A review on Deep Learning approaches in Speaker Identification

Sreenivas Sremath Tirumala[1], Seyed Reza Shahamiri[2]
Faculty of Business and Information Technology
Manukau Institute of Technology
Auckland, New Zealand
[1]ssremath@aut.ac.nz, [2]admin@rezanet.com

ABSTRACT
Deep learning (DL) is becoming an increasingly interesting and powerful machine learning method with successful applications in many domains, such as natural language processing, image recognition, hand-written character recognition, and computer vision. Despite of its eminent success, limitations of traditional learning approach may still prevent deep learning from achieving a wide range of realistic learning tasks. DL approaches has shown success in speech recognition and speaker identification over traditional approaches such as those that use Mel Frequency Cepstrum Coefficients for feature extraction with Gaussian Mixture Models. However, speaker identification research community are not fully aware of the DL process and its application with respect to speaker identification. This paper is motivated to reduce this knowledge gap and to promote the research of implementing deep learning techniques for speaker identification. In this paper, we present a review of the DL methodologies used for speaker identification and surveys important DL algorithms that can potentially be explored for future works. We categorised the applications of DL for speaker identification according to the process of speaker identification and presented a review of these implementations.

CCS Concepts
•Computing methodologies→ Machine learning approaches;Neural networks;

Keywords
Deep learning, speaker identification, feature extraction

1. INTRODUCTION
Deep Learning is a popular topic with high research interests due to its extensive application in natural language processing, image recognition and computer vision. Big corporates like Google, Microsoft, Apple, Facebook, Yahoo etc. established their deep learning research groups for implementing this concept in their products. Deep learning approaches were successful in machine learning competitions which were earlier dominated by other machine learning approaches like Support Vector Machines and typical Artificial Neural Networks (ANNs). The main problem with implementation of ANNs was its training mechanism with multiple layers (more than one hidden layer) using Back Propagation (BP) algorithms. In particular, as the ANN gets deeper with multiple hidden layers, the training process turned to be very time consuming due to the error propagation mechanism of BP algorithm [1]. To overcome this, a layer greedy-wise training is used to train multiple hidden layered neural networks termed as Deep Neural Networks. Deep Neural Networks (DNNs) use multiple layers (at least with 3 hidden layers) with a new layer-wise training mechanism.

This training mechanism of DNN is called 'deep learning' (DL). With DL, each layer learns from previous layer (unsupervised), i.e. the output of one layer acts as an input to the next layer considering each layer as an individual one layered ANN. This process is continued for all the layers. Typically, the layer-wise training is un-supervised followed by overall supervised training. For instance, few Auto-encoder networks can be stacked together to form a deep architecture where each auto-encoder will be similar to a layer of DNN. This principle is also followed for other types of ML paradigms like SVMs. These types of implementations that does not use a DNN is termed as unconventional deep learning paradigms. However, DNN is considered to be most suitable for DL algorithms [2].

Speaker Recognition (SR) is the process of recognizing the words or statements uttered. Automating this process, usually with AI techniques, is called Automatic Speaker Recognition or ASR. In terms of Machine Learning (ML), SR is considered as a pattern recognition problem. Speaker Identification (SID) is another NLP technique similar to SR but different objective: SID's objective is to identify the speaker based on her voice prints by comparing the voice profile of the speaker against existing profiles of various speakers[3].SID systems have various applications like user authorization (voiced password), personalized assistant, automatic mail direction, etc.

The process of SID involves extracting and identifying unique characteristics of speech features from a group or set of speakers; hence, it is important to select the most efficient feature extraction approaches that best represent the speech features. One of the most complex aspects of feature extraction for SR is when the input utterances are infected with noise [4]. With layer-wise training, each layer of DNN extracts features at different levels (hierarchically). Deep architecture is a hierarchical structure of multiple layers with each layer being self-trained to learn from the output of its preceding layer.

DL algorithms were applied for hierarchical feature extraction to overcome this complexity [5] proving their effectiveness in improving SR performance [6-8]. DL has been successful in various applications involving feature extraction for analysis and comparison [9-13]. The importance of feature extraction is well

ICSPS2016, November 21 to 24, 2016, Auckland, New Zealand
©2016 ACM ISBN: 978-1-4503-4790-7/16/11…$15.00
DOI: http://dx.doi.org/10.1145/3015166.3015210

defined and implemented with DNNs. This makes it important to understand the various implementations of DL for SID.

In this paper, we present a review on various approaches for SID using DL. We categorised the implementations of DLs based on the SID implementations and processes. Furthermore, we presented details of major DL implementations which attained state of art results. This paper will introduce these approaches to SR community to remove the knowledge gap and promote the research of implementing deep learning techniques.

The paper is organized as follows. A brief introduction to deep learning is presented in section 2followed by section 3with detailed process of speaker identification along with explaining various phases of SID. Section 4introduces categories of DL approaches for SID along with various implementations which attained state of art results. Finally, conclusion and future work is presented as section 5.

2. DEEP LEARNING

A simple form of deep architecture implementation is a multilayer feed forward neural network with many hidden layers (typically 5 to 7) that is also termed as Deep Neural Network or DNN.[14]. However, the training mechanism in DNN is performed layer-wise using gradient descent. This layer-wise training enables DNN to learn the 'deep representations' that transform from one hidden layer to another hidden layers. Typically, the layer-wise training is unsupervised. The weights in the DNNs are updated using stochastic gradient descent as defined below

$$\Delta w_{ij}(t+1) = \Delta w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \quad (1)$$

where η represents the learning rate, C is the cost function associated with the weights and w_{ij}represent weight. For larger training sets, DNNs may be trained in multiple batches of small sizes without losing the efficiency [15]. However, it is very complex to train DNNs with many layers and many hidden units since the number of parameters to be optimized are very high.

The layer-wise training is sufficient to learn the representation. However, a fine-tuning is performed for the entire DNN with BP.This entire training process is termed as Deep Learning or DL. DNNs are trained with BP by discriminative probabilistic models that calculate the difference between target outputs and actual output.

There are three major implementations for DL namely convolutional neural networks (CNN or ConvNets), deep belief networks (DBN), and stacked auto-encoders (SAE). ConvNets are a special type of feed-forward ANNs that can be used to perform feature extractions by applying convolution and sub sampling at each layer. The principle application of ConvNets is for feature identification. DL is based on distributed representation learning with multiple levels of representation for various layers. In simple terms, each layer learns a new feature from its preceding layer which makes the learning process concretized. By this, the learning procedure follows a hierarchy by transforming a low level representation at the first layer to a very high level feature at the last layer with multiple intermediate stages. The learning of these intermediate stages can also be utilized. Deep architectures empower deep learning strategy using greedy-layer-wise training mechanism which enables to extract only those features that are useful for learning. Apart from layer-wise training, an overall supervised training for fine-tuning the networks is responsible for the success of DL.

Deep Belief Networks (DBN) proposed by Hinton in 2006 [16] has multiple interconnected hidden layers whereeach layer acting as an input to the next layer and is visible only to the next layer. Each layer in a DBN has no lateral connection between its nodes presented in that layer. The nodes of DBN are probabilistic logic nodes thus allowing the possibility of using activation function. Initially, the first layer receives the input, and the output after applying activation function is treated as an input to the second layer and the process continues. Thus, in DBN methodology, each layer is treated as a separate neural network with one hidden layer. The transformation of data can be done using activation function or sampling. In this way, the subsequent hidden layer becomes a visible layer for current hidden layer so as to train it as a RBM.

An auto-encoder neural network is used for reducing dimensionality of data by identifying efficient method to transform high dimensional data, which is complex to optimize, into a lower dimensional data. A deep auto-encoder (DAE) comprises of multiple auto-encoder networks as layers which are stacked together. DAE comprises of encoder and decoder layers with multiple hidden layers. Initially both encoder and decoder networks are assigned with random weights and trained by observing the discrepancy between original data and output obtained from encoding and decoding. Next, the error is back-propagated initiallythrough the decoder network and then followed by an encoder network; this entire system is names as auto-encoders [16].

3. SPEAKER IDENTIFICATION PHASES

Based on the application requirements, SID may be distinguished in different types as shown in figure 1. The first categorization of SID is based on the presence of speaker's voice print [17]. To put it differently, the SID can be a closed-set approach where the speaker is verified with the existing voice prints in the database. The other approach is an open set approach in which the identification process is done for a new speaker whose voice prints do not exist in the database.

The second categorization can be based on the level of user cooperation and control in application. In other words, this category depends on the content of speech. This type is further divided into two types: text dependent approach and text independent approach. In the text dependent approach, the speaker repeats the same text for identification which he / she used duringenrolment (i.e. training) [5]. In text independent SID systems, the speaker is identified irrespective of content of the utterance [18]. The identification approach in a text dependent SID can be either the same for all users or user specific.

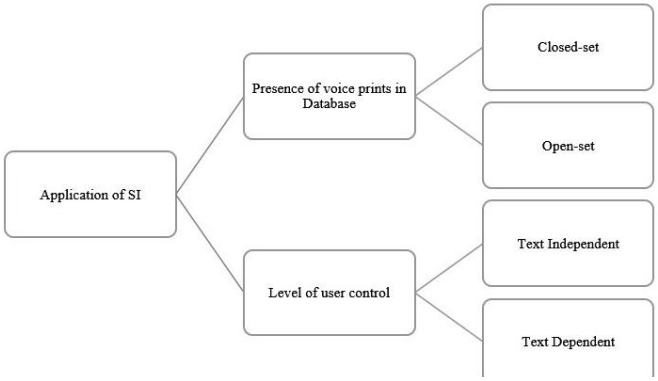

Figure 1: classification of speaker identification

The key aspect of SID is identifying the test speaker from the group of speakers. The input / test speaker characteristics are compared with available speaker models (existing data) in the database.

The SID process is conductedthroughtwo phases. The initial phase is called training phase or enrolment followed by a second phase

called matching. This process is shown in figure 2. A brief description of these two phases are presented in the following.

3.1 Training / Enrolment

The enrolment or training phase is the initial phase where the input speaker signal is pre-processed and its features are extracted. Pre-processing is a form of cleansing to make it suitable to identify and extract characteristic features of the speaker signal. The process of feature extraction will enable the presentation of the speaker vocal characteristics to construct a model for that particular speaker.

The extracted features may need to be normalized before constructing a repository of speaker specific models. The process of training involves algorithms and background models that can be performed offline (offline training). This offline training is followed by adapting the model for that particular speaker (model adaptation) which is performed online. The outcome of this phase is generating speaker specific models with unique characteristics for each speaker and creating a database of these models.

Feature extraction is important for SID since the characteristic features reflect the identity of the speaker. The more accurate the extracted features are the more differentiable the speaker will be in a group of speakers. This accuracy will directly influence the second phase (matching) since inaccurate speaker models reduce the quality of SID systems. It is important to identify that the extracted features have enough discriminative data to distinguish various speakers [19]. Moreover, voice input includes both linguistic data at the high-energy levels and other data (surrounding sounds, noise etc.) at the low energy level. Thus, to build an accurate speaker model it is necessary to implement these energy level guidelines to differentiate the speakers according to the frequency wrapping present in the speech signals [20].

A speaker is identified by extracting characteristic features of the speaker via processing the input speech signal. The spoken text of the speaker is identified in the form of frequency patterns in the complete band of frequencies. The differences in the frequency band is crucial in comparing various speakers since different speakers inhabit different frequencies. The patterns in the band is obtained by applying Fourier transform and the extraction of Mel-Frequency Cepstrum Coefficients (MFCCs) for each segment [21]. MFCC is widely used approach in SR, SID and its variants. MFCC is a collection of coefficients that is used as speech features. MFCC features are constructed using frequencies of vocal track information. A time-series analysis is performed on vocal track information for the features by segregating the input signal into different components (Filter bank approach). The designing of filters applied in MFCC is similar to human auditory frequency perception [22].

Each speaker has a different subspace within the acoustic signal cluster. This subset describes the direction of the vectors of a speaker sample depending on the speech. Hence, in order to extract speech features representing individual speakers, it is necessary to analyse these vectors to group characteristic features. The vectors that are used to identify unique characteristics of a speaker are called identity vectors or i-vectors [23]. Typically, the i-vectors are extracted with low speech utterances with low dimensional values 400 and 600compared with super vectors for speech which are over 1000 dimensions. Recent implementations of i-vector based approaches showed promising results for SID systems by converting speech signals into i-vectors [24].

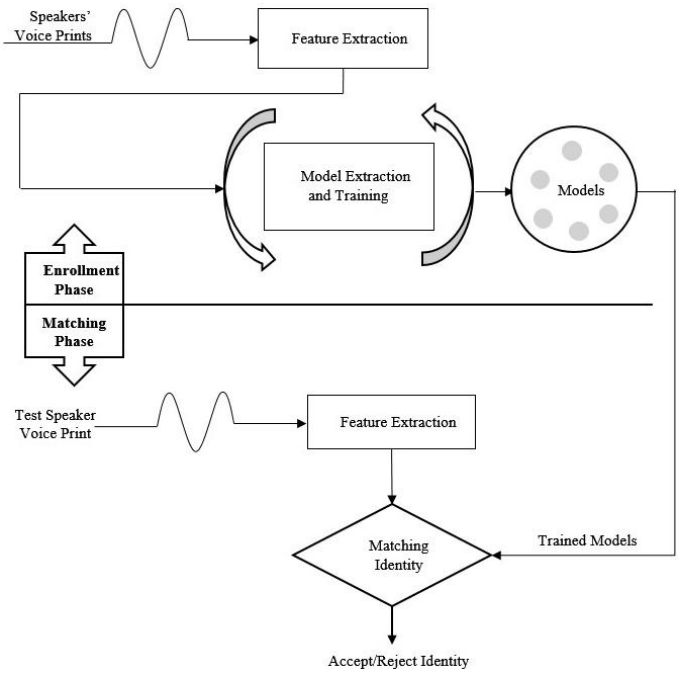

Figure 2: Phases of Speaker Identification

3.2 Matching

The matching phrase is responsible to identity the speaker by comparing the test speaker's voice prints with the existing models stored in the database during the enrolment phase; the comparison of unique characteristic features is what defines 'matching'. There have been various approaches for matching phase with Gaussian mixture models (GMMs) and pattern recognition techniques [25].

The first part of the matching phase is similar to enrolment phrase where features are extracted from the test speakers voice print. The test model constructed with these features is compared for a 'match' in the existing models presented in the repository. GMM based approaches are very efficient where there is no prior knowledge of spoken text (text independent).

For text-dependent speaker recognition, where there is a prior knowledge about the spoken text, additional knowledge can be incorporated to help the matching process to be more accurate. This additional knowledge for text dependent speaker recognition is incorporated as likelihood functions. Among various approaches to build likelihood function, Hidden Markov Models (HMMs) are widely used considering the spatio-temporal nature of the data. For i-vector based approaches, the speaker models are extracted using GMMs that are adopted using Universal Background Model (UBM).

Prominent SR and SID systems are developed using MFCC based approaches for enrolment and GMM usually with HMM for matching. GMM-HMM based algorithms determine the speaker by extracting frames of coefficients that are presented in the input. Alternatively, ML algorithms are also used for matching. Recently, other approaches for feature extraction using ML methods, especially neural networks, have been implemented successfully. In particular, MFCC was used for feature extraction followed by ML algorithms for matching. Though many ML paradigms are quite successful in SR as well as SID, using ML for feature extraction attained little success till recently. With the success of DNNs in various applications, Hinton proposed a novel approach for speaker recognition at the frame level [26]. DNNs with many

hidden layers have outperformed GMM and MFCC based models in SR for various benchmark problems [27].

4. DEEP LEARNING FOR SPEAKER IDENTIFICATION

The process of SID has two phases (as presented in the previous section) which forms the basis for categorization of DNN based approaches for SID as well. The DNN approaches for SID can be divided into three categories:

1.Enrolment / feature extraction
2.Matching
3.Both enrolment and matching

The first category applies DNN for extracting features in the enrolment phase. These features are then utilized in the matching phase using conventional SID approaches like GMM. In the second category, DNN is employed as a classifier for matching. For this category, the feature extraction is done using conventional approaches like MFCC. In the third category, DNN is appliedfor both enrolment and matching. In other words, using DNN for the entire SID process. The following sections provide more information about the DNN-based SID categories and their variations.

4.1 DNN as feature extractor

Using DNN to extract features from acoustic speech signals was initially proposed in 2014 [28]. This approach uses DNN for extracting features instead of regular models for representing voice frames (similar to i-vector based approach). This approach used supervised training instead of regular convolutional networks. A DNN topology is designed with each layer working at the acoustic frame level. Each frame of speaker voice input is fed to this DNN and the output activation values of the last layer is accumulated as a representation of that particular speaker. This representation of speaker is termed as d-vector. For regular DNN approaches, Softmax layer is used for output. This approach uses the output from the last layer instead of regular Softmax layer. Removing output layer enables reducing the size by one layer which is quite significant for DNN. Further, removing Softmax (supervised) layer will enable better generalization to extract compact speaker model for unknown speakers [28].

Different to regular SID approaches, this approach does not use any adaptation technique for extracting known features in the training phase. Instead, this approach uses DNN model for extracting specific features in both enrolment and matching phases.

4.2 DNN as classifier

A typical DNN classifier for speaker recognition uses a set of stacked features as input typically from feature extraction approaches like MFCC [15]. The features used in initial DNN approaches are short frame based with 20 ms with a context of ten frames for each segment of input. Each DNN is expected to predict probability of speaker for the input frames that are fed to DNN. To obtain the overall decision comprising of multiple frames, each prediction of DNN can be averaged out to find the speaker class. An alternate approach is using two different DNNs one for frame level prediction followed by the second one for classification.

4.3 DNN approach for feature extraction and Prediction

A Deep convolution neural networks based approach was proposed for SID with two different DNNs one for feature extraction and other one for classification [34]. Each hidden layer activation wasused to extract feature vectors which were weight values after applying activation function. In case of high dimensionality of hidden layers, dimensionality reduction techniques like principal component analysis can be applied before passing the weights to the next layer. The dimensionality of the output layer is reduced using matrix factorization techniques. This work was consideredas the inspiration for bottleneck features extraction approach which is presented in the next section.

4.4 Stacked Bottleneck Features (SBN)

Stacked bottleneck features (SBN) based approach uses DNNs with multiple hidden layers [30-31]. This type of DNN with bottleneck layers is called Bottleneck Neural Network (BN-NN). In BN-NN, the dimensionality of the one of the layers is reduced and termed as bottle-neck layer.

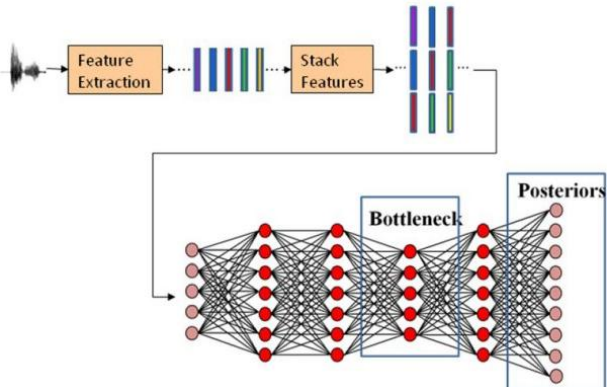

Figure 3: Stacked Bottleneck features (SBN) approach [26]

The dimensionality of the bottle-neck layer is significantly lower than the layers before and after. The input features obtained from MFCC are fed to BN-NN and the values obtained at the bottleneck layer are used for further processing. In SBN approach, two cascading neural networks are used. The output of the first SBN is fed to the second SBN justifying the name stacked bottleneck features. SBN approach is shown in figure 3.

4.5 i-vector based approaches

Among most SID approaches i-vector based approaches provided better results. Due to the success of DL in speech recognition, researches implemented i-vector based approach for SID as well [8,27,32-33]. There are two scenarios that were implemented with DL and i-vector. The first scenario is to use DL to extract i-vectors [27, 34-35], and the second approach is to use DL s a classifier after extractingi-vectors [33, 36].

In the approach of single and multi-session SID, deep belief networks (DBNs) are used instead of regular DNNs [8]. The main advantage of DBNs is the non-requirement of labelled data for training which makes it unsupervised and more suitable for identifying unknown speakers. These models of DBNs with unsupervised training is called universal DBN or UDBN [38]. UDBN uses i-vectors extracted from different speakers and fed into DBN.

4.6 DNN posteriors

GMM models are used to extract i-vector statistics which is also called universal background model or UBM [31]. The value of i-vector is calculated by reducing dimensionality transformations applying first order statistics on the frame level GMM component densities for each frame. A novel approach for extracting i-vectors using phonetically-aware deep neural network was proposed in 2014 [37]. This DNN based approach followed the similar lines of GMMs-based posteriors accumulating i-vector values for each layer. This phonetically-aware deep neural network has attained considerable success over GMM based models particularly improving the performance for SID systems [35].

5. CONCLUSION

This paper presents a comprehensive analysis of various deep learning (DL) based approaches for speaker identification (SID). DL has attained state of art results for various machine learning problems which created interest in research community. However, many SID researchers may not be fully aware of the DL implementations for SID and DNN applications in improving SID performance. This paper tried to address this knowledge gap and presented a review on DL implementations for SID.

To conclude, the process of SID was briefly introduced including few conventional approaches for feature extraction and matching phrases. Its noteworthy to observe that MFCC is considered to be most popular feature extraction approach for SID whereas GMMs models are widely implemented for matching speaker voice prints. Further, this paper presented various DL approaches for SID based on implementation with respect to SID process.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[2] S. S. Tiruala. Deep learning: Fundamentals, methods and applications. In J. Porter, editor, *DEEPLEARNING USING UNCONVENTIONALPARADIGMS*, chapter 1, pages 11–. NOVA publishes, New York, 2014.

[3] R. Rajesh, K. Ganesh, S. C. L. Koh, N. Singh,R. Khan, and R. Shree. International conference on modelling optimization and computing applications of speaker recognition. *Procedia Engineering*, 38:3122 – 3126, 2012.

[4] S. R. Shahamiri and S. S Binti Salim.Real-time frequency-based noise-robust Automatic Speech Recognition using Multi-Nets Artificial Neural Networks: A multi-views multi-learners approach. *Neurocomputing*, 129:199-207, 2014

[5] H. Kekre, A. Athawale, and M. Desai. Speaker identification using row mean vector of spectrogram. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, pages 171–174. ACM, 2011.

[6] F. Richardson, D. Reynolds, and N. Dehak. A unified deep neural network for speaker and language recognition. *arXiv preprint arXiv:1504.00923*, 2015.

[7] M. McLaren, Y. Lei, and L. Ferrer. Advances in deep neural network approaches to speaker recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4814–4818. IEEE, 2015.

[8] O. Ghahabi and J. Hernando. Deep learning for single and multi-session i-vector speaker recognition. *arXiv preprint arXiv:1512.02560*, 2015.

[9] Y. LeCun, K. Kavukcuoglu, and C. Farabet.
Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256, May 2010.

[10] Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In NIPS, 2012.

[11]M. Pobar and I. Ipˇsi´c. Online speaker de-identification using voice transformation. In *Information and Communication Technology, Electronics andMicroelectronics (MIPRO), 2014 37th International Convention on*, pages 1264–1267. IEEE, 2014.

[12]T. Justin, V. Struc, S. Dobriˇsek, B. Vesnicer, I. Ipˇsi´c,ˇ and F. Miheliˇc. Speaker de-identification using diphone recognition and speech synthesis. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 4, pages 1–7. IEEE, 2015.

[13]M. Dutta, C. Patgiri, M. Sarma, and K. K. Sarma. Closed-set text-independent speaker identification system using multiple ann classifiers. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, pages 377–385. Springer, 2015.

[14]G. Tesauro. Practical issues in temporal difference learning. In *Machine Learning*, pages 257–277, 1992.

[15]Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[16]Y. LeCun, K. Kavukcuoglu, and C. Farabet.
Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256, May 2010.

[17]D. Reynolds. An overview of automatic speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075)*, 2002.

[18]G. K. Verma. Multi-feature fusion for closed set text independent speaker identification. In International Conference on Information Intelligence, Systems, Technology and Management, pages 170–179. Springer, 2011.

[19]C. Zhao, H. Wang, S. Hyon, J. Wei, and J. Dang.Efficient feature extraction of speaker identification using phoneme mean f-ratio for chinese. In Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on, pages 345–348. IEEE, 2012.

[20]S. K. Sarangi and G. Saha. A novel approach in feature level for robust text-independent speaker identification system. In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on*, pages 1–5. IEEE, 2012.

[21] S. R. Shahamiri and S. S Binti Salim. Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. *Advanced Engineering Informatics*, 28 (1): 102-110, 2014

[22]N. Sen and T. Basu. Features extracted using frequency-time analysis approach from nyquist filter bank and gaussian filter bank for text-independent speaker identification. In *European Workshop on Biometrics and Identity Management*, pages 125–136. Springer, 2011.

[23]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*, 19(4):788–798, 2011.

[24]Y. Qian, T. Tan, D. Yu, and Y. Zhang. Integratedadaptation with multi-factor joint-learning for far-field speech recognition. In *2016 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pages 5770–5774. IEEE, 2016.

[25]K. Kumar, Q. Wu, Y. Wang, and M. Savvides. Noise robust speaker identification using bhattacharyya distance in adapted gaussian models space. In *Signal Processing Conference, 2008 16th European*, pages 1–4. IEEE, 2008.

[26]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*, 29(6):82–97, 2012

[27]O. Ghahabi, A. Bonafonte, J. Hernando, and A. Moreno. Deep neural networks for i-vector language identification of short utterances in cars. *Interspeech 2016*, pages 367–371, 2016.

[28]E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE*International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4052–4056. IEEE, 2014

[29]K. Vesely`, M. Karafia ́t, and F. Gr ́ezl. Convolutive bottleneck network features for lvcsr. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pages 42–47. IEEE, 2011.

[30]P. Matejka, L. Zhang, T. Ng, S. H. Mallidi, O. Glembek, J. Ma, and B. Zhang. Neural network

[31]F. Richardson, D. Reynolds, and N. Dehak. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10):1671–1675, 2015.

[32]F. Richardson, D. Reynolds, and N. Dehak. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10):1671–1675, 2015.

[33]V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis. I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6334–6338. IEEE, 2014.

[34]Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu. Deep feature for text-dependent speaker verification. *Speech Communication*, 73:1–13, 2015.

[35]D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey. Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 378–383. IEEE, 2014.

[36]O. Ghahabi and J. Hernando. Global impostor selection for dbns in multi-session i-vector speaker recognition. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 89–98. Springer, 2014.

[37]Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1695–1699. IEEE, 2014.

[38]O. Ghahabi and J. Hernando. Deep belief networks for i-vector based speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1700–1704. IEEE, 2014.