SEMI-SUPERVISED LEARNING WITH GENERATIVE ADVERSARIAL NETWORKS FOR ARABIC DIALECT IDENTIFICATION

Chunlei Zhang, Qian Zhang [§]*, John H.L. Hansen*[♠]

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, Texas, USA

{chunlei.zhang,john.hansen}@utdallas.edu

ABSTRACT

Dialect Identification (DID) refers to the process of identifying different dialects within the same language class. Compared with more general language identification (LID), DID is a more challenging task because of the substantial similarity between dialects. For an i-vector based LID/DID, prior studies have shown advancements with deep neural networks (DNNs) over Gaussian Mixture Models (GMMs) in acoustic modeling. In this study, a novel i-vector representation which is based on unsupervised bottleneck features is examined as the feature to identify dialects from Arabic broadcast speech. To utilize the unlabeled training data, semi-supervised learning with generative adversarial networks (GANs) are incorporated in the back-end classifier development. Experiments with the proposed method in the third release version of the Multi-Genre Broadcast (MGB-3) Challenge yields the best single system performance among all submitted systems. An overall classification accuracy of 73.8% achieves a +28.8% relative improvement over the MGB-3 baseline with an accuracy of 57.3%, which is the state-of-the-art performance in this DID task. The fused system further achieves an improvement of +39.4% in accuracy.

Index Terms— Semi-supervised learning, language identification, generative adversarial networks, i-vector, Arabic Dialect identification.

1. INTRODUCTION

As a special case of Language Identification (LID), Dialect Identification (DID) requires us to identify different dialects within the same language class. Because of the remarkable similarity between dialects, DID is considered to be a more challenging problem than general LID. In [1], 100% accuracy on an Arabic/English language identification task is achieved using an i-vector framework [2], while only a 59.2% accuracy is reported in Arabic dialect identification (ADI) task from the same system. This indicates that DID remains as a difficult problem even when recent advancements have been reported from current LID systems [3, 4, 5, 6].

Modern LID systems can be categorized into two general classes, (i.e., text/lexical feature based LID, and acoustic feature based LID [1, 7, 8]). Acoustic based LID has drawn more attention with its flexibility, relatively better performance and less demanding annotations. As a contrast, text/lexical based LID systems often rely on parallel transcriptions, which are difficult to satisfy in practical LID/DID applications.

As one competitive representative of the acoustic based systems, the i-vector framework provides a way to map an arbitrary length utterance into a fixed length vector, where information (e.g., channel, speaker, language etc) from the utterance is encoded [2, 9]. Similar to that for speaker recognition, replacing Gaussian Mixture models (GMMs) with supervised deep neural networks in the acoustic modeling has been shown to be effective in LID [5]. Richardson et al. found that bottleneck features (BNF) from an ASR DNN acoustic model combined with GMM posteriors achieved the best performance in the LRE11 post-evaluations [5]. The result confirms the observation that the advantages of ASR DNN posteriors have been largely constrained to English language speech where a large amount of text information can be utilized [10]. To relax the limitation of unsatisfying text for non-English speech, Zhang et al. [11, 12] proposed to perform acoustic partitioning with GMMs and assign phonetic labels according to the GMM posteriors. With this operation, an unsupervised "phonetic alignment" is incorporated into the BNF deep neural networks, with an impressive gain in performance, for which an $unsupervised$ BNF is obtained for the DID task.

For i-vector based LID/DID, a secondary classifier is required. Gaussian back-end (GB) and Support Vector Machine (SVM) solutions are reported to have state-of-the-art performance in several different tasks [3, 5, 13]. Deep neural network based classifiers are also effective [14]. In the case of unlabeled training/development data, semi-supervised learning with Ladder Networks has been successfully applied to language recognition [15]. In our study, we develop a classifier based on a semi-supervised generative adversarial networks (GANs) framework, which takes advantage of unlabeled training/development data.

In addition to the advancements based on the i-vector modeling, studies in DNN based embedding/metric learning from speech or end-to-end systems have also draw increasing attentions. State-of-the-art performance has been reported in speaker recognition, language recognition and even DID tasks [16, 17, 18, 19, 20]. However, those methods are generally considered as (labeled) data demanding, which is difficult to be satisfied for low resource language cases.

In this paper, we present a systematic study for DID task with unsupervised/semi-supervised learning concepts in both front-end i-vector modeling and back-end classifier sides. The primary contribution of this study is integrating recent advancements in front-end features and back-end classifiers for a challenging DID task. An overview of our recent proposed unsupervised bottleneck features (UBNF) for i-vector extraction is presented in Sec. 2. A new semi-supervised learning framework with GANs for identifying Arabic dialects is detailed in Sec. 3 [21]. Experiments with the proposed methods are examined using the MGB-3 Arabic dialect identification corpus in Sec. 4. Finally, we conclude our work in Sec. 5.

[§] Job was done when Qian Zhang was with CRSS, UT Dallas, now she is with Google Inc.

[♠] This project was funded in part 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.

2. UBNF FOR I-VECTOR MODELING

This section describes i-vector modeling with our UBNF features. The motivation of replacing traditional MFCC features with the alternative BNF is to employ the impressive phonetic discriminative ability with DNNs [1]. In the case of LID/DID, training a supervised ASR DNN acoustic model for all languages is difficult (e.g., absence of sufficient parallel text transcriptions). To overcome this, a novel *unsupervised* version of BNF is introduced in this study.

2.1. UBNF paradigm

The concept of UBNF is similar to traditional bottleneck features, but without the need for text information. In the UBNF paradigm, a universal background model (UBM) is trained to align acoustic features with mixture indexes, which are assumed to represent phonetic-like information. As shown in Fig.1, a UBM is trained using MFCCs with Shifted Delta cepstral (SDC) features in the training set. Similar to [11], the universal phonetic space is partitioned into N mixtures. Subsequently, frame level phonetic labels are estimated according to highest GMM posteriors. With the estimated phonetic labels, a deep bottleneck network is trained on the filter-bank features. The DNN architecture is illustrated in Fig.1, a 40-D normalized activation vector is extracted from the bottleneck layer as the replacement for MFCC-SDC features in the conventional GMM/i-vector framework [2, 9].

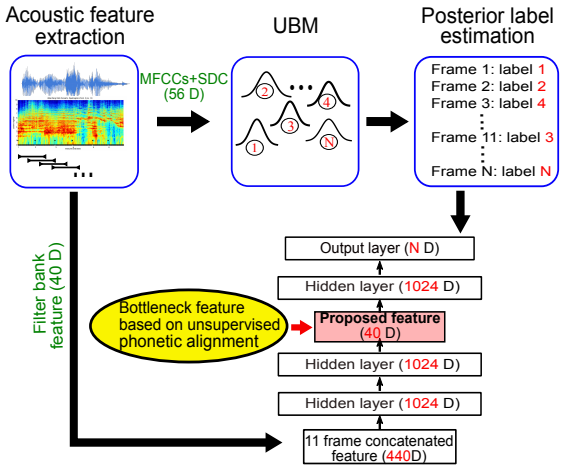

Fig. 1. The UBNF feature extraction diagram. *Sigmoid* nonlinearity is used with softmax normalization for the output layer of DNN.

2.2. i-vector modeling with UBNF

To estimate the i-vector for a given speech utterance, the Baum-Welch statistics are needed and obtained by:

$$N_k = \sum_t p(k|\mathbf{x_t}) \\ F_k = \sum_t p(k|\mathbf{x_t})\mathbf{x_t}, \tag{1}$$

where $\mathbf{x_t}$ is t-th frame of the utterance, $p(k|\mathbf{x_t})$ corresponds to the posterior probability of Gaussian mixture k generating the vector $\mathbf{x_t}$. Here, $\mathbf{x_t}$ is the proposed UBNF that goes through the standard i-vector modeling process, which includes UBM training, UBM posterior calculation, Baum-Welch statistics extraction, total variability matrix training etc. The generative model for the i-vector can be expressed as:

$$M = m + Tx, x \sim N(0, I) \tag{2}$$

where M is a supervector constructed by appending together the first order statistics for each mixture component k, m is the universal supervector usually concatenated from UBM means, T is a row rank total variability matrix derived from zero and first-order Baum-Welch statistics from Equ.(1), x is the i-vector which retains most of the high-level information of the utterance. In this study, i-vector is the feature for DID back-end classifier development.

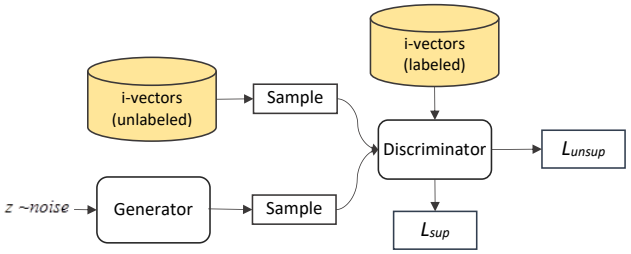

Fig. 2. A conceptual semi-supervised learning framework with GANs. The "feature matching" trick is also employed to construct the generator loss, as proposed in [22].

3. SEMI-SUPERVISED LEARNING WITH GANS

We incorporate the concept of semi-supervised adversarial training for dialect identification, given the unlabeled data for the MGB-3 ADI challenge.

3.1. Semi-supervised learning with Generative Adversarial Networks (GANs)

In the original GANs design, there is a generative network $G(z; \theta_G)$ that produces samples from noise to fool a discriminator network $D(\mathbf{x})$, where the discriminator tries to identify the generated samples. By jointly optimize both G and D, a powerful unsupervised generative model is learned that can produce samples close to the real data [21].

In this study, the GANs discriminator network is not simply for detecting whether the sample is real or "generated". We wish to ensure that the discriminator network is also a classifier that distinguishes features from different dialects, like a multi-task within GANs framework [23, 22]. To do so, we add the samples from the generator G to the real K classes data set, labeling them with a new "generated" class $y = K + 1$, so that the classifier is expanded to $K + 1$ classes. The loss function for training our classifier becomes:

$$L = -\mathbb{E}_{\mathbf{x},y\sim p_d(\mathbf{x},y)}[\log p_m(y|\mathbf{x})] \\ -\mathbb{E}_{\mathbf{x}\sim G}[\log p_m(y = K + 1|\mathbf{x})], \tag{3}$$

where $\mathbf{x}$ is the i-vector, y is the corresponding label, $p_d(\mathbf{x}, y)$ is the real data distribution, and $p_m(\mathbf{x}, y)$ is the distribution modeled by the discriminator. We formulate two losses from the cross-entropy loss of the classifier:

$$L_{sup} = -\mathbb{E}_{\mathbf{x},y\sim p_d(\mathbf{x},y)}[\log p_m(y|\mathbf{x}, y < K + 1)], \tag{4}$$

$$L_{unsup} = -\{\mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x})} \log[1 - p_m(y = K + 1|\mathbf{x}) \\ + \mathbb{E}_{\mathbf{x}\sim G} \log[p_m(y = K + 1|\mathbf{x})]\}, \tag{5}$$

substituting $D(\mathbf{x}) = 1 - p_m(y = K + 1|\mathbf{x})$ into Equ.(5), we reach Equ.(6) which is a standard GAN game-value.

$$L_{unsup} = -\{\mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{\mathbf{x}\sim G} \log(1 - D(\mathbf{x})\}, \quad (6)$$

Thus, the discriminator D is optimized with all labeled data using the supervised loss L_{sup}; at the same time, the discriminator also maximizes the probability of the real data to K real data classes with the unsupervised GANs objective L_{unsup}. In this way, unlabeled data is utilized with a semi-supervised framework. A conceptual semi-supervised learning framework with GANs is depicted as Fig. 2.

4. EXPERIMENTS

In this section, we first introduce MGB-3 ADI Challenge corpus, and then briefly describe the baseline systems. Finally, the experimental setup and results are detailed and analyzed here, showing advancements of our proposed method for i-vector based ADI task.

4.1. MGB-3 Arabic Dialect Identification Corpus

The ADI training dataset was collected from the Broadcast News domain from 5 Arabic dialects inclusing Egyptian (EGY), North African or Maghrebi (NOR), Gulf or Arabian Peninsula (GLF), Levantine (LAV), and Modern Standard Arabic (MSA). Data recordings were digitized at 16Khz. The recordings were segmented into utterances with random durations. The test and the development datasets came from the same broadcast domain, but the recording setup was different from the training data. Text information (transcription from a ASR system [1, 24]) is also provided. Since our system is based on acoustic features, we remove the text part to avoid redundancy. The corpus statistics are listed in Table 1. It is noted that unlabeled MBG-2 data is also provided for system development, which is useful in UBM, T-matrix as well as semi-supervised GANs classifier training.

Table 1. Statistics of MGB-3 ADI Corpus. **Utt #** stands for number of utterances, **Dur.** stands for duration in hours.

	Training		**Development**		**Eval**	
Dialect	Utt #	Dur.	Utt #	Dur.	Utt #	Dur.
EGY	3093	12.4	298	2.0	302	2.0
GLF	2744	10.0	264	2.0	250	2.1
LAV	2851	10.3	330	2.0	334	2.0
MSA	2183	10.4	281	2.0	262	1.9
NOR	2954	10.5	351	2.0	344	2.1
Total	13825	53.6	1524	10.0	1492	10.1

4.2. ADI Baselines

Two ADI systems from the MGB-3 challenge organizers are provided as baselines. An additional one MFCC-SDC/i-vector + GB system is also developed by here for cross-system comparisons.

4.2.1. MGB-3 baselines

One lexical and one BNF/i-vector system are provided by the ADI challenge organizers [24]; both text and acoustic features are followed by a multi-class SVM classifier to find the probabilities of Arabic dialects. We refer to these systems as $\text{I}_{\text{bs-t}}$ and $\text{II}_{\text{bs-a}}$, respectively.

4.2.2. MFCC-SDC/i-vector + GB

The third baseline is a standard UBM/i-vector system, with 56-D MFCC-SDC (7 static cepstra appended to 49 shifted delta cepstra), energy based speech activity detection and a 1024-mixture UBM (with both MGB-2 and MGB-3 training data). The dimension of i-vector is 600, with a Gaussian back-end implemented to identify each dialect. The system is labeled with $\text{III}_{\text{bs-gb}}$.

4.3. Hyper-parameter setup for UBNF/i-vector+ GANs system

4.3.1. UBNF/i-vector

As illustrated in Fig.1, the same 56-D MFCC-SDC features with an energy based SAD is provided for unsupervised acoustic partitioning; a 2048-mixture GMM is only trained using MGB-3 training set. The dimension of output layer of the deep BNF network is also 2048. The network architecture is summarized as 440-1024-1024-40-1024-2048, where the input layer is the concatenation of 11-frame filter-bank features. 40-D BNF features extracted from MGB-2 and MGB-3 training set are the input of standard GMM/i-vector pipeline, and 600-D i-vector used for the final feature for back-end classifiers.

4.3.2. Semi-supervised GANs

The Semi-supervised GANs architecture is shown in Table 2. The generator G is a feed-forward dense network with two hidden layers. A Gaussian noise vector z is the input to G, and the output is the "generated i-vector". The discriminator D is also the supervised classifier for ADI. It has three hidden layers, where 50% of each hidden layer parameters is randomly dropped out to address overfitting problem [25].

In our submissions to MGB-3 ADI Challenge, duration information is also appended to the 600-D i-vector space as auxiliary features for classifier training. The intuition behind this operation is to add uncertainty to the estimated i-vector before back-end modeling, a performance gain is expected with this simple duration calibration [3]. Our experiments on the MGB-3 ADI Dev set confirms a relative $\sim$3% accuracy improvement achieved for different front-end features. Therefore, a 601-D i-vector+duration solution is used for all our ADI systems.

Table 2. Semi-supervised GANs architecture for i-vector features.

Generator architecture			**Discriminator architecture**		
Input	2×hidden	Output	Input	3×(hidden & dropout)	Output
100	500	601	601	1024 & 0.5	5

4.4. Performance on the MGB-3 ADI Dev set

Throughout the experiments on the ADI Dev and Eval sets, three metrics (e.g., Accuracy (ACC), Recall (RCL), and Precision (PRC)) are used to report system performance (note: this is the same as [1, 24]). The single system performance is presented in this section. In addition to three baselines, the UBNF/i-vector + GB and UBNF/i-vector + GANs proposed solutions are also reported in Table 3. These two systems are noted as $\text{IV}_{\text{u-gb}}$ and $\text{IV}_{\text{u-gans}}$, respectively.

As shown in from Table 3, the baseline $\text{III}_{\text{bs-gb}}$[1] with MFCC-SDC/i-vector and a Gaussian back-end outperforms two MGB-

[1]Our MFCC-SDC/i-vector front-end feature outperforms $\text{II}_{\text{bs-a}}$ with BNF front-end because of MGB-2 data augmentation in the UBM training.

Table 3. Dev results of 3 baseline and 2 proposed systems.

System	ACC	RCL	PRC
$I_{\text{bs-t}}$	48.26%	50.33%	49.13%
$II_{\text{bs-a}}$	58.09%	61.37%	58.83%
$III_{\text{bs-gb}}$	63.37%	63.82%	64.29%
$IV_{\text{u-gb}}$	65.88%	66.49%	65.31%
$IV_{\text{u-gans}}$	**69.42%**	**69.83%**	**69.02%**

3 ADI challenge baselines. Compared with other features, the UBNF/i-vector generally has a better characterization for the ADI task when we fix other conditions (i.e., the same Gaussian back-end is applied on $III_{\text{bs-gb}}$ and $IV_{\text{u-gb}}$). The GANs based semi-supervised learning outperforms all the systems by a substantial margin, which shows the advantages of our proposed method.

4.5. Performance on the MGB-3 ADI Eval set

4.5.1. Performance with only training set

We apply classifiers which are trained with only training data for the Eval set. Table 4 lists performance of the 5 single systems. In terms of ACC, overall performance is slightly lower than the Dev set. This maybe attribute to relatively more short utterances in the Eval set, which is also a motivation to perform duration calibration in the i-vector space described above. The Eval set result does confirm that the UBNF feature and Semi-supervised GANs classifier have consistent advantages over traditional methods.

Table 4. Eval results of 3 baseline and 2 proposed systems.

System	ACC	RCL	PRC
$I_{\text{bs-t}}$	47.64%	48.33%	47.23%
$II_{\text{bs-a}}$	57.33%	59.57%	58.83%
$III_{\text{bs-gb}}$	63.23%	63.62%	63.91%
$IV_{\text{u-gb}}$	65.45%	66.37%	66.89%
$IV_{\text{u-gans}}$	**69.16%**	**70.27%**	**69.76%**

4.5.2. Data augmentation

Although the training, Dev and Eval sets are from the same broadcast domain, there is still differences between training and the Dev/Eval set. Here, we note that the Dev and Eval sets are more similar to each other. So, introducing Dev set data for classifier training will benefit more on Eval set. In this experiment, we randomly introduce 2/3's of the Dev data (around 1000 utterances, we leave the other 1/3 of Dev data for score level fusion) to the training set. The performance after Dev data augmentation is detailed in Table 5. We see a $\sim 6\%$ relative improvement achieved by simple data augmentation, which shows the importance of in-domain data for the ADI task. An impressive accuracy of 73.86% is achieved with only a single system $IV_{\text{u-gans}}$, which outperforms most of the reported fusion systems to the MGB-3 ADI Challenge [24, 26].

4.5.3. System fusion

In order to predict final scores combining our multiple single systems. We build a fused model by training a logistic regression model for fusion. Let $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ be the score features by concatenating each single system output. In the logistic regression model, the target binary variable y is a Bernoulli random variable of which the probability of occurrence is dependent on the prediction

Table 5. Eval results of 3 baseline and 2 proposed systems–data augmented.

System	ACC	RCL	PRC
$I_{\text{bs-t}}$	52.61%	53.63%	52.23%
$II_{\text{bs-a}}$	59.78%	62.07%	60.73%
$III_{\text{bs-gb}}$	65.81%	66.62%	66.91%
$IV_{\text{u-gb}}$	69.64%	70.12%	69.89%
$IV_{\text{u-gans}}$	**73.86%**	**74.67%**	**73.52%**

given in Equation 7. Regression coefficients $\boldsymbol{\omega}$ are estimated using the maximum likelihood estimation. Scores from each single system are combined with the estimated coefficients to get the fusion score $\hat{y}$.

$$p(y = 1|\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{1 + \exp\left(-\boldsymbol{\omega}^T \mathbf{x}\right)} \tag{7}$$

$$\hat{y} = \boldsymbol{\omega}^T \mathbf{x} \tag{8}$$

The system fusion together with the latest updates on this task is reported in Table 6. Among all results, the "UBNF/i-vector+GANs" single system achieves slightly better results compared with current state-of-the-art single system solution [20], which again shows the advancement of UBNF feature for capturing accurate acoustic unit and semi-supervised GANs based classifier for utilizing unlabeled data. Finally, our fused system gives 79.86% accuracy on ADI task with a relative improvement over the MGB-3 ADI baseline by +39.4%. The overall performance is 1.5% worse than the recent mega fusion system [20], where more systems are utilized in [20].

Table 6. Comparison with the recent results on MGB-3 ADI task.

System	ACC	RCL	PRC
Shon et.al (fusion) [20]	**81.36 %**	/	/
Our system fusion	**79.86%**	80.27%	79.87%
Shon et.al (fusion) [27]	75.0%	75.1%	75.5%
$IV_{\text{a-gans}}$-single	**73.86%**	74.67%	73.52%
Shon et.al [20]	73.39 %	/	/

5. CONCLUSIONS

This study described the development of a UBNF i-vector system and a semi-supervised GANs classifier and demonstrated substantial performance gains when the system is applied to the MGB-3 ADI Challenge. The main focus of this study is to develop a single system which can incorporate both discriminative and generative abilities from deep neural networks. GANs based semi-supervised learning has been employed as a back-end classifier for a DID task. The "UBNF/i-vector+GANs" single system improves performance over the MGB-3 ADI baseline with a relative +28.8% accuracy gain, which is state-of-the-art performance reported. Further more, system fusion further boosts the performance to a $\sim 80\%$ for ADI task.

Also, we believe this newly proposed semi-supervised learning framework is promising for different tasks. For example, it can be applied to the NIST 2015 LRE i-vector Machine Learning Challenge, where unlabeled data is the main point to address; It can be introduced to ASR acoustic modeling, where unsupervised domain adaptation (without transcripts) can be implemented using this framework. The study therefore highlights effective methods to advance i-vector based language/dialect identification, as well as fundamental observations for future speech and language technologies.

6. REFERENCES

[1] Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals, "Automatic dialect detection in arabic broadcast speech," *arXiv preprint arXiv:1509.06928*, 2015.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[3] Chengzhu Yu, Chunlei Zhang, Shivesh Ranjan, Qian Zhang, Abhinav Misra, Finnian Kelly, and John H. L. Hansen, "Utd-crss system for the nist 2015 language recognition i-vector machine learning challenge," in *IEEE ICASSP*, 2016, pp. 5835–5839.

[4] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno, "Automatic language identification using deep neural networks," in *IEEE ICASSP*, 2014, pp. 5337–5341.

[5] 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.

[6] Abhinav Misra, Qian Zhang, Finnian Kelly, and John H. L Hansen, "Between-class covariance correction for linear discriminant analysis in language recognition," in *ISCA Odyssey*, 2016.

[7] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu, "Language identification: A tutorial," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82–108, 2011.

[8] Douglas A Reynolds, William M Campbell, Wade Shen, and Elliot Singer, "Automatic language recognition via spectral and token based approaches," in *Springer Handbook of Speech Processing*, pp. 811–824. Springer, 2008.

[9] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *ISCA INTERSPEECH*, 2011.

[10] C. Zhang, F. Bahmaninezhad, S. Ranjan, C. Yu, N. Shokouhi, and J. H. L. Hansen, "UTD-CRSS systems for 2016 NIST speaker recognition evaluation," in *ISCA INTERSPEECH17*, 2017.

[11] Qian Zhang and John H. L. Hansen, "Dialect recognition based on unsupervised bottleneck features," in *ISCA INTERSPEECH*, 2017.

[12] Qian Zhang and John HL Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 5, pp. 873–882, 2018.

[13] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. L. Hansen, "Joint information from nonlinear and linear features for spoofing detection: an i-Vector/DNN based approach," in *IEEE ICASSP*, 2016, pp. 5035–5039.

[14] Shivesh Ranjan, Chengzhu Yu, Chunlei Zhang, Finnian Kelly, and John H. L. Hansen, "Language recognition using deep neural networks with very limited training data," in *IEEE ICASSP*, 2016, pp. 5830–5834.

[15] Ehud Ben-Reuven and Jacob Goldberger, "A semisupervised approach for language identification based on ladder networks," in *ISCA Odyssey*, 2016, pp. 319–325.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018.

[17] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *ISCA INTERSPEECH17*, 2017.

[18] C. Zhang, K. Koishida, and J. H.L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[19] Alan Mccree, David Snyder, Greg Sell, and Daniel Garcia-Romero, "Language recognition for telephone and video speech: The jhu hltcoe submission for nist lre17," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 68–73.

[20] Suwon Shon, Ahmed Ali, and James Glass, "Convolutional neural network and language embeddings for end-to-end dialect recognition," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 98–104.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[23] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *ICML*, 2008, pp. 160–167.

[24] Ahmed Ali, Stephan Vogel, and Steve Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *IEEE ASRU*.

[25] N. Srivastava, G. E Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] Bulut Ahmet, Zhang Qian, Zhang Chunlei, Bahmaninezhad Fahimeh, and John H. L. Hansen, "UTD-CRSS submission for MGB-3 arabic dialect identification: Front-end and back-end advancements on broadcast speech," in *IEEE ASRU*, 2017.

[27] Suwon Shon, Ahmed Ali, and James Glass, "Mit-qcri arabic dialect identification system for the 2017 multi-genre broadcast challenge," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 374–380.