Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey

Nishtha H. Tandel
Dept. of Information Technology
Dharmsinh Desai University,
Nadiad, India
nishtha2710@gmail.com

Harshadkumar B. Prajapati
Dept. of Information Technology
Dharmsinh Desai University,
Nadiad, India
prajapatihb.it@ddu.ac.in

Vipul K. Dabhi
Dept. of Information Technology
Dharmsinh Desai University,
Nadiad, India
vipuldabhi.it@ddu.ac.in

***Abstract*—Voice comparison is a variant of speaker recognition or voice recognition. Voice comparison plays a significant role in the forensic science field and security systems. Precise voice comparison is a challenging problem. Traditionally, different classification and comparison models were used by the researchers to solve the speaker recognition and the voice comparison, respectively but deep learning is gaining popularity because of its strength in accuracy when trained with large amounts of data. This paper focuses on an elaborated literature survey on both traditional and deep learning-based methods of speaker recognition and voice comparison. This paper also discusses publicly available datasets that are used for speaker recognition and voice comparison by researchers. This concise paper would provide substantial input to beginners and researchers for understanding the domain of voice recognition and voice comparison.**

Keywords—voice comparison, speaker recognition, deep learning, Siamese NN

I. INTRODUCTION

Voice comparison [1] is a difficult problem to solve because the voice of a person may change due to the emotion, age-gap, and throat infection [2]. On the other hand, when a speaker tries to say precisely the same utterance twice, a measurable difference occurs in the speaker's voices. However, a robust voice comparison is necessary because it can be used in many fields, such as forensic science [1], authentication/verification [30], [39] surveillance, etc. Though voice comparison is a hard problem for researchers, newer machine learning techniques, such as deep learning [3], have the capability to provide an appropriate solution for the problem.

We highlight differences among different voice processing operations that are used in the literature. Speaker recognition is the method of recognizing who is the speaker by using speaker's unique information. The recognition of speakers is typically divided into two categories: (1) speaker identification [9] and (2) speaker verification or authentication [30]. Speaker identification is the process of determining an unknown speaker's identity by matching his or her voice to the voices in the database of registered speakers. Speaker verification can determine whether a person is what he or she claims to be based on his or her voice sample. There is an additional variant of speaker recognition called voice comparison [1] in which two voices are supplied as input to the voice comparison system and the system determines the similarity score between two input voices. On the basis of words or text used in speech, the speaker recognition and voice comparison system are divided into two categories: (1) text-dependent and (2) text-independent. Text-dependent employs the same text for training and testing whereas text-independent employs different text for training and testing.

Saquib et al. [5] and Singh et al. [6] presented a survey of speaker recognition techniques in 2010 and 2017, respectively, which contain traditional approaches of speaker recognition. For speaker recognition and voice comparison, most of the works, e.g., [1], [7], [8], [9], [37] have been carried out using the traditional approaches by various researchers. Less amount of research exists on the use of deep learning methods on the topic of speaker recognition and voice comparison. Therefore, there is a need for such a survey that explores both traditional approaches as well as deep learning-based approaches of speaker recognition (identification and verification) and voice comparison.

This paper explores and analyzes various traditional and deep learning-based approaches to discuss potential solutions to the problem of voice comparison. This paper conducts a survey of major works carried out on speaker recognition and voice comparison to discuss all major issues and their solutions. Furthermore, the paper also analyzes the suitability of the Siamese Neural Network for the problem of voice comparison. The paper also discusses and analyzes the datasets used by various researchers for speaker recognition and voice comparison.

This paper is arranged as follows: Section II includes introduction on voice comparison, speaker identification, and speaker verification. Furthermore, a general pipeline for voice comparison is discussed, and traditional and deep learning approaches for speaker recognition and voice comparison are studied. Section III presents a detailed literature survey on speaker recognition (identification and verification) and voice comparison. Section III includes analyses of different datasets and Siamese NN (Siamese Neural Network). Finally, Section IV concludes the paper.

II. VOICE RECOGNITION AND VOICE COMPARISON

This section presents variants of speaker recognition, descriptions of the voice comparison, and traditional v/s deep learning-based approaches for voice comparison. Additionally, for voice comparison, Siamese Architecture is also studied.

978-1-7281-5197-7/20/$31.00 ©2020 IEEE

A. *Variants of Speaker Recognition*

Speaker recognition can be divided into two types: (1) speaker identification (2) speaker verification. Fig. 1 shows an illustration of speaker identification and speaker verification.

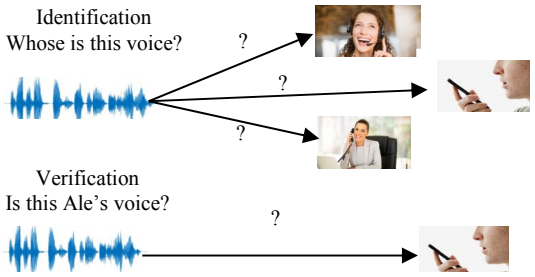

Fig. 1. Variants of speaker recognition

1) Speaker identification: The speech of an unknown speaker is processed and is compared to established speaker voice models. The unknown speaker is defined as the one that best suits. Thus, input to speaker identification is an unknown voice and the output is the name or id of the speaker.

2) Speaker verification: In this variant of speaker recognition, an unknown speaker claims an identity whose speech is compared to the registered speaker model claiming identity. Thus, input to speaker verification is the name of the speaker and his or her voice and the output is Yes or No.

There is one more variant of speaker recognition called voice comparison.

3) Voice comparison: Voice comparison is a task to analyze two recordings of the speaker and make a decision whether the voices belong to the same speaker or to different speakers. Fig. 2 shows an illustration of voice comparison. Thus, input to voice comparison is two voice recordings and the output is similarity score in the range 0 to 1.

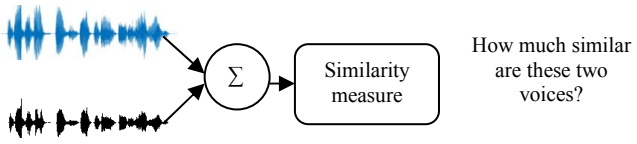

Fig. 2. Voice comparison

As discussed earlier, based on the text or words used in voice, types of system that address the problem of speaker recognition (identification and verification) and comparison can be classified into two types [40]: 1) text-dependent and (2) text-independent. Text-dependent system is connected to a predefined text used for training and testing and the text-independent system should be capable of using any text. This paper explores both the text-dependent and text-independent methods of recognition.

B. *Different approaches of voice comparison*

The forensic voice comparison is based on four specific approaches: (1) auditory, (2) spectrographic, (3) acoustic, and (4) automatic approach [11]. In all the approaches, for comparing a voice, at least two recordings of a speaker are needed. The result of the auditory approach is the experts' (machines') subjective judgment on the basis of listening of speech recording. A spectrographic approach is an image-based approach in which speech recordings are transformed into speech images, called a spectrogram. In general, the spectrogram reflects the frequency spectrum, which is also known as "voiceprints". In a spectrographic approach, an expert will pay attention at multiple words or phrases in both recordings. The expert can then look at a specific pattern of information in the image to see how close they are. In Fig. 3, one example of a spectrogram is given. Lukic et al. [12] used a voiceprint (spectrogram) as input to CNN.

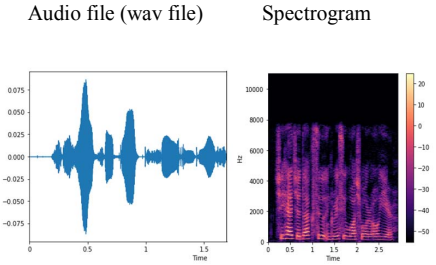

Fig. 3. Spectogram of Voice data

The acoustic-phonetic approach needs making quantitative estimates of the acoustic properties (pitch, formant, fundamental frequency, and HNR) on equivalent phonetic units in both recordings of the speakers. Cardoso et al. [1] proposed a technique to improve the performance of the Forensic Voice Comparison (FVC) system using fundamental frequency and formant. In an automatic approach, frame-wise speech features are automatically extracted. Unlike an acoustic-phonetic approach, the automated approach does not use different acoustic features on a specific part of the signal. Examples of automatic approaches are MFCC [13], LPCC [14], etc.

C. *Difference between traditional and deep learning-based techniques*

The traditional methods of speaker recognition and voice comparison system such as HMM (Hidden Markov Model), GMM (Gaussian Mixture Model), and VQ (Vector Quantization) use unique characteristics of speech features from a collection of speakers; therefore, it is necessary to choose the most successful feature extraction approaches that truly represent the characteristics of speech. There are many feature extraction techniques available such as MFCC [13], LPCC [15], and pitch [16]. As per the researchers' analysis, e.g., in [3] and [4], traditional methods are very time-consuming. Therefore, deep learning-based approaches are preferred for an automatic system to save time.

A generalized process to perform voice comparison efficiently using deep learning model is shown in Fig. 4.

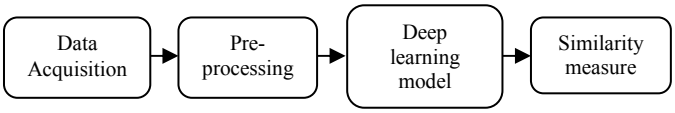

Fig. 4. Pipeline of voice comparison System

1) Data Acquisition: Forensic scientists or researchers generally avoid modeling raw audio because it ticks so often. Generally, the text-independent voice comparison system

requires a type of datasets that contain audios of the same subject having different dialogues and a text-dependent voice comparison system requires datasets that contain audios of the same subject having the same dialogue. Some datasets are available for recognition and comparison tasks; such datasets include microphone audio data (TIMIT [20]), telephone speech (NTIMIT [21]), age-wise speech corpus (VoxCeleb [22]).

2) Preprocessing: Preprocessing of audio data [17] is a very important step after data acquisition because real-world audio data is noisy. Generally, VAD (Voice Activity Detection) is used to separate voiced data and unvoiced data, i.e., VAD is used to find out the presence and absence of a human in speech. The VAD strategies utilize the prompt proportions of the dissimilarity separation among speech and noise. In the past, VAD was based on extracting features such as short-time energy [23], zero-crossing rate [24], and pitch analysis [16]. Nowadays, the classification of voiced and unvoiced segments is done based on cepstral coefficients [13], [15], and wavelet transforms [25]. The important methods of VAD and its applications are listed in Table I.

TABLE I. VOICE ACTIVITY DETECTION METHODS

VAD Methods	Application	References
Linear predictive coding (LPC)	Speech coding and speech synthesis	[14] [15]
Formant shape	Speech recognition speaker recognition	[1]
Zero crossing rate (ZCR)	Find out human presence	[23][24]
Cepstral feature	Speech recognition and speaker recognition	[13] [18]
Periodicity measure	Visualizing structural periodic changes	[51]
Pattern recognition	Voice-based personal verification	[19]

In a multi-speaker environment, we may need the answer to "who spoke when". In such a context, audio data often contains recordings of more than one person talking (i.e. telephone and meeting conversation). Speaker diarization is the method of splitting an input audio into homogeneous segments according to the speaker identity. Wang et al. [26] proposed a novel speaker diarization technique based on LSTM-based d-vector audio embeddings. Speaker diarization itself is a wide domain and hence is out of the scope of this paper. However, a recent review on speaker diarization is available in [27], which interested readers can refer.

3) Deep learning model: After preprocessing, the inputs are fed into the model. As per our understanding of the literature, the Siamese Neural Network (Siamese NN) is well adapted for the problem of comparison. Siamese NN learns a similarity function that takes two inputs (i.e. spectrogram or voiceprint) as input and shows how identical the two inputs are. The Siamese architecture's goal is not to classify input objects, but to distinguish between the two.

The Architecture of Siamese NN is shown in Fig. 5. For the first time, Siamese NN was used for Signature verification (whether the signatures belong to one person) in [28]. Recently, Siamese NN has been designed for one-shot image recognition [29]. In one-shot image recognition, the researchers have to make predictions right with only one instance of each new class. In one shot image recognition [29], the researchers discuss a method for learning Siamese NN that uses a unique structure to rank similarities between inputs naturally and experimentally conclude that Siamese NN provides better results than CNN when training is less. In the human speech domain, every human's voice has a unique formant structure and vocal pattern. Therefore, there is no need to train the network with more samples of one specific speaker's speech.

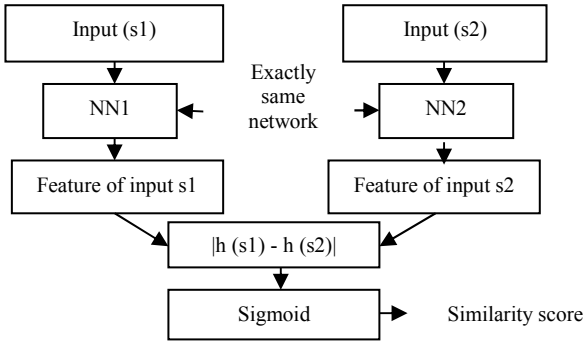

Fig. 5. Architecture of Siamese NN

III. SURVEY ON SPEAKER RECOGNITION AND VOICE COMPARISON

This section presents a broad survey of speaker identification, speaker verification, and voice comparison. We divide the survey into two subsections: a traditional approach based and deep learning-based.

A. Survey on Speaker Recognition and Voice Comparison System based on the Traditional approach

Many papers dealing with the problems and difficulties of speaker recognition and voice comparison systems have been published in recent years. Several of the papers are reviewed and analyzed in Table II. Most of the researchers used traditional based methods. Reynolds and Rose [7] proposed speaker identification based on GMM (Gaussian Mixture Model). The whole procedure of speaker identification is divided into two parts: (1) feature extraction and (2) classification. For feature extraction, the authors introduced a Mel-frequency filter bank for short utterance of speaker. For classification, GMM model is used. The result of the speaker identification technique is, the accuracy decreases when the quality of sound is degraded, i.e. GMM attains 96.8 % and 80.8% accuracy, when the speech is clean-speech and telephone speech, respectively.

Reynolds [30] proposed speaker identification and verification systems providing superior performance based on Gaussian mixture speaker models. The author [30] tested their work on publicly available datasets such as TIMIT [20], NTIMIT [21], Switchboard [45], and YOHO [46]. In their work, the whole procedure of speaker identification is divided into two parts: (1) feature extraction and (2) classification. In

the feature extraction, the speech signal of the speaker is first divided into separate speech frames, and then Mel-scale Cepstral feature (MFCC) vectors are extracted from the speech frame. For classification, GMM is used. For feature extraction, MFCC is widely used by various researchers, e.g., Chakroborty et al. [9], Tolba et al. [34], Krishnamoorthy et al. [36] and Saeidi et al. [10].

Cardoso et al. [1] proposed a new technique of voice comparison system, in which they extracted Vq (Voice Quality) features with MFCC features from speech. In their work [1], 97 Speakers are used from the DyViS corpus [52]. Out of 97 speakers, 32 speakers are used for training 33 speakers are used for testing purposes and 32 speakers for references. In DyViS corpus, the speech is recorded in four ways: (1) HQ (high-quality recording), (2) TEL (telephone recording), (3) MOBHQ (mobile high-quality recording), and (4) MOBLQ (mobile low-quality recording). For feature extraction, the author chose MFCC as a feature. To improve the performance of the system, Vq features are added with MFCC. For Vq, four methods are used, that is (1) F0 (Fundamental frequency), (2) CPP (Cepstral Peak Prominence), (3) HNR (Harmonic to Noise Ratio), and (4) H1-A1; H1-A2; H1-A3. All these four features are also called the vocal characteristic of humans. These four features with MFCC are extracted from speech. However, as transmission quality degraded (HQ > TEL > MOBHQ > MOBLQ), the contribution of Vq to system performance became much more impressive. In their work, EER (Equal Error Rate) is 2.85% using MFCC alone and EER is 0.09% using MFCC combined with Vq.

Zhang et al. [37] explored the effectiveness of the formant trajectory technique applied to tokens of the standard Chinese triphthong /iau/. Chinese token is extracted from speech and for extracting features from tokens acoustic-phonetic approach is used. The test scores from the acoustic-phonetic and automatic systems were fused using logistic-regression fusion. Another similar work is carried out in [38], in which Morrison et al. [38] proposed a forensic-voice-comparison system for standard Chinese monophthongs /i/,/e/, and /a/.

TABLE II. ANALYSIS OF SPEAKER RECOGNITION AND COMPARISON METHODS BASED ON TRADITIONAL APPROACH

Researchers	Dataset	No. of speakers	Feature Extraction	Model	Text-type	System-type	Accuracy or EER (in %)	
Reynolds and Rose (1995) [7]	KING speech database	49	Mel-frequency filter bank for short utterance	GMM	TI	SI	AC: 96.3	
Reynolds(1995) [30]	TIMIT, NTIMIT, Switch-board, YOHO	TIMIT: 630 NTIMIT: 630, Switchboard: 113 YOHO: NA	Mel-scale Cepstral	GMM	NA	SI	TIMIT	AC: 99.5
							NTIMIT	AC: 60.7
							switchboard	AC: 82.8
							YOHO	AC: NA
						SV	TIMIT	EER:0.24
							NTIMIT	EER:7.19
							switchboard	EER:5.15
							YOHO	EER:0.51
Adami et al. (2001) [31]	Random	30	LPCC, FOR, PIT, LPC	MLP	NA	SI	LPCC- AC: 100	
Rabha et al. (2003) [32]	Random	10	LPC/ Cepstral	SVD-based algorithm	TI	SI	Clean speech- AC: 99.5	
							Noisy speech- AC: 77.5	
Shahin (2009) [33]	Non professional database	40(20 male + 20 female)	LFPC	HMM	TD	SI	AC: 61.4	
				CHMM			AC: 66.4	
				SPHMM			AC: 69.1	
Revathi et al. (2009) [8]	TIMIT	50	MF-PLP	Iterative clustering approach	TI	SR	AC: 91.0	
			PLP				AC: 88	
Chakroborty and saha (2009) [9]	Dataset	>130	MFCC, IMFCC	GMM	TI	SI	TF	GF
	YOHO						AC: 97.26	AC: 97.42
	POLYCOST						AC: 81.16	AC: 82.76
Saeidi et al. (2010) [10]	Speech separation challenge corpus	34(18 male + 16 female)	MFCC	GMM-UBM	TI	SI	AC: 97.0	
Tolba et al. (2011) [34]	Arabic speaker	10	MFCC	CHMM	TI	SI	AC: 80	
Ajmera et al. (2011) [35]	TIMIT	630	Spectrographic acoustic feature	DCT	TI	SI	AC: 96.69	
	SGGS	151					AC: 98.41	
Krishnamoorthy et al. (2011) [36]	TIMIT	100	MFCC	GMM-UBM	TI	SR	AC: 80	
Zhang et al. (2011) [37]	Chinese female speakers	60	Formant, MFCC	GMM-UBM	TD	VC	AC: NA	
Morrison et al. (2011) [38]	Chinese male speakers	64	Formant	Likelihood ratio	TD	VC	AC: NA	
Cardoso et al. (2019) [1]	DyViS corpus	97	HNR, CPP, f0, formant, MFCC	GMM-UBM	NA	VC	EER: 0.09	

[a.] Result: EER-Equal Error Rate, AC-Accuracy, TF-Triangular Filter, GF-Gaussian Filter [b.] Text-type: TI-Text Independent, TD-Text Dependent, [c.] feature extraction: FOR-Formant, PIT-Pitch, [d.] System-type: SV-Speaker Verification, SI-Speaker Identification, SR- Speaker Recognition.

B. Survey on Speaker Recognition (Identification and Verification) and Voice Comparison System based on Deep learning-based approach

Deep learning is becoming an interesting and powerful method of machine learning. Deep learning strategies have been effective in recognizing speakers. Few of the researchers have worked on voice domains using deep learning-based methods. Table III presents a survey on the deep learning-based approach.

Variani et al. [39] proposed the DNN (Deep Neural Network) based method of speaker verification. The DNN is trained to classify speakers with acoustic characteristics at the frame level. The average features of these speakers, or called d-vector features, are then used to verify other speakers. Lukic et al. [12] proposed a new approach for optimizing the pipeline of speaker identification and evaluated on TIMIT dataset. The authors have used the Convolution Neural Network (CNN) on spectrograms to learn speaker-specific characteristics from a rich representation of acoustic sources. The CNN consists of several such layers of convolution that apply a wide range of filters to subsequent small local input sections (e.g. a 3X3-area, which is then repeated throughout the entire input space). That convolution layer is followed by a max-pooling layer, which produces a lower resolution version of the activations of the convolution layer by removing the total filter activation from e.g. a 2X2 window. At the end, fully connected layers eventually integrate all outputs of the last max-pooling layer to classify speakers.

Plchot et al. [40] presented a DNN-based auto-encoder (DAE) of speaker recognition systems for microphones and noisy information. The function of auto-encoder is to enhance the speech signal (i.e., to de-noise and de-reverberate). Plchot et al. [40] concluded that an audio enhancement method offers good compensation for distortions caused by reverberation, whereas multi-condition training can very well handle the distortion caused by additive noise. Torfi et al. [42] proposed a novel method for text-independent speaker verification using 3D convolution neural network (3D-CNN) architecture. In their work, the authors proposed an adaptive learning feature by using the 3D-CNN to directly create speaker model. In the process, an identical number of spoken utterances per speaker are flowed into the network.

TABLE III. ANALYSIS OF SPEAKER RECOGNITION AND COMPARISON METHODS BASED ON DEEP LEARNING APPROACH

Researchers	Dataset		No. of speakers	Input	Model	Text-type	System type	Accuracy or EER (in %)
Variani et al. (2014) [39]	NA		646	Energy features of frame	DNN	TD	SV	EER : 2.00 (For 20 utterances)
Lukic et al. (2016) [12]	TIMIT		630	Spectrogram of voice data	CNN	NA	SI	AC: 97
Plchot et al. (2016) [40]	PRISM	Fisher corpora	13916	MFCC, PNCC	DNN auto encoder	TD, TI	SR	NA
		Switch board	1991					
		SRE	2740					
Chung et al. (2017) [41]	Voxceleb		1251	Spectrogram	CNN	NA	SI	AC: 80.5
							SV	EER: 7.8
Torfi et al. (2018) [42]	WVU-Multimodal 2013		1083	Frame-wise MFEC	3D-CNN	TI	SV	EER: 21.1
Muckenhirn et al. (2018) [43]	Voxforge		Selected 300	Raw speech data	CNN, MLP	NA	SI	EER: 1.18
							SV	EER: 1.20
Dhakal et al. (2019) [44]	ELSDSR		22	Statistical, Gabor feature and CNN based	SVM	NA	SR	AC: 98.07
					RF			AC: 99.41
					DNN			AC: 98.14

[a.]Result: EER-Equal Error Rate, AC-Accuracy[b.] System-type: SV-Speaker Verification, SI-Speaker Identification, SR-Speaker Recognition [c.] Text-type: TD-Text Dependent, TI-Text Independent

C. Analysis of different datasets

In Table IV, we analyze widely used datasets such as TIMIT [20], NTIMIT [21], Switchboard [45], YOHO [46], VoxCeleb [22], ELSDSR [47], POLYCOST [48], ICSI Meeting speech [49], and 2010 NIST SRE [50]. For analyzing datasets we use attributes like a number of subjects, utterances, types of speech, sample rate, dataset size, and application. The TIMIT [20] corpus (440 MB) is created to provide speech data for acoustic-phonetic studies and automated speech recognition systems. VoxCeleb [22] is a dataset for the recognition of speakers on a large scale, which is prepared from celebrities' YouTube videos. Most of the data [22] are gender-balanced (males are 55%). The videos include a range of backgrounds, professions, age, and gender.

D. Analysis of traditional and deep learning-based approach

Table V provides a comparison of the traditional and deep learning-based models. From an analysis, we can state that the traditional methods take a lot of time for feature extraction because traditional approaches measure frame-wise features such as fundamental frequency, formant, pitch, etc. Instead of extracting features manually, many automatic feature extraction techniques such as MFCC [13] and LPCC [15] are available. Traditional approaches are usually two-step procedures: first, calculate the feature (e.g., MFCC) and then feed them into the classifier (e.g., GMM, HMM, and VQ). However, in a deep learning method, we give voiceprint images directly as input to the model. For example, a spectrogram or voiceprint was used as an input to CNN by Lukic et al. in [12]. CNN directly learn from the input spectrogram. The main advantage of using any deep learning-based system is that the system is fully automatic. In a deep learning-based approach, CNN is perfect for classification. However, for comparison, Siamese NN is one of the popular approaches as compared to CNN because Siamese NN performs well for a limited dataset. In Siamese NN, we can use CNN as a sub-network as a feature vector generator.

We emphasize the following points related to CNN and Siamese NN architecture:

- CNN is good for classification problems while Siamese NN is good for comparison problems.
- Even if the training is less, Siamese NN can estimate well as compared to CNN.

IV. CONCLUSION

This paper focused on an elaborated survey of two useful voice processing operations: voice recognition and voice comparison. Voice comparison can become a very important task in the Human Machine Interface based systems. Before presenting the survey and analysis, this paper explained the essential concepts of speaker identification, speaker verification, and speaker comparison. Furthermore, the paper also presented a whole pipeline of voice comparison with enough details. In the survey, the paper studied and analyzed both traditional and deep learning-based approaches for speaker recognition and voice comparison and suggested the use of deep learning-based approaches for the voice processing domain. Furthermore, the paper also surveyed and presented various datasets used for automated voice processing. At the end, the suitability of Siamese NN combined with CNN, which is popular for the classification problems, for voice comparison is discussed.

TABLE IV. ANALYSIS OF DATASETS

Datasets	# Subject (speaker)	# Utterances	Types of speech	Sample rate (in HZ)	Dataset Size	Application
TIMIT [20]	630	6300	Microphone speech	16000	440 MB	Speaker identification and verification
NTIMIT [21]	630	6300	Telephone speech	16000	(25200 files)	Speaker identification and verification
Switchboard [45]	3114	33039	Telephone talking	8000	NA	Speaker identification
YOHO [46]	NA	NA	Microphone speech	8000	1500 MB	Speaker verification
VoxCeleb [22]	1251	≈ 100000	From YouTube videos	NA	150 MB	Speaker classification
ELSEDSR [47]	22	198	MARANTZ PMD670 recording speech	16000	NA	Speaker identification and verification
POLYCOST [48]	131	> 1285	Telephone speech	8000	≈ 1246 MB	Speaker identification and verification
ICSI Meeting speech [49]	53	922	Microphone conversation	16000	NA	Speaker segmentation
2010 NIST SRE [50]	> 2000	NA	Microphone and telephone speech	8000	NA	Speaker identification

TABLE V. ANALYSIS OF SPEAKER RECOGNITION AND VOICE COMPARISON MODELS

Approaches	Model	Advantages	Disadvantages
Traditional Approaches	HMM	● Provide good result when system is text-dependent, ● High computation burden in pattern recognition	● Not suitable for text-independent voice comparison system, ● Suitable for small speech recording
	VQ	● Easy to use, ● For pattern recognition, VQ have low computational burden then HMM	● Performance is degraded when recording of speaker is too large ● Slow generation of code book
	GMM	● The modeling of mixture is very versatile, ● It is a probabilistic approach to achieve a fuzzy observation classification	● Computationally costly if there is a huge number of distributions and need large datasets
Deep learning-based approach	CNN	● Fully automatic ● Easy model construction involving fewer formal statistics ● Capacity to capture non-linearity between predictors-results	● Due to the complexity of the model structure, prone to over-fitting
	3D-CNN	● Offer direct modeling of speaker	● Required optimized structure
	DAE	● It include de-noising	● Problem of over-fitting
	Siamese NN	● Required small training, ● Easy to label, ● More Sophisticated to irregularity in class	● Both sub-network are required to calculate same hyper parameter

References

[1] A. Cardoso, P. Foulkes, J.P. French, A.J. Gully, P.T. Harrison, and V. Hughes, "Forensic voice comparison using long-term acoustic measures of voice quality," InProceedings of the 19th International Congress of Phonetic Sciences (ICPhS) 2019 Feb 12. York.

[2] K. Kolhatkar, M. Kolte, and J. Lele, "Implementation of pitch detection algorithms for pathological voices," In2016 International Conference on Inventive Computation Technologies (ICICT) 2016 Aug 26 (Vol. 1, pp. 1-5). IEEE.

[3] H. Lee, P. Pham, Y. Largman, and A.Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," InAdvances in neural information processing systems 2009 (pp. 1096-1104).

[4] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," In2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014 May 4 (pp. 1695-1699). IEEE.

[5] Z. Saquib, N. Salam, R.P. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," InSignal Processing and Multimedia 2010 Dec 13 (pp. 134-145). Springer, Berlin, Heidelberg.

[6] N. Singh, A. Agrawal, and R.A. Khan, "Automatic speaker recognition: current approaches and progress in last six decades," in last six decades. Global J Enterp Inf Syst. 2017 Jul 1;9(3):45-52.

[7] D.A. Reynolds, and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE transactions on speech and audio processing. 1995 Jan;3(1):72-83.

[8] A. Revathi, R. Ganapathy, and Y.Venkataramani, "Text independent speaker recognition and speaker independent speech recognition using iterative clustering approach," International Journal of Computer science & Information Technology (IJCSIT) 1.2 (2009): 30-42.

[9] S. Chakroborty, and G. Saha, "Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter," International Journal of Signal Processing. 2009 Nov 26;5(1):11-9.R.

[10] R Saeidi, Mowlaee, T. Kinnunen, Z.H. Tan, M.G. Christensen, S.H. Jensen, and P. Franti, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," In2010 20th International Conference on Pattern Recognition 2010 Aug 23 (pp. 4565-4568). IEEE.

[11] G.S. Morrison, and W.C. Thompson, "Assessing the admissibility of a new generation of forensic voice comparison testimony," Colum. Sci. & Tech. L. Rev. 18 (2016): 326.

[12] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," In2016 IEEE 26th international workshop on machine learning for signal processing (MLSP) 2016 Sep 13 (pp. 1-6). IEEE.

[13] M.A. Hossan, S. Memon, and M.A. Gregory, "A novel approach for MFCC feature extraction," In2010 4th International Conference on Signal Processing and Communication Systems 2010 Dec 13 (pp. 1-5). IEEE.

[14] L.R. Rabiner, and M.R.Sambur , "Voiced-unvoiced-silence detection using Itakura LPC distance measure," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., May 1977, pp. 323–326.

[15] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of LPCC and MFCC," In2010 IEEE International Conference on Intelligent Computing and Intelligent Systems 2010 Oct 29 (Vol. 3, pp. 765-767). IEEE.

[16] A.M. Noll, "Cepstrum pitch determination," The journal of the acoustical society of America 41.2 (1967): 293-309.

[17] T. Kinnunen, and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," In2013 IEEE international conference on acoustics, speech and signal processing 2013 May 26 (pp. 7229-7233). IEEE.

[18] J.A. Haigh, and J.S. Mason, "A voice activity detector based on cepstral analysis," InEurospeech 1993 Sep (Vol. 9, pp. 1103-1106).

[19] B. Atal, and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing 24.3 (1976): 201-212

[20] J.S. Garofolo, "TIMIT Acoustic-Phonetic Continuous SpeechCorpus," LDC93S1: Linguistic Data Consortium, 1993. 1993.

[21] Fisher, and M. William, "NTIMIT LDC93S2," Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[22] J.S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition INTERSPEECH," arXiv preprint arXiv:1806.05622. 2018 Jun.

[23] M. Jalil, F.A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," In2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE) 2013 May 9 (pp. 208-212). IEEE.

[24] N.N. Lokhande, N.S. Nehe, and P.S. Vikhe, "Voice activity detection algorithm for speech recognition applications," InIJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012), vol. iccia 2012 Mar (No. 6, pp. 1-4).

[25] J. Stegmann, and G. Schroder, "Robust voice-activity detection based on the wavelet transform," In 1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding 1997 Sep 7 (pp. 99-100). IEEE.

[26] Q. Wang, C. Downey, L. Wan, P.A Mansfield, and I.L. Moreno, "Speaker diarization with lstm," In2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018 Apr 15 (pp. 5239-5243). IEEE.

[27] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," IEEE Transactions on Audio, Speech, and Language Processing 20.2 (2012): 356-370.

[28] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," InAdvances in neural information processing systems 1994 (pp. 737-744).

[29] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," ICML deep learning workshop. Vol. 2. 2015.

[30] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech communication 17.1-2 (1995): 91-108.

[31] A.G. Adami, and D.A. Barone, "A speaker identification system using a model of artificial neural networks for an elevator application," Information Sciences 138.1-4 (2001): 1-5.

[32] R.W. Aldhaheri, and F.E. Al-Saadi, "Text-independent speaker identification in noisy environment using singular value decomposition," InFourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint 2003 Dec 15 (Vol. 3, pp. 1624-1628). IEEE.

[33] I. Shahin, "Speaker identification in emotional environments," (2009): 41-46.

[34] H. Tolba, "A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach," Alexandria Engineering Journal 50.1 (2011): 43-47.

[35] P.K. Ajmera, D.V. Jadhav, and R.S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram," Pattern Recognition 44.10-11 (2011): 2749-2759.

[36] P. Krishnamoorthy, H.S. Jayanna, and S.M. Prasanna, "Speaker recognition under limited data condition by noise addition," Expert Systems with Applications 38.10 (2011): 13487-13490.

[37] C. Zhang, G.S. Morrison, and T. Thiruvaran, "Forensic Voice Comparison Using Chinese/iau/," in InICPhS 2011 Aug 17 (pp. 2280-2283).

[38] G.S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system," Forensic science international 208.1-3 (2011): 59-65.

[39] E. Variani, X. Lei, E. McDermott, I.L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," In2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014 May 4 (pp. 4052-4056). IEEE.

[40] O. Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with DNN autoencoder for speaker recognition," In2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 Mar 20 (pp. 5090-5094). IEEE.

[41] A. Nagrani, J.S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612. 2017 Jun 26.

[42] A. Torfi, J. Dawson, and N.M. Nasrabadi, "Text-independent speaker verification using 3d convolutional neural networks," In2018 IEEE International Conference on Multimedia and Expo (ICME) 2018 Jul 23 (pp. 1-6). IEEE.

[43] H. Muckenhirn, M.M. Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," In2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018 Apr 15 (pp. 4884-4888). IEEE.

[44] P. Dhakal, P. Damacharla, A.Y. Javaid, and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," Machine Learning and Knowledge Extraction. 2019 Mar;1(1):504-20.

[45] J. Godfrey, and E. Holliman, "Switchboard-1 Release 2 LDC97S62," DVD. Philadelphia: Linguistic Data Consortium. 1993.

[46] J. Campbell, and A. Higgins, "YOHO speaker verification. Linguistic Data Consortium," Philadelphia. 1994.

[47] L. Feng, "Speaker recognition," (Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark).

[48] J. Hennebert, H. Melin, D. Petrovska, and D. Genoud, " POLYCOST: a telephone-speech database for speaker recognition," Speech communication. 2000 Jun 1;31(2-3):265-70.

[49] Janin, Adam, et al. ICSI Meeting Speech LDC2004S02. Web Download. Philadelphia: Linguistic Data Consortium, 2004

[50] Greenberg, Craig, et al. 2010 NIST Speaker Recognition Evaluation Test Set LDC2017S06. Hard Drive. Philadelphia: Linguistic Data Consortium, 2017.

[51] R. Tucker, "Voice activity detection using a periodicity measure," IEE Proceedings I (Communications, Speech and Vision). 1992 Aug 1;139(4):377-80.

[52] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research," International Journal of Speech, Language & the Law. 2009 Jun 1;16(1).