Speaker Recognition for Robotic Control via an IoT Device

Zhanibek Kozhirbayev[1,3,*], Berat A. Erol[2,*], Altynbek Sharipbay[1], PhD and Mo Jamshidi[2], PhD

***Abstract*—Speaker Recognition is considered as one of the primary tasks in speech processing. Nowadays, the speaker identification method has been extensively appealing for its broad application in many fields, such as smart environments, securing the cyber-physical systems, speech communications, and robotic controls. Researchers are targeting to perform an effective method that makes it possible to obtain the recognition ability that is close to the hearing of human. In order to get high accuracy, challenges of large-scale applications of speaker identification are overcome through applying techniques not only traditional models based on the GMM, but also deep learning methods. Aiming at effectively dealing with this challenge, in this paper, we present a novel model to increase the recognition accuracy of the short utterance speaker recognition system. We developed a technique to train a Neural Network (NN) on the extracted Mel-Frequency Cepstral Coefficient (MFCC) features from audio samples. Therefore, the recognition system gains the significant accuracy. The model was trained using open-source high-level neural networks API Keras.**

***Index Terms*—Human robot interactions, Speaker identification and recognition, Neural networks, Internet of robotic things, Amazon Echo.**

I. Introduction

Automatic Speaker Recognition systems are able to be used for a practical purpose for identifying a speaker; enable automatic voice control over Human-Robot Interaction (HRI), such as controlling an assistive robot. Especially, smart digital assistance devices, such as Amazon Echo platforms with Alexa and Google Home, has helped to increase the number of applications on this regard and create a new domain that focusing on to control a system by vocal interactions while monitoring the user demands spontaneously. Speech can be considered as a biometric characteristic of a speaker that can be gained with or without the speakers knowledge. A general speaker recognition system consists of two components such as feature extraction and classification.

There might be serious issues while gathering suitable speech data in some application cases. The recent speaker identification systems can be reasonably successful only when the training samples are long enough. The accuracy decreases in a short environment scenario. More precisely, short samples mean the inputs with insufficient acoustic features to the model. However, deep learning techniques can be effectively used by virtue of its deep feature learning possibility. Nowadays, they are widely applied in speech processing tasks such speech recognition [1], spoken term detection [2] and automatic language identification [3].

There are two types of speaker recognition systems such as text dependent or text independent. The former system utilizes a fixed utterance for training and testing a person, whereas the later one does not employ a fixed phrase for any cases.

The usage of speaker recognition is rapidly increasing, some of them are:

- Access Control: confidential information or control of services can be accessed through legitimated persons voice.
- Online Transactions: person's voice features can be used as biometric information to the extra-layer of security.
- Law Enforcement: forensic analysis can be performed using speaker identification.
- Speech Data Management: speaker diarization systems are able to label speakers in meeting video/audio recordings.
- Multimedia and personalization: speaker identification technique is able to label soundtracks with metadata about a singer.

In more details, it can achieve the strong anti-interference goal by excavating a larger number of voiceprint features. However, the training process requires a large number of samples and the characteristic specificity is not obvious, which results the deep learning cannot work directly. Considering the instability of the acquisition of short utterance features and the possible shortage of training samples, we focus on how to design the voiceprint model and deep-learning model for the short utterance recognition, which can effectively overcome the shortcomings of less training samples and susceptible to interference. To be precise, deep learning can attain the robust anti-interference aim by extracting a larger quantity of sound features. Therefore, model building relies on the quantity of dataset. Taking into account the unsteadiness of the acquisition of short samples features, we concentrated on how to extract an adequate training and testing sets as well as design NN.

Speaker identification system defines who among a set of known speakers is making available for use the given utterance as illustrated Fig. 1. Speaker sound features are obtained from the speech data, and decided in comparison with models built using special techniques on speech samples of the engaged

[1]: Faculty of Information Technologies, L.N.Gumilyov Eurasian National University, Astana, Kazakhstan.
[2]: Autonomous Control Engineering Laboratories, The University of Texas at San Antonio, San Antonio, TX 78249 USA.
[3]: National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan.
∗: Contributed equally on this work.
Email: zhanibek.kozhirbayev@nu.edu.kz, berat.erol@utsa.edu, sharalt@enu.kz and mo.jamshidi@utsa.edu

978-1-5323-7791-4/2018/ ©2018 TSI Press, Inc

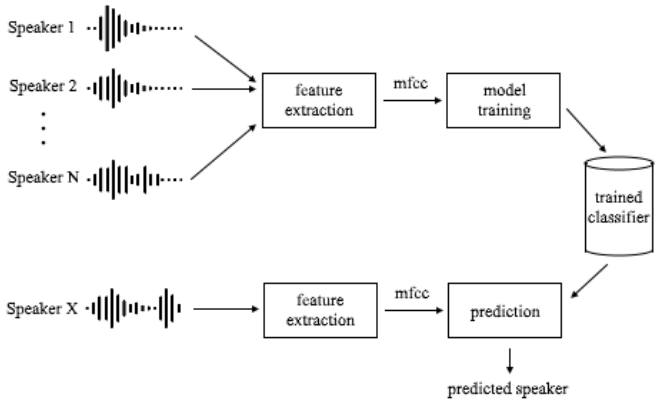

Fig. 1. Speaker Recognition System

speakers. The system identifies the legitimate speaker if his voice features are got high likelihood score. As a rule, the model creates a probabilistic score and the model that creates the maximum likelihood score is chosen.

This paper is organized as follows: Section II presents an overview of various techniques for speaker identification applications. More precisely, it describes methods on the text-dependent and text-independent as well as modern approaches based on Neural Networks. The dataset descriptions and features extraction and configuration parameters are described in Section III. Section IV presents the NN, as a speaker identification system as well as its conceptual architecture. Proposed system for voice activated robotic control via user re-dentification is described in Section V. Experimental details and obtained results of the speaker identification task are stated in Section VI. Summary of the performed experiments and areas of further research are given in Section VI.

II. Related Work

Significant advancement has been made in speaker recognition research. This Section presents brief review about speaker identification techniques. The research conducted by [4] in 1960 is considered as one of the basements in speaker recognition. [4] evaluated a likeness by utilizing filter banks and correlating two digital spectrograms. Averaged autocorrelation and time domain methods are used to improve speaker recognition [5], [6]. Texas Instruments developed an automatic speaker verification system in 1970. New approaches such as HMM methods as a substitute to template matching [7], melfrequency cepstral coefficients as features [8], vector quantization (VQ) and HMM methods [9], Gaussian mixture models (GMM) [10] are examined and employed to produce more efficient speaker recognition.

From 2000s, researchers investigated new directions by supplementing higher level information such as ideolectic features [10], temporal trajectories of fundamental frequencies and short term energies to segment and label speech [11], pronunciation models, prosodic dynamics, pitch and duration features, phone streams and conversational interactions [12].

A combination of GMMs, SVMs and NGrams is applied to the speaker features obtained from short time acoustics, pitch, duration, prosodic behavior, phoneme and phone usage [13], whereas phonetically designed GMMs and speaker adaptive modeling were utilized to model MFCC in [14]. The approach presented by [14] was tested on YOHO and Mercury speech databases. GMM, GMM-Universal background model (GMM-UBM) and SVM were utilized in research by [15] and their methods can precisely depict the target speaker. This research method was conducted on the Mercury and Orion databases including 44 speakers. [16] applied a combination of a GMM system and a syllable-based HMM to the NTT dataset. A smoothed fundamental frequency contour at different time scales was utilized by [17] as the speech features extracted from SRI and NIST2001 databases. [18] modeled a NN to classify 6 speakers using a shortened TIMIT database. [18] used Formant trajectories and gender as features.

Nowadays, Joint Factor Analysis (JFA) [19] as well as ivector models [20] have also been offered to speaker recognition. Scientists are integrating speaker identification mechanisms to user authorization [21], data privacy [22] and offer similar approaches for cloud computing architectures, performance and security [23]–[25]. Nevertheless, the techniques referred to above suffers from the short language conditions. For instance, it is hard to obtain effective result by applying the GMM method on short samples. With the progress of neural networks technology, applying machine learning to speech features may overcome the speaker recognition obstacles for short samples. [26] used the convolutional neural networks (CNN) and got remarkable results in short speaker samples recognition problems.

In this work, we focused on a neural networks speaker recognition system, to achieve corresponding performance observed in above mentioned techniques. However, we used two different speech datasets and fused their scores to have a better performing system. However, there are a number of moments that distinguish this work from previous research: database types, database sizes and multilayer NN.

III. Dataset

In order to train and test the system for speaker recognition, we can use prerecorded speech databases which are already used in speech recognition. There are many of them and some of them need a license and some are open source. In this experiment, TIMIT and VoxForge are utilized to train and test the speaker recognition system.

A. TIMIT

The TIMIT (Texas Instruments Massachusetts Institute of Technology) is a dataset s of read speech is originated to ensure corpus for acoustic-phonetic research as well as for the design and assessment of ASR systems [27]. It includes recordings of 630 speakers and each of them read 10 sentences. The distribution of speakers varies depending on gender as 70% of men and 30% of women and on dialect regions in America as New England, Northern, North Midland, South

TABLE I
DIALECT DISTRIBUTION OF SPEAKERS

Dialect Region	#Male	#Female	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
Total	438 (70%)	192 (30%)	630 (100%)

TABLE II
TIMIT SPEECH MATERIAL

Sentence Type	#of Sentences	#of Speakers	Total	#of Sentences
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342	638	6300	10

Midland, Southern, New York City, Western and Army Brat. The distribution of the dataset can be seen in the Table I. The speech is planned to obtain an intensive phonetic corpus that includes diverse sentences as shown in Table II.

B. VoxForge

The VoxForge database provides a set of transcribed speech to be utilized in Open Source Speech Recognition Engines [28]. It includes multilingual speech samples such as 25420 English samples, 4021 French samples and 2963 German samples. Each speech sample is approximately 5 seconds and combined with various metadata. The quality of the speech samples are low and it changes notably among samples since they were recorded by users with their own microphones.

C. Feature extraction and configuration parameters

MFCC (Mel-Frequency Cepstral Coefficient) is a way to represent the short-term power spectrum of an audio signal. It is designed on the basis of a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. MFCC is mainly extensively utilized features in Automatic Speech Recognition (ASR), Language Identification and it can also be employed in Speaker Recognition purpose as well. As an input to our system, Mel-Frequency Cepstral Coefficients were extracted from audio signals convolved with 32ms Hamming windows shifted every 16ms. We used the following parameters to extract the acoustic features from raw audio data, shown in Table III.

IV. USING NN TO MODEL SPEAKER RECOGNITION SYSTEM

This section presents the conceptual architecture of the effective speaker identification system on the basis of NN. Experiments are conducted on the machine with 32 Gb memory and 8 cores. The software utilized to develop the NN model is Keras [29] on the basis of Tensorflow; which is a high-level straightforward Python library for developing as well as evaluating deep learning models. It wraps the efficient numerical computation libraries Theano and TensorFlow and offers consistent and simple APIs.

TABLE III
MFCC PARAMETERS

Common parameters	MFCC parameters	LPC Parameters
Frame size: 32ms	number of cepstral coefficient: 15	number of coefficient: 23
Frame shift: 16ms	number of filter banks: 55	
Preemphasis coefficient: 0.95	maximal frequency of the filter bank: 6000	

The system consists of two hidden layers on top of the input layer. We used Feedforward Neural Network as a hidden layer each having 300 units. This was the maximum size that could fit our hardware constraints.

The softmax layer was added to the network as an output layer. It has the same number of units as the speakers in the training set. The softmax function is usually utilized in the output layer of network to solve a classification task.

To train the network, we exploited categorical cross-entropy which was optimized using stochastic gradient descent (SGD), initial learning rate of $1e-3$ and decay factor of $1e-4$. Additionally, a momentum with value 0.9 and Nesterov accelerated gradient were used to accelerate SGD.

In order to train the model and integrate it with robot control, we will create a dataset by recording several audio sequences while saying the same word repetitively, and then passing them through the framework, to obtain a dataset.

V. PROPOSED SYSTEM FOR VOICE ACTIVATED CONTROL VIA USER IDENTIFICATION

The important part of integrating an HRI aspect into a system is make the robot to learn. A key way to do this for a task assigned robot is to recognize the the user. An overview of the prototype system is provided in Figure 2. The smart environment system includes interfaces for voice, vision, cloud-based computation, and robotic platforms. In this section, preparation of robotic hardware and the control loop are discussed.

The assistive robot used in this system is a hybrid of an Unmanned Ground Vehicle (UGV) and a humanoid robot. The model of UGV utilized in this work is a Kobuki Turtlebot 2 research platform from YujinRobot. For the humanoid portion of the hybrid machine, the 3D printed Poppy humanoid torso is used which is the result of an open source platform. The hybrid robotic platform is compatible with the Robot Operating System (ROS). A camera mounted in the head is used to detect objects in the environment. The *kobuki* ROS package handles control of the base robot. A custom designed head unit was designed to support addition of a five-inch touchscreen LCD display (ODROID-VU5), monocular

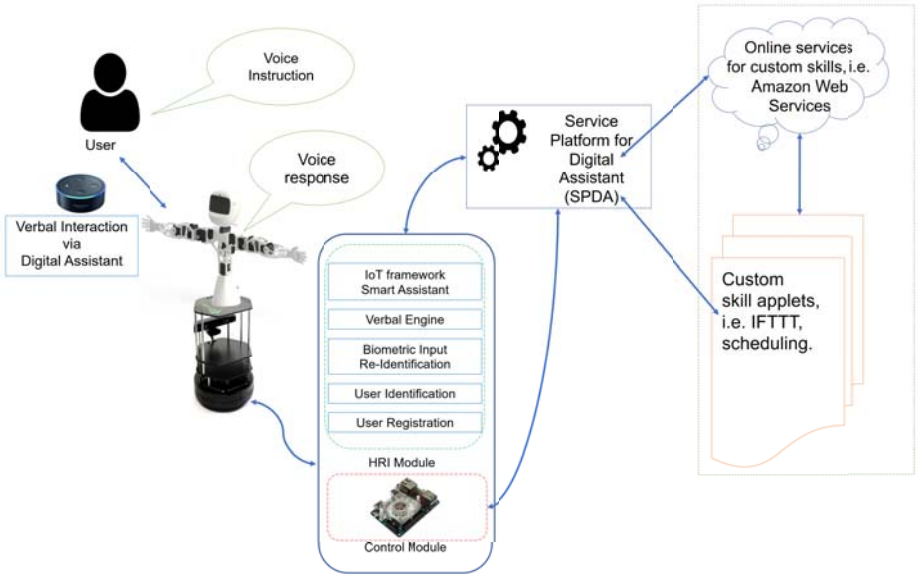

Fig. 2. Proposed system test-bed for user recognition and re-identification approach with digital assistant (IoT device) integration. The system illustration includes smart home assistive robot use case; a user-centered voice-activation control focused on high-level human-robot interactions.

camera, stereo speakers for synthesized auditory feedback, and a microphone for obtaining commands from the user.

The control loop starts with the user asking the robot to perform a task, basically uttering the smart assistant by calling its name, and asking for initiating the system for further requests. If user has registered previously as a legit user, then the device will provide a voice feedback since the user is already recognized by their voice. A voice assistance device combined with the computing service extracts the requested action from the user and provides the action plan to the robot. The requested actions are decoded by the robot, where any action that has already been completed or can be extracted from its local database are reported immediately to the system.

VI. Results and Discussion

In this work we examine slightly different architectures and configurations to speaker recognition. Also there are a number of moments that distinguish this work from previous research: database types and database sizes. We performed two types of experiments. For each type of experiment, we examined the performance of our system for number of speakers. The speakers were selected from given datasets TIMIT and VoxForge and correspond directly to segments of the talk. Therefore, the training set is built from the selected speakers, whereas the test set contains of unseen speech samples of the selected speakers.

- 10 speakers: 5 male and 5 female speakers from randomly selected dialect areas for the TIMIT, whereas 10 randomly selected speakers chosen for the VoxForge.
- 20 speakers: In this setup, we selected 10 male and 10 female speakers from randomly selected dialect areas for the TIMIT dataset, whereas 20 randomly selected speakers chosen for the VoxForge.

TABLE IV
Size of the datasets

Dataset	# speech records per speaker	Average duration of records per speaker in seconds	Total duration of dataset in seconds
TIMIT 10 speakers	10	30.3	302.9
TIMIT 20 speakers	10	31.55	631.4
VoxForge 10 speakers	10	58.9	597.6
VoxForge 20 speakers	10	57.1	1141.9

Table IV shows the number of data used in the experiments. More precisly, it explains the speech records per speaker, average duration of the speech records per speaker in seconds and total duration of the dataset in seconds.

The performance of NN system regarding accuracy and loss on the test set can be seen in Table V. The Figures from 3 till 6 show these metrics on the train and evaluation sets for the used datasets with different numbers of speakers. We highlighted that we wanted to explore the performance of the system in terms of applying different datasets.

Experiment results on the VoxForge datasets show reasonble performnce rather than on the TIMIT datasets. The speech voiceprints from the TIMIT are recorded in clean conditions, therefore, it creates difficulties to thoroughly distinguish speakers. Whereas the quality of the VoxForge recordings leaves much to be desired. Moreover, there is a significant

TABLE V
RESULTS FOR PROPOSED SYSTEM WITH NN IMPLEMENTATION

Dataset	Accuracy(%)	Loss (%)
VoxForge 10 speakers	96	13
VoxForge 20 speakers	90	31
TIMIT 10 speakers	89	39
TIMIT 20 speakers	76	87

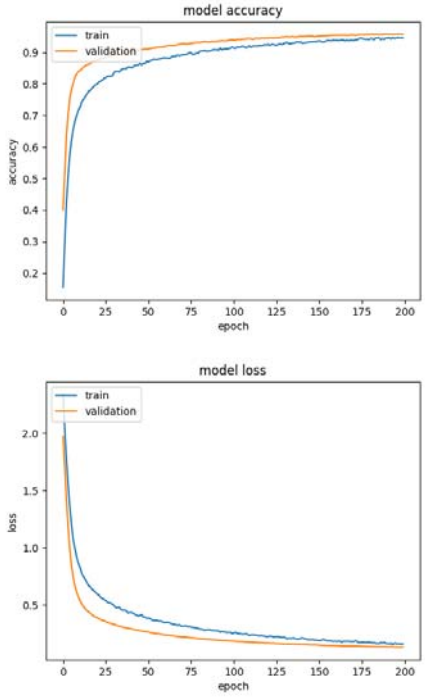

Fig. 3. Model accuracy(left) and loss(right) on VoxForge for 10 speakers.

difference in an average duration of the speech recordings per speaker.

VII. CONCLUSION

As seen in the literature, the number of studies of new applications on Human-Robot Interactions and Internet of Things devices have been increasing dramatically. Due to their improved connectivity, mobility and interoperability, applications of new approaches for IoT and smart monitoring devices on smart environments brought more prospects on effective solutions in both HRI and robotic control domains.

In this paper, we presented a system that identifies, recognize and re-identifies it's users by their voice. The results are applicable to our proposed system that controls a robotic platform with an integrated IoT device, a digital smart assistant. The purpose was creating a system for home and assistive robotics that interacts with users and perform their requests, while collecting, computing and sharing the gathered data as a part of the system.

The results supports that the components of the systems can work as planned and fulfills the human-in-the-loop scenarios. This is a positive outcome for the research by not relying

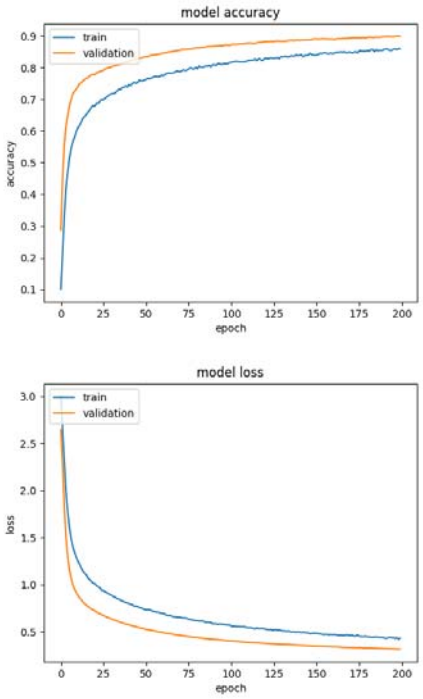

Fig. 4. Model accuracy(left) and loss(right) on VoxForge for 20 speakers.

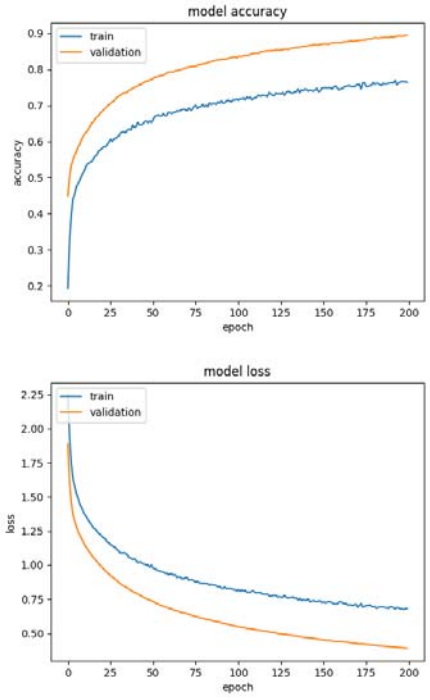

Fig. 5. Model accuracy(left) and loss(right) on TIMIT for 10 speakers.

on processing power as used to be expected; however, being able to perform identification and recognition of users with enough accuracy. The voice recognition analysis suggested by our work clearly shows how digital assistants can be used in

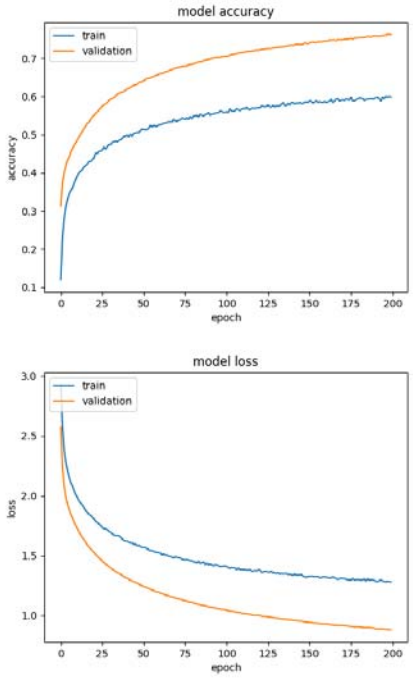

Fig. 6. Model accuracy(left) and loss(right) on TIMIT for 20 speakers.

a smart environment to identify users by their voices and uses the advantages of cloud computing back-end from the well-known service providers. This is more useful in a HRI situation where the robots are designed to interact with a limited set of people, e.g. an assistive robot, or heterogeneous robotic platforms with and IoT module developed for smart facilities based on Industry 4.0 perspectives.

ACKNOWLEDGMENT

This work has been conducted under the Project "Building a Kazakh Dependency Treebank" funded by Nazarbayev University within contract 144-2018/010-2018 dated 12.02.2018.

REFERENCES

[1] M. Karabalayeva, Z. Yessenbayev, and Z. Kozhirbayev, "Spoken term detection for kazakh language," in *Computer Processing of Turkic Languages TurkLang 2017, 5th International Conference on*, 2017, pp. 113–129.

[2] Z. Kozhirbayev, M. Karabalayeva, and Z. Yessenbayev, "Spoken term detection for kazakh language," in *Computer Processing of Turkic Languages TurkLang 2016, 4th International Conference on*, 2016, pp. 47–52.

[3] Z. Kozhirbayev, Z. Yessenbayev, and M. Karabalayeva, "Kazakh and russian languages identification using long short-term memory recurrent neural networks," in *Application of Information and Communication Technologies (AICT), 2017 IEEE 11th International Conference on*. IEEE, 2017, pp. 342–347.

[4] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 354–358, 1963.

[5] P. Bricker, R. Gnanadesikan, M. Mathews, S. Pruzansky, P. Tukey, K. Wachter, and J. Warner, "Statistical techniques for talker identification," *Bell Labs Technical Journal*, vol. 50, no. 4, pp. 1427–1454, 1971.

[6] B. Atal, "Text-independent speaker recognition," *The Journal of the Acoustical Society of America*, vol. 52, no. 1A, pp. 181–181, 1972.

[7] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker verification over long distance telephone lines," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 524–527.

[8] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[9] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using vq-distortion and discrete/continuous hmm's," *IEEE Transactions on speech and audio processing*, vol. 2, no. 3, pp. 456–459, 1994.

[10] T. J. Hazen, D. A. Jones, A. Park, L. C. Kukolich, and D. A. Reynolds, "Integration of speaker recognition into conversational spoken dialogue systems," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker-specific gmm with speaker adapted syllable-based hmm," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–81.

[12] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[13] F. Farahani, P. G. Georgiou, and S. S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–89.

[14] M. M. Tanabian, P. Tierney, and B. Azami, "Automatic speaker recognition with formant trajectory tracking using cart and neural networks," in *Electrical and Computer Engineering, 2005. Canadian Conference on*. IEEE, 2005, pp. 1225–1228.

[15] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[16] S. Gray and J. H. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 35–40.

[17] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3. IEEE, 1996, pp. 1784–1787.

[18] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using gaussian mixture models," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 343–346.

[19] R. J. Vogt, B. J. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," 2008.

[20] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[21] Y. Jiang, H. Song, R. Wang, M. Gu, J. Sun, and L. Sha, "Data-centered runtime verification of wireless medical cyber-physical system," *IEEE transactions on industrial informatics*, vol. 13, no. 4, pp. 1900–1909, 2017.

[22] P. Li, J. Li, Z. Huang, C.-Z. Gao, W.-B. Chen, and K. Chen, "Privacy-preserving outsourced classification in cloud computing," *Cluster Computing*, pp. 1–10, 2017.

[23] A. Sahba, R. Sahba, and W. M. Lin, "Improving ipc in simultaneous multi-threading (smt) processors by capping iq utilization according to dispatched memory instructions," in *2014 World Automation Congress (WAC)*, Aug 2014, pp. 893–899.

[24] A. Sahba and J. J. Prevost, "Hypercube based clusters in cloud computing," in *2016 World Automation Congress (WAC)*, July 2016, pp. 1–6.

[25] J. Li, Y. Zhang, X. Chen, and Y. Xiang, "Secure attribute-based data sharing for resource-limited users in cloud computing," *Computers & Security*, vol. 72, pp. 1–12, 2018.

[26] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "Gmm and cnn hybrid method for short utterance speaker recognition," *IEEE Transactions on Industrial Informatics*, 2018.

[27] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[28] K. MacLean, "Voxforge," *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Acedido em 2012]*.

[29] F. Chollet *et al.*, "Keras," 2015.