuses on details concerning system development and the corpus used for evaluation. In the baseline system, 13 dimensional static acoustic MFCC features are extracted using a 25 ms analysis window with 10 ms shift. SDC features are added afterwards. In addition, voice activity detection is applied based on log mel energy. A UBM with 256 mixtures is trained on the given enrollment data, since more mixtures does not bring significant benifit on our corpora. Specifically, KALDI toolkit [17] is adopted for both acoustic feature extraction and UBM training which uses 20 iterations per mixture split. Based on the UBM, a total variability (TV) matrix is trained with same enrollment data. Finally, 600 dimensional i-Vectors are extracted for each utterance. In terms of proposed system diagram, it is similar to baseline, since the proposed feature is extracted at the frame level. Instead of MFCCs, unsupervised BNFs are utilized for UBM training and i-Vector extraction.

The corpora utilized for evaluation in this study consists of a Chinese dialect dataset [18] and the Pan-Arabic corpus.

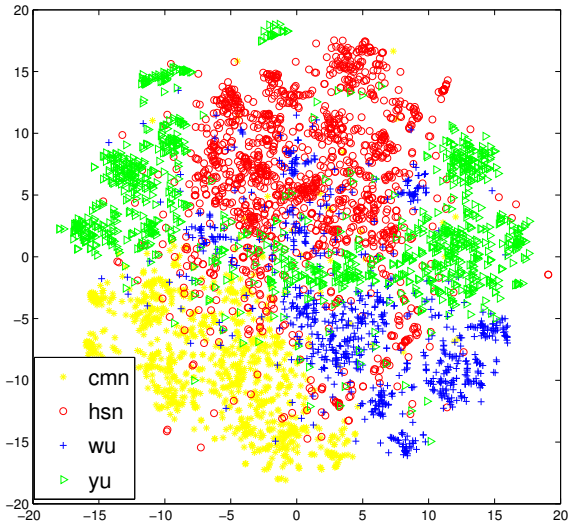

Figure 3: *Baseline system: i-vector distribution visualization (t-SNE iter=200).*

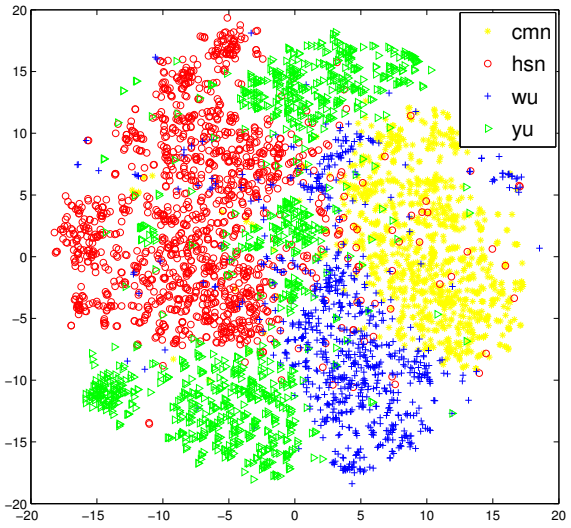

Figure 4: *Unsupervised bottleneck feature based system: i-vector distribution visualization (t-SNE iter=200).*

The Chinese corpus consists of four Chinese dialects (sub-languages): Mandarin (CMN), Cantonese (YU), Xiang (HSN), and Wu (WU). All data in this corpus is based on spontaneous conversational noise-free speech, without duration mismatch between training and test data. More detailed information about training and test data is shown in Table 1. In addition, the condition with limited training data is also considered in our study (i.e., swap the training and test dataset). Generally speaking, acoustic similarity among the target Chinese dialects is lower than that among regular dialects. For example, Mandarin speakers cannot understand Cantonese unless they are dedicated to learning this dialect. The Pan-Arabic corpus consists of Arabic dialect data from five different regions, including United Arab Emirates (AE), Egypt (EGY), Iraq (IRQ), Palestine (PS), and Syria (SY). Each dialect set captures conversations of 100 speakers (gender balanced). To be consistent with the Chinese corpus statistics, around 7 hours and 2 hours data per Arabic dialect are random selected from the original corpus to make up training and test set, respectively. Our previous study [2] shows that Pan-Arabic is more challenging than Chinese because of linguistic similarity. With respect to consistent DID system evaluation, both corpora were collected in the countries/regions with the exact same recording system/equipment. A previous study by Boril and Hansen [19] "Is the secret in the silence?" showed that consistent recording conditions are needed for DID task.

4. Result and analysis

This section focuses on analysis of LID/DID system performance with the proposed unsupervised learning methods. To evaluate experiments in terms of different perspectives, three types of measurement criteria were adopted. The first is averaged accuracy across classes. To better analyze binary classification performance for each language, a DET curve and EER are employed to illustrate more details concerning false alarms and missing rates. In addition, an evaluation of the overall classification performance using the standard NIST LRE criterion average cost function (C_{avg}) [20] is employed.

4.1. Feature visualization

To better visualize the impact of the proposed bottleneck with unsupervised phonetic label, t-Distributed Stochastic Neighbor Embedding (t-SNE)[21] is adopted. It is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. The visualized features are 600 dimensional i-Vectors which are extracted based on either acoustic feature MFCCs or proposed BNFs, since it is the finalized input feature to Gaussian classifier back-end. Here, t-SNE is set up with 200 iterations for both experiments. Fig. 3 shows the distribution of i-Vector from the baseline system. It can be noted that the four dialects are basically separated, but there are some overlap between dialect Wu and Hsn, which are very hard to be distinguished perfectly. In contrast, Fig. 4 shows the i-Vector based on the proposed BNF. There is now a clear boundary between each pair of dialects. Notably, the proposed feature is more discriminative than the acoustic feature MFCCs.

4.2. Performance analysis across corpora

In this section, the proposed BNF are evaluated on both Chinese and Pan-Arabic corpora. The consistent performance improvement demonstrates the effectiveness of proposed feature. Table 2 shows the evaluation results from three aspects.

For Chinese DID, the EER drops from 2.3 to 1.3 with a relative +48% improvement with the proposed BNF. More details about evaluation comparison in terms of detection error tradeoff (DET) curve are showed in Fig. 5 (i.e., Chinese DID under normal condition: Train with 28 h and test with 9.4 h). Meanwhile, The overall performance $C_{avg} * 100$ is deceased from 2.7 to 1.3 with a relative +52% improvement, and the average accuracy is increased from 95.5% to 97.8%. In addition, the traditional BFN based on Switchboard is also evaluated on Chinese corpus. It achieves a better performance with around 99% accuracy, since much more resources are used for acoustic modeling.

Table 2: *Unsupervised BNF impact (in %).*

		EER	C_{avg}	Accuracy
Chinese	Baseline	2.3	2.7	95.5
	Proposed BNF	**1.3**	**1.3**	**97.8**
Pan -Arabic	Baseline	12.9	16.2	72.0
	Proposed BNF	**9.2**	**12.8**	**81.3**

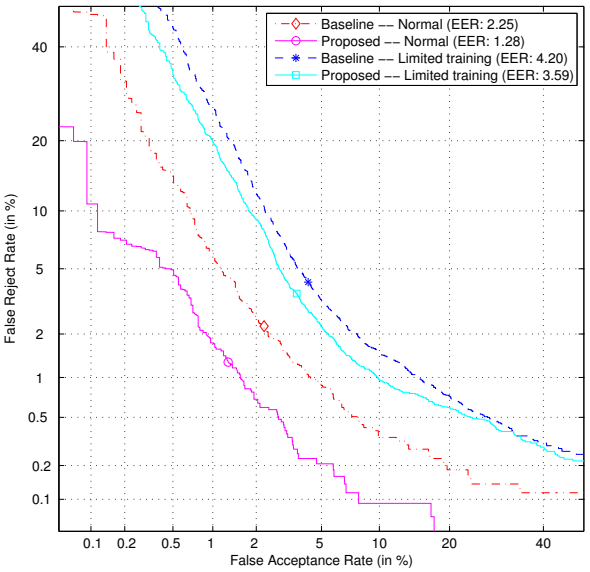

Figure 5: *Chinese LID DET curve with EER.(Normal condition: Train with 28 h and test with 9.4 h; Limited training condition: Train with 9.4 h and test with 28 h)*

Meanwhile, it can be noted that the Pan-Arabic DID performance is worse than Chinese overall, since closely space Arabic DID is more challenging than Chinese. Similar to Chinese, the proposed BNF solution achieves a 9.2% EER with a relative +29% improvement and the overall performance $C_{avg} * 100$ is deceased from 16.2 to 12.8 with a relative +21% improvement. Therefore, the proposed BNF is effective across DID corpora.

4.3. Performance with limited training data

Without loss of generality, evaluation with limited training data is also considered in our study. For simplicity, only Chinese corpus are adopted for evaluation. The new training data contains 9.4 hours speech which is the original test data, and the new test data is the original training data which is around 27 hours. According to Table 3 and Fig. 5, it can be noted that the proposed feature continues to outperform the baseline system. Specifically, EER drops from 4.2 to 3.6 with a relative +15% improvement and overall performance $C_{avg} * 100$ is deceased from 5.4 to 4.1 with a relative 24% improvement. Meanwhile, the average accuracy is increased from 91.0% to 93.1%.

It can be seen that improvement based on sufficient training is more notable than that with limited training data. Since a well trained senone DNN usually requires hundreds or thousands hours of training data, limited training data (i.e., 9.4 hours in total) definitely impacts the acoustic modeling with phonetic alignment. However, proposed unsupervised bottleneck still shows the benefits in system performance even with very limited training data.

Table 3: *Limited training on Chinese DID (in %; Limited training condition: Train with 9.4 h and test with 28 h)*

	EER	C_{avg}	Accuracy
Baseline	4.2	5.4	91.0
proposed BNF	**3.6**	**4.1**	**93.1**

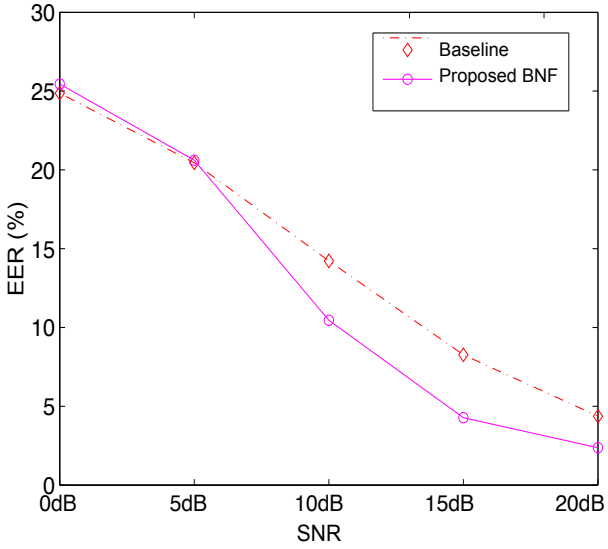

Figure 6: *Noise robustness analysis on Chinese (EER %).*

4.4. Noise robustness analysis

The unsupervised BNF have already been shown to be effective on clean data on two corpora. However, noisy conditions are a challenge in real scenarios. To simulate noisy speech signals, speech-shape noise (SSN) is employed in our study whose long-term average spectrum is similar to that of speech. In order to evaluate the robustness of our proposed methods under controlled noisy conditions, the performance with different speech-to-noise-ratios (SNR) is evaluated on Chinese DID and illustrated in Fig. 6. It can be noted that the propose BNF outperforms the baseline MFCCs on noisy speech with 10dB to 20dB SNR. In particular, it achieves a relative +27% improvement compared to baseline at the 10dB SNR. However, the proposed BNF does not bring any improvement on severely noisy condition, such as 0dB to 5dB.

5. Conclusion

Bottleneck features (BNF) with i-Vector strategy has been widely used for language/dialect identification (LID/DID). Since traditional bottleneck extraction requires an additional transcribed corpus, an unsupervised BNF extraction diagram was proposed in this study. The proposed method was evaluated on a 4-way Chinese dialect dataset and a 5-way Pan-Arabic corpus. Compared to a baseline i-Vector with MFCC feature system, the proposed feature achieves sustained performance gain across corpora. Specifically for Chinese DID, the EER drops from 2.3 to 1.3 with a relative +48% improvement and the overall performance $C_{avg} * 100$ is deceased from 2.7 to 1.3 with a relative +52% improvement. Even under the condition with limited training data, the EER and overall performance $C_{avg} * 100$ are still improved by relatively +15% and +24%, respectively. In addition, the proposed feature are demonstrated effective under noisy condition.

6. References

[1] D Martinez, Lukás Burget, Luciana Ferrer, and Nicolas Scheffer, "iVector-based prosodic system for language identification," in *Proc. ICASSP, Kyoto, Japan*, Mar. 2012, pp. 4861–4864.

[2] Qian Zhang, Hynek Boril, and John H L Hansen, "Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification," in *Proc. ICASSP, Vancouver, Canada*, May 2013, pp. 7363–7367.

[3] Leena Mary and Bayya Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech communication*, vol. 50, no. 10, pp. 782–796, 2008.

[4] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. ICASSP, Florence, Italy*, May 2014, pp. 4047–4051.

[5] Gang Liu and John HL Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1978–1992, 2014.

[6] Maryam Najafian, Saeid Safavi, Philip Weber, and Martin Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems," in *ODYSSEY*, 2016, pp. 213–218.

[7] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP, Florence, Italy*, May 2014, pp. 1695–1699.

[8] Patrick Kenny, Vishwa Gupta, Themos Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey: Speaker and Language Recognition Workshop, Joensuu, Finland*, Jun. 2014.

[9] Qian Zhang, Gang Liu, and John H L Hansen, "Robust langauge recognition based on diverse feature," in *Proc. Odyssey: Speaker and Language Recognition Workshop, Joensuu, Finland*, Jun. 2014.

[10] Pavel Matejka, Le Zhang, Tim Ng, Sri Harish Mallidi, Ondrej Glembek, Jeff Ma, and Bing Zhang, "Neural network bottleneck features for language identification," pp. 299–304, Jun. 2014.

[11] Kong Aik Lee, Ville Hautamäki, Anthony Larcher, Wei Rao, Hanwu Sun, Trung Hieu Nguyen, Guangsen Wang, Aleksandr Sizov, Ivan Kukanov, Amir Poorjam, et al., "Fantastic 4 system for nist 2015 language recognition evaluation," *arXiv preprint arXiv:1602.01929*, 2016.

[12] Fred Richardson, Douglas Reynolds, and Najim Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[13] Radek Fér, Pavel Matějka, František Grézl, Oldřich Plchot, and Jan Černockỳ, "Multilingual bottleneck features for language recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[15] Steve J Young, Julian J Odell, and Philip C Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[16] Mitchell McLaren, Luciana Ferrer, and Aaron Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition," in *Proc. ICASSP, Shanghai, China*, Mar. 2016.

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *in Proc. ASRU: IEEE workshop on automatic speech recognition and understanding, Waikoloa, USA*, Dec. 2011.

[18] Yun Lei and John HL Hansen, "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 85–96, 2011.

[19] Hynek Bořil, Abhijeet Sangwan, and John H. L. Hansen, "Arabic dialect identification - 'Is the secret in the silence?' and other observations," in *INTERSPEECH 2012*, Portland, Oregon, September 2012, pp. 30–33.

[20] Alvin Martin and Craig Greenberg, "The 2009 NIST language recognition evaluation," in *Proc. Odyssey: Speaker and Language Recognition Workshop, Brno, Czech Republic*, Jun. 2010, pp. 165–171.

[21] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.