RESEARCH ARTICLE OPEN ACCESS

Automatic Speaker Recognition: A Survey

Rizwan K Rahim[1],Tharikh Bin Siyad [2], Muhammed Ameen M.A [3], Muhammed Salim K.T [4] , Selin M [5]

[1], [2], [3], [4], [5] Computer Science and Engineering, APJ Abdul Kalam Technological University – India

ABSTRACT

Speaker recognition is the task of identifying persons from their voices. Recently, deep learning has dramatically revolutionized speaker recognition. This paper, reviews several major subtasks of speaker recognition, including speaker verification, identification, and robust speaker recognition, with a focus on deep learning-based methods. An Automatic Speaker Recognition is a biometric system that allows you to identify and verify people, using voice as a discriminatory feature. Automatic Speaker Recognition (ASR) using Autoencoder is discussed here. This paper discusses the Deep Learning methodologies for ASR followed by different Feature Extraction techniques. Then the Autoencoder technology, its working, and its architecture and, how the ASR works using Deep Learning are discussed. Finally, A survey about robust speaker recognition from the perspectives of domain adaptation and speech enhancement, which are two major approaches to dealing with domain mismatch and noise problems is done.

Keywords- Deep Learning, Automatic Speaker Recognition, Auto-encoder, Feature Extraction, MFCC.

I. INTRODUCTION

An Automatic Speaker Recognition (ASR), is a non-invasive biometric system because it manipulates the voice as a discriminatory feature, also it presents a great versatility during the evaluation so this is a process that only requires that the user speaks, which constitutes a natural act of human's behavior [3]. It is known that a speaker's voice contains personal traits of the speaker, given the unique pronunciation organs and speaking manner of the speaker, e.g. the unique vocal tract shape, larynx size, accent, and rhythm. Therefore, it is possible to identify a speaker from his/her voice automatically via a computer system. This technology is termed automatic speaker recognition, which is the core topic of this paper. Speaker recognition is a fundamental task of speech processing and finds its wide applications in real-world scenarios. For example, it is used for the voice-based authentication of personal smart devices, such as cellular phones, vehicles, and laptops. It guarantees the transaction security of bank trading and remote payment. It has been widely applied to forensics for investigating a suspect to be guilty or not guilty, or surveillance and automatic identity tagging. It is important in audio-based information retrieval for broadcast news, meeting recordings, and telephone calls. It can also serve as a frontend of automatic speech recognition (ASR) for improving the transcription performance of multi-speaker conversations [4].

Here, the reader gets a comprehensive overview of the deep learning-based speaker recognition methods in terms of the vital subtasks and research topics, including speaker identification, voice diarization, and obust speaker recognition. From this study, we hope to provide a useful resource for the speaker recognition community. The main contributions of this article are to summarize deep learning-based feature extraction techniques for speaker verification and identification, Make an overview of the deep learning-based speaker diarization, with an emphasis on recent supervised, end-to-end, and online diarization, and Survey robust speaker recognition from the perspectives of domain adaptation and speech enhancement, and Domain adaptation and speech enhancement are two major approaches to dealing with domain mismatch and noise problems.

Many studies have proposed techniques to improve the accuracy of ASR in noisy and reverberant conditions. One approach is to enhance the noisy features by applying noise removal techniques, Others designed discriminative, handcrafted features that are more robust against noise and reverberation. Many works also propose adapting the acoustic models to noisy conditions. For deep learning frameworks, various architectures are investigated to find better systems such as recurrent neural networks (RNN) and convolutional neural networks (CNN).

But here, We are trying to deal with Automatic Speaker Recognition using the Auto-encoder technology [5]. For the feature extraction techniques, Mel Frequency Cepstral Coefficients(MFCC) predominates. Even though MFCC is the most cited and used, there are some robust feature extraction techniques that will work more accurately and efficiently.

1.1 Overview and scope

This summary outlines four major research branches of speaker recognition, which are speaker verification, identification, and robust speaker recognition respectively. The flowcharts of the first three branches are burst; speaker recognition deals with the challenges of noise and domain mismatch troubles. The topics of the overview are organized in Fig. 2, which are characterized briefly as follows.

Speaker verification aims at verifying whether an utterance is pronounced by a hypothesized speaker based on his/her pre-recorded utterances [6]. Speaker verification algorithms can be classified into stage-wise and end-to-end ones. A stage-wise speaker verification system usually consists of a front-end for the extraction of speaker features and a back-end for the resemblance calculation of speaker features. The front-end transforms an utterance in the time domain or time-frequency domain into a high-dimensional feature vector. It accounts for the recent advantage of deep learning-based speaker recognition.

The back-end first computes a similarity score between enrollment and test speaker features and then compares the score with a threshold:

$$f(\mathbf{x}^e, \mathbf{x}^t; \mathbf{w}) \underset{H_1}{\overset{H_0}{\gtrless}} \xi \qquad \text{.....(1)}$$

where f(·) indicates a function for calculating the similarity, w stands for the parameters of the back-end, x^e and x^t are the enrollment and test speaker features respectively, ξ is the threshold, H_0 represents the hypothesis of x^e and x^t belonging to the same speaker, and H^1 is the opposite hypothesis of H^0. One of the major responsibilities of the back-end is to compensate for channel variability and reduce interferences, e.g. language mismatch. Because most back-ends aim at alleviating the interferences, which belongs to the problem of powerful speaker recognition.

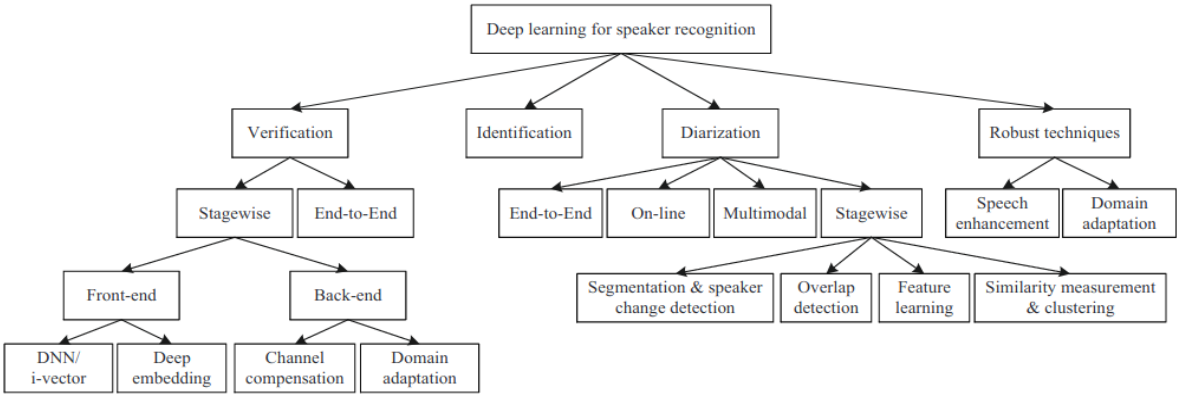

Fig 1.1 Overview of deep learning-based speaker recognition

II. DEEP LEARNING

Deep learning also called Deep Neural Network comprises many layers with various neurons in each layer. These layers can vary from a few to thousands and each layer may further comprise thousands of neurons (processing unit). The simplest function in a neuron is to multiply the input values with the

allocated weight to each input and sum up the result [8].

Deep learning approaches are reasonable for us to solve many problems. In the future, it is foreseeable that deep learning could demonstrate perfect theories to explain its performance. Meanwhile, its capacities for unsupervised learning will be enhanced since there are millions of pieces of data in the world but it is not applicable to add labels to all of them. It is also predicted that neural network structures will become more complex so that they can extract more semantically significant features [7]. What is more, deep learning will combine with reinforcement learning and we can use these benefits to accomplish more tasks.

Deep learning models usually recognize hierarchical structures to connect their layers [9]. The output of a lower layer can be considered as the input of a massive layer via simple linear or nonlinear computations. These models can transform low-level features of the data into high-level abstract features. Owing to this characteristic, deep learning models can be more powerful than shallow machine learning models in feature representation.

There are many kinds of Deep Learning technologies that we can use for ASR programs. Even though the most cited and used one is Convolutional Neural Network (CNN), it has so many drawbacks. So we are using Auto-encoder as the Deep Learning technology.

III. METHODOLOGY

A systematic workflow of the proposed Automatic speaker recognition system is shown in Fig. 3.1. Given an input speech signal, voice activity detection is accomplished to identify speech presence or speech absence in the given speech signal [10]. An Auto-Encoder is used to denoise the noisy input and enhance the quality & intelligibility of distorted speech signals. Then audio feature vectors are extracted and used to train the models using the Gaussian mixture model. Lastly, the network recognizes the speaker by testing the sample with the trained model.

3.1. VOICE ACTIVITY DETECTION

Voice Activity Detection is a strategy used in speech processing to recognize speech existence of speech absence in audio. This procedure processes the speech signals to rule out the silence fraction, otherwise, the training might be biased [11]. Long Term Spectral Divergence (LTSD) algorithm [22] was used concurrently with a noise compression script from SOX1 to perform this task. LTSD algorithm breaks an utterance into overlapped frames and gives scores for each frame on the probability that there is voice activity in the frame. The probability is then developed to extract all the duration with voice activity.

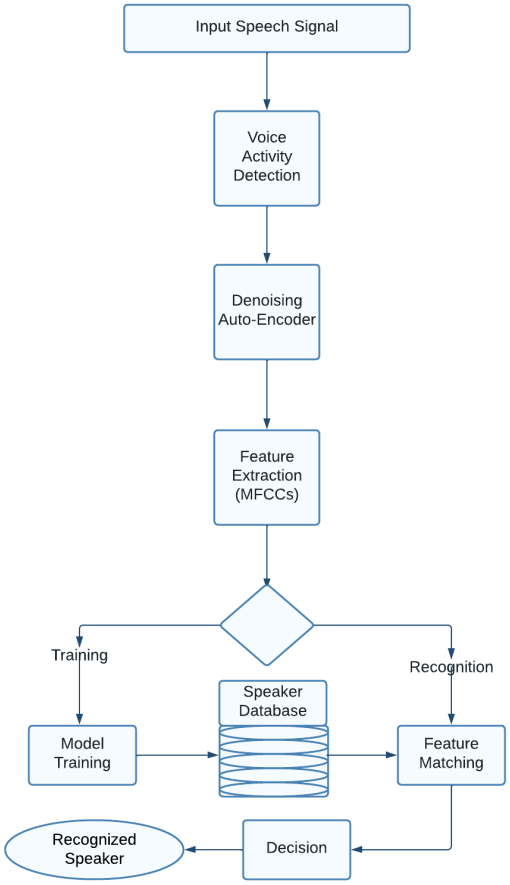

Fig 3.1 Systematic workflow of the proposed system

3.2. AUTO-ENCODER

Speech enhancement (SE) aims to improve the quality and intelligibility of deformed speech signals, which may be caused by background noises, interference, and recording accessories [12]. SE strategies are generally used as pre-processing in

various audio-related applications, such as speech communication, automatic speech recognition (ASR), speaker recognition, hearing assistance, and cochlear implants Denoising Autoencoders (DAE) has been widely explored in the field of speech signal processing. documented the usefulness of DAE on dereverberation and distant-talking speech recognition. [10] investigated the performance of DAE on unsupervised domain adaptation for speech emotion recognition.

Auto-encoder allows our speaker verification system to quickly adapt to the release of any new models of smart speakers. Finally, as the next step, we plan to extend the exploration beyond smart speakers to other fields in the industry where labeled speakers are scarce but unlabeled data is abundant.

An autoencoder-based semi-supervised curriculum learning scheme is proposed to automatically accumulate unlabeled data and iteratively update the corpus during training. This new training technique allows us to (1) progressively improve the size of the training corpus by using unlabeled data and rectifying previous labels at run-time; and (2) improve robustness when generalizing to numerous conditions, such as out-of-domain and text-independent speaker verification tasks. It is also discovered that a denoising autoencoder can considerably enhance the clustering accuracy when it is trained on a carefully-selected subset of speakers.

An autoencoder-based semi-supervised curriculum learning approach is proposed to rapidly adapt the speaker verification system to the unseen new domain in which no labeled data is available [13].

3.3. FEATURE EXTRACTION

An Automatic Speaker Recognition (ASR), is a non-invasive biometric system because they use the voice as a discriminatory feature, also it illustrates a great versatility during the examination so this is a process that only requires that the user speaks, which constitutes a natural act of human's behaviour [3]. The voice has six information statuses, from spectral (lower level) to semantic level (upper level), and the complexity during the information extraction procedure rises proportionally respecting the level on which it's worked [14]. In real circumstances, the voice can be supported by all kinds of noise, such as public transport sound, channel distortion, and even reverberation, because it is significant to use reliable techniques in noisy conditions.

Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is one of the most cited and used methods in the speech processing community. It's established on the simulation of cochlear auditory capability, with the design of a uniformly spaced filterbank in the Mel frequency scale, which when altered to the linear frequency scale, the spacing between filters is linear in the range of first 1000 Hz.

Feature extraction acts as a crucial position in training a model. It is essential to extract a set of features from audio signals. A group of extracted features is provided as input to the classifier. In speech recognition feature vector represents the speech waveforms. There are various feature extraction strategies available to extract the features from audio signals such as MFCC, delta MFCC, LPCC, PCA, etc [15].

The Fourier transformation of the time-domain audio signal into the frequency domain is called a spectrum. By using fast Fourier transformation samples from each frame are converted into frequency domain i.e. spectrum. Mel scales for frequency f find out by using the equation:

$$Mel\ f = 2595 \log_{10}(f/700 + 1)$$

Log magnitude of mel is called the mel spectrum. DCT (Discrete Cosine Transform) applied to mel spectrum and mel frequency cepstral coefficients features are computed. The computation of MFCC features comprises numerous phases such as pre-processing, framing, windowing, estimation of discrete Fourier transform, mel frequency, and inverse document frequency.

Functionally, this scheme is established on the introductory process of windowing and overlapping, then the signal power spectrum is assessed and distributed into sub-bands through a Mel filterbank, after that is logarithmically compressed, and finally the Discrete Cosine Transform (DCT) is applied for accumulating information in the first coefficients.

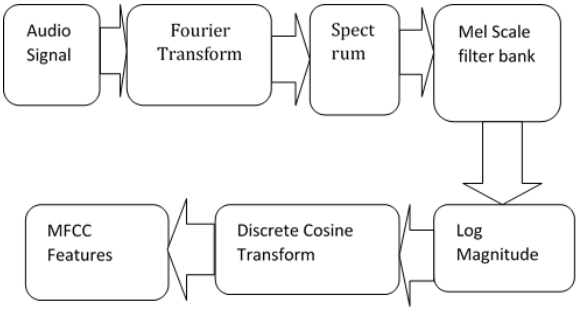

Fig 4.1 Process of MFCC feature extraction

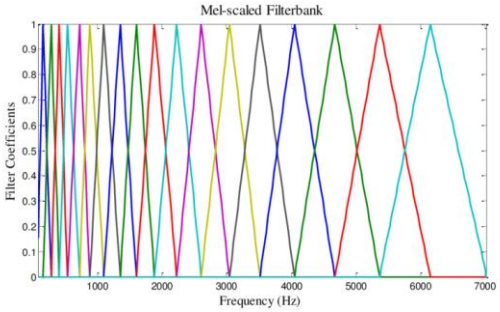

Fig 4.2 Scaled Filterbank on Mel Frequency

3.4. SPEAKER RECOGNITION

Speaker recognition is a technique used to automatically understand a speaker from a recording of their voice or speech utterance. It has evolved into an economical and reliable method for person identification and verification. This paper presents the advancement of an automatic speaker recognition system that incorporates the classification and recognition of speakers. Four classifier models, namely, Support Vector Machines, K-Nearest Neighbors, Multilayer Perceptrons (MLP), and Random Forest (RF), are trained using the WEKA data mining tool[16]. Auto-WEKA is assigned to specify the best classifier model together with its best hyper-parameters. The performance of each model is assessed in WEKA manipulating 10-fold cross-validation. The following evaluation measurements are used; RMSE, Accuracy, Precision, and Recall are used to evaluate the performance of the models. The Speaker verification is again divided into two, Speaker Verification and Speaker Enrollment.

3.4.1. Recognition: It is a one-on-one matching process. Here we have prior information that this is speaker "X" (This is the recognition phase), then the voice is matched to the speaker "X" (Which we get from the enrollment phase) voice print only. Based on the amount of similarity we can set the threshold for matching.

3.4.2. Speaker Enrollment: In this phase when a new user comes into the system their voice samples are stored and the d-vector is calculated of all the samples and an average is taken and stored as that user's voiceprint. so that next time the same user comes we can match it with this stored voiceprint. Here longer voice samples help to capture features better and more samples help to show the variation of the user's voice. A good voice sample falls in the range of 3–5 seconds. Speaker Enrollment is also known as Speaker Identification.

Figure 4.4 shows both Speaker Verification and Speaker Identification.

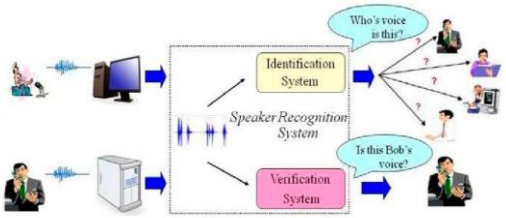

Fig 4.4 Speaker Verification and Speaker Identification

IV. CONCLUSION

Here presented a light introduction on an Automatic Speaker Recognition System using Deep Learning and the present scenario where it is now. In the proposed system, the first phase is voice activity detection and it is done by Long Term Spectral Divergence (LTSD) algorithm. Then we discussed the Auto-encoder technology as the denoising or Speech Enhancement technique we use in the ASR system. We familiarize ourselves with the feature extraction procedure and the most important phase in the system. A well-known feature extraction method is used i.e., MFCCs (Mel Frequency Cepstral Coefficient). Then we looked into the Speaker Verification procedures and their classifications.

V. REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to super vectors," Speech communication, vol. 52, no. 1, pp. 12-40, 2010.

[2] R. R Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition general classifier approaches and data fusion methods," Pattern Recognition, vol. 35, no. 12, pp. 2801-2821, 2002.

[3] Campbell, Edward & Lara, José & Hernández-Sierra, Gabriel. (2018). Feature extraction of Automatic Speaker Recognition, analysis,s and evaluation in a real environment.

[4] Červa, Petr & Silovský, Jan & Zdánský, Jindrich & Nouza, Jan & Seps, Ladislav. (2013). Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives. Speech Communication. 55. 1033-1046.

[5] Yu, Dong & Li, Jinyu. (2017). Recent progress in deep learning-based acoustic models. IEEE/CAA Journal of Automatica Sinica.

[6] Das, Rohan & Prasanna, S.. (2017). Speaker Verification from Short Utterance Perspective: A Review. IETE Technical Review. 35. 1-19.

[7] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications, and Research Directions. *SN COMPUT. SCI.* 2, 160 (2021).

[8] Steven Walczak, Narciso Cerpa, Artificial Neural Networks. Encyclopedia of Physical Science and Technology (Third Edition), 2003

[9] Jianzhu Ma,1, Michael Ku Yu,1,2, Samson Fong,1,3, Keiichiro Ono,1 Eric Sage,1 Barry Demchak,1 Roded Sharan,4 and Trey Ideker1,2,3,* Using deep learning to model the hierarchical structure and function of a cell

[10] Dharm Singh Jat, ... Charu Singh, Voice Activity Detection-Based Home Automation System for People With Special Needs. Intelligent Speech Signal Processing, 2019

[11] rVAD: An unsupervised Segment-based Robust Voice Activity Detection Method Zheng-Hua Tana, Achintya kr. Sarkara,b, Najim Dhaka

[12]Cheng Yu, Ryandhimas E. Zezario, Syu-Siang Wang, Jonathan Sherman, Yi-Yen Hsieh, Xugang Lu, Hsin-Min Wang, Senior Member, IEEE, and Yu Tsao, Senior Member, IEEE. Speech Enhancement based on Denoising Autoencoder with Multi-branched Encoders

[13] Siqi Zheng, Gang Liu, Hongbin Suo, Yun Lei Machine Intelligence Technology, Alibaba Group. Autoencoder-based Semi-Supervised Curriculum Learning For Out-of-domain Speaker Verification.

[14] Kiran Adnan, Rehan Akbar. An analytical study of information extraction from unstructured and multidimensional big data Journal of Big Data 6, Article number: 91 (2019)

[15] Vaisali A. Kherdekar, Dr.Sachin A.Naik (2021)Convolution Neural Network Model for Recognition of Speech for Words used in Mathematical Expression.

[16] Tumisho Billson Mokgonyane,Tshephisho Joseph Sefara,Thipe Isaiah Modipa,Mercy Mosibudi Mogal,Madimetja Jonas Manamela .(2019) Automatic Speaker Recognition System based on Machine Learning Algorithms

[17] D. Ferbrache, "Passwords are the broken-the future shape of biometrics," Biometric Technology Today, vol. 2016, no. 3, pp. 5-7, 2016.

[18] L. Hamid, "Biometric technology: not a password replacement, but a compliment," Biometric Technology Today, vol. 2015, no. 6, pp. 7-10, 2015.

[19] N. Singh, R. Khan, and R. Shree, "Applications of speaker recognition," Pmcedia engineering, vol. 38, pp. 3122-3126, 2012.

[20] 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.

[21] E. Aliyu, O. Adewale, and A. Adetunmbi, "Development of a text-dependent speaker recognition system," International Journal of Computer Applications, vol. 69, no. 16, 2013.

[22] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. GonzalezDominguez, "Deep neural networks for small footprint text-dependent speaker verification." in ICASSP, vol. 14, 2014, pp. 4052-4056.