3-Step Speaker Identification Approach in Degraded Conditions

Sana Boujnah[1,3] , Xianfang Sun[2], David Marshall[2], Paul. Rosin[2] and Mohamed Lassaad Ammari[1]
[1] NOCCS, National Engineering School of Sousse, University of Sousse, Tunisia
[2] HFTC, Cardiff School of Computer Science, Cardiff University, Wales UK
[3] National Engineering School of Tunis, University of Tunis El Manar, Tunisia

***Abstract*—The human voice is a perfect source of data for person identification in many applications. With a growing need for security in various public places, voice biometrics seems to be a good solution, since voice records can be easily taken. This paper provides a brief overview of approaches used in speaker recognition. It also presents a novel approach for speaker recognition in degraded conditions in the context of smart homes. The proposed method includes a pre-processing phase, feature extraction phase and classification phase.The second phase is based on the extraction of the sound spectrum energy maxima of the speech sound called formant. The next step is the use of a well-known technique called Dynamic Time Warping to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. Complimentary steps are needed to complete the processing. The experiments are performed on a 1248-sample database to validate the proposed approach. The novel approach leads to better results regarding the state-of-the-art with an accuracy of 94.5%.**

I. Introduction

In the current technological context where the world is reduced to a small connected village [1], several new requirements have emerged and have even become vital. We quote, for example, the need for access control and the management of bank accounts or their highly confidential databases, through smart phones [2]. Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. The areas of speaker recognition applications are very large and varied. Actually there is no limit for them. It can be said that if an audio recording is involved, one or more branches of speaker recognition will be used. We can mention among these applications: financial applications, forensics, legal applications, security applications, applications for audio and video indexing, monitoring, smart homes [3]. The recognition of the speaker is one of the sub-disciplines of voice processing. It is divided into several sub-areas [4]: verification, grouping into speakers, and identification (we are interested in this sub-area in this paper). Indeed, speaker recognition is certainly not the least intrusive or the most reliable biometric. However, when the voice of a person is the only physical characteristic available, the biometric identity verification is possible only through the discipline of speaker recognition. The speech signal conveys many levels of information to the listener [4]. At the primary level, speech conveys a message via words. However, at other levels, speech conveys information about the language being spoken and the emotion, gender, and generally the identity of the speaker. While speech recognition aims to recognize the word spoken in speech, the goal of automatic speaker recognition systems is to extract, characterize and recognize the information in the speech signal conveying the speaker identity. In this paper a hybrid system composed of formant extraction and DTW method is proposed. Furthermore, a refinement process is introduced in order to enhance the proposed approach.

The experiments show that the new approach can produce better results than the state-of-the-art approaches. The remaining parts of this paper are organized as follows. Section 2 highlights the well-known approaches in the state-of-the-art. Section 3 describes the proposed method. Eventually, section 4 summarizes our research results and discusses potential future improvement as well as future directions. We will finish by giving the conclusion and the prospects.

II. Related Work

The general area of speaker recognition encompasses two more fundamental tasks, which are speaker identification and speaker recognition. Speaker identification is the task of determining who is talking from a set of known voices or speakers. The unknown person makes no identity claim, so the system must perform a l:N classification. Normally, it is assumed that the unknown voice must come from a fixed set of known speakers [4]. Thus, the task is often referred to as closed-set identification. Speaker verification (also known as speaker authentication or detection) is the task of determining whether a person is who he/she claims to be (a yes/no decision). Since it is generally

978-1-5386-5305-0/18/$31.00 ©2018 IEEE

assumed that imposters (those falsely claiming to be a valid user) are not known to the system, this is referred to as an open-set task. By adding a *none-of-the-above* option to closed-set identification task one can merge the two tasks for what is called open-set identification. The authors in [5] present a comparative study of speaker recognition methods based on DTW, GMM and SVM. There are several feature extraction techniques. We will consider those extracted from the individual physiological characteristics, e.g. the frequency of vibration of the vocal cords. This frequency is known as the fundamental frequency [6], pitch or F0. We can categorize the existing classification approaches into three categories:

- Vector-based approaches: Such as Dynamic Time Warping (DTW), Vector Quantization (VQ) and Support Vector Machine (SVM) [7], [5].
- Static approaches: As Hidden Markov Model chains (HMM) and Gaussian Mixture Model (GMM) [8].
- Connectionist approaches: Like neural networks [9].

In [10], the authors used Deep Neural Network (DNN) to provide bottleneck features for speaker recognition. The work was compared to the standard Mel Frequency Cepstral Coefficients (MFCC) features in i-vector/Probabilistic Linear Discriminant Analysis (PLDA) speaker recognition system. The authors in [11] present a forensic automatic system for speaker recognition named Voice Comparison and Analysis of the Likelihood of Speech Evidence (VOCALISE). This system used different algorithms such as the MFCC, phonetic features (such as formants), or features of their own choice (such as voice quality metrics and articulation rate). It also includes GMM with [and without] Maximum A Posteriori (MAP) adaptation, i-vector extraction with PLDA and cosine distance comparison. The author in [12] describes the significant corpora available to support the speaker recognition research and evaluation, along with details about the corpora collection and design. The author in [13] adapted the PLDA based i-vector speaker recognition systems. Two of the proposed approaches are applicable to a larger class of models (low-rank across-class). The authors in [14] present speaker identification using LPC and formant analysis. Authors in [15] explored Joint Factor Analysis (JFA) for text-dependent speaker recognition with random digit strings using the RSR2015 (part III) dataset. The authors in [16] assessed the potential improvement in the performance of the MFCC-based Automatic Speaker Recognition (ASR) systems with the inclusion of linguistic-phonetic information. Testing was run over 20 replications utilizing randomized sets of speakers. In [17], the authors used the estimated quality parameters of speech recording. they validated their experiments on the ELSDSR database.

III. Proposed Approach

Our main goal is to provide a robust and suitable speaker identification system in degraded conditions. For doing so, a set of steps is considered in each phase. Speaker recognition consists of three parts: preprocessing, feature extraction and classification. The aim of this study is to present a novel combination of speech processing and classification approaches in order to recognize speakers in degraded conditions. The steps of the proposed scheme are summarized in Fig. 1.

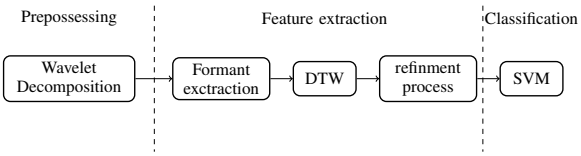

Fig. 1. Flowchart of the proposed approach.

A. Preprocessing

Several pre-processing operations can be carried out in order to extract the useful information and prepare it for the characterization phase. Among these operations, we quote the wavelet decomposition [18] and the filters with pulse response to denoise the raw signals. In this study, we opt for wavelet decomposition to denoise the raw signal. The stationary wavelet decomposition is used in the denoising process. The silence at the beginning and the end of the signal is removed to ensure a better performance.

B. Processing

It is composed from different steps. The purpose of this block is to extract the wanted features with the proposed methods and prepare them to be classified.

1) Formant Extraction: A formant of a speech sound is one of the sound spectrum energy maxima of the speech sound [16]. There are several definitions of the word "forming" (vocal tract resonances, poles, etc.). We denote the formants by the physical notation Fi (measured in Hertz) starting from the first F1 in the low frequencies [19]. The notation F0 is reserved for the fundamental frequency, whose variations over time constitute the intonation of speech. The first two formants, F1 and F2, are not sufficient enough for the vowel description if the language contrasts, for example between drawn and rounded anterior vowels as in French) [20]. Vowels, which are not diphthongs, are more or less stationary speech sounds. Each vowel is therefore characterized by its specific

timbre, determined as a first approximation by F1 and F2. The phonetic discipline correlates the measured values in Hertz of F1, F2, the third formant, and articulations of the phonetic apparatus necessary for the realization of vowels (including lips). Thus, F1 is correlated with the aperture (mouth opening) and F2 with the previous position (high value of F2), posterior (lower value of F2) of the language, as well as with the configuration of the lips [16]. The third formant F3 is also an interesting feature, correlated with the configuration of the lips for the anterior vowels. We extract the first formant as a feature; this formant will be treated by the dynamic time warping.

2) Dynamic Time Warping (DTW): DTW [7] is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. Intuitively, the sequences are warped in a nonlinear fashion to match each other. The DTW algorithm [5] has been used in video, audio, computer graphics, and bioinformatics. Moreover, it can be applied in any situation where data can be transformed into a linear representation. A famous application is in ASR, where it is necessary to take into account very variable speech rates. DTW [8] measures the similarity between two sequences that can vary over time. For instance, similarities between gait in videos can be detected even if the subject in one or other videos walks faster or slower, or if the subject accelerates or slows down. In general, the DTW [8] is a method that seeks optimal matching between two time series, under certain restrictions. The temporal series are deformed by the non-linear transformation of the temporal variable, in order to determine a measure of their similarity, independently from certain non-linear transformations of time. This method of time series alignment is often used in the context of HMM. With the extracted formants, we calculate the DTW. The two used sequences are: the testing sequence and all the remaining samples in the database. Then a refinement process is necessary to enhance the results.

3) refinement process: To ensure a better classification, we opt for enhancing the results of the output of the DTW system. It is a novel method to enhance the results. First, we binarize the matrix output of the DTW. After that, we threshold the number of samples of each block. Every block contains 12 samples related to one person.

IV. Experimental results

Comprehensive experiments were conducted in order to evaluate the performance of the proposed approach compared to a state-of-the-art approach. One should bear in mind that the experiments are carried out on the Ear and Voice Database in Degraded Conditions(EVDDC). The EVDDC database is acquired

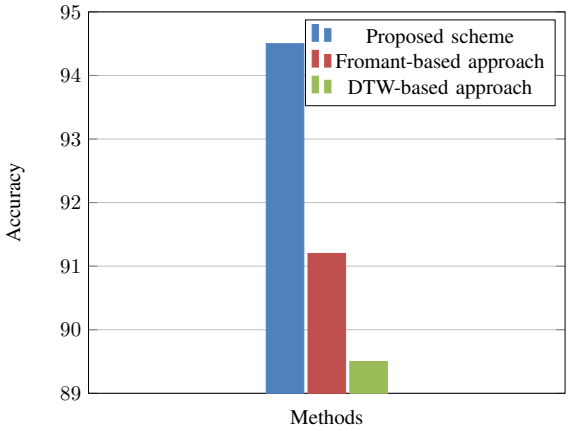

Fig. 2. Performances of the proposed scheme and the state-of-the-art methods

from 111 different subjects (50 in the first session and the rest in the second one) with varied ages and from different places of one country with various tones. The database contains images and voices from some twins. For each individual, a number of samples, between 12 and 50 images and between 12 and 21 voice records, are taken. The database contains some video recording for the ear and the voice of 6 candidates. The EVDDC contains 2742 ear images, 2106 voice records, and 40 video records. It is sequentially numbered for every subject with an integer identification number ("Nperson-Nsample"). This database contains samples from volunteers: 53 females and 58 males. The subjects are from different ages: 40 are between 6 and 14 years, 45 are between 15 and 35 years old, 24 are between 36 and 59 years, and 2 are more than 60 years.

Different types of classifications [8] exist, but one of the most intuitive and widely used is the supervised classification. The idea of supervised classification is to learn a ranking rule from a set of data whose classification is already known. Once the rule is learned, it is possible to apply it to categorize new unknown data. Based on the SVM, a one-leave-out classification [5] was performed. SVMs are a class of learning algorithms initially defined for the discrimination such as the prediction of a qualitative binary variable. SVM [21] solves classification problems (sorting individuals according to their characteristics) and arouses much interest, both for its elegance and its good performances. Since there is no partition of the dataset into training and testing parts in the one-leave-out classification, the accuracy of the whole dataset is set to a constant.

To better demonstrate the advantage of our proposal, we compared it with state-of-the-art works [5] and [14]. In order to assess the efficiency of our

proposed scheme, we have selected the classification accuracy. Fig. 2 provides the comparative results of these methods as well as our results. The method based on the DTW [5] presented in the state-of-the-art records an accuracy of 90.2% whereas the method based on the formant [14] presented on the state-of-the-art records an accuracy of 89.9%. As it can be noticed, our approach achieves higher performance with an accuracy of 94.5%. This performance is achieved thanks to the combination of the formants extraction method, the DTW approach and refinement process.

V. Conclusion

Identity recognition from voice records is an active field of research within the biometric community. The ability to record from a distance and in a covert manner makes voice recognition technology an appealing choice for surveillance and security applications as well as related application domains. In contrast to other biometric modalities, where large datasets captured in uncontrolled settings are readily available, datasets of voice records in degraded conditions are still limited in size and mostly of laboratory-like quality. In this paper we present an approach to recognize speaker in degraded conditions in a smart home context. The result has shown that the combination of two approaches, which are the formant extraction and the DTW, improves the global performances of the system to 94.5%. A possible application for this approach lies in developing a multi-model system for person identification in degraded conditions.

References

[1] Basma M Mohammad El-Basioni, Sherine M Abd El-kader, and Mahmoud Abdelmonim, "Smart home design using wireless sensor network and biometric technologies," *information technology*, vol. 1, pp. 2, 2013.

[2] Rubén Vera Rodríguez, Richard P Lewis, John SD Mason, and Nicholas WD Evans, "Footstep recognition for a smart home environment," *International Journal of Smart Home*, vol. 2, no. 2, pp. 95–110, 2008.

[3] Robert J Orr and Gregory D Abowd, "The smart floor: A mechanism for natural user identification and tracking," in *CHI'00 extended abstracts on Human factors in computing systems*. ACM, 2000, pp. 275–276.

[4] Florent Perronnin and Jean-Luc Dugelay, "Introduction à la biométrie-authentification des individus par traitement audio-vidéo," *Traitement du signal*, vol. 19, no. 4, 2002.

[5] Jr Ding, Chih-Ta Yen, and Da-Cheng Ou, "A method to integrate GMM, SVM and DTW for speaker recognition," *International Journal of Engineering and Technology Innovation*, vol. 4, no. 1, pp. 38–47, 2014.

[6] Makrem Ben Jdira, Imen Jemâa, and Kaïs Ouni, "Speaker recognition system based on pitch estimation," in *Electrical Sciences and Technologies in Maghreb (CISTEM), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[7] Maruti Limkar, B Rama Rao, and Vidya Sagvekar, "Speaker recognition using VQ and DTW," in *International Conference on Advances in Communication and Computing Technologies*, 2012, pp. 18–20.

[8] Loh Mun Yee and Abdul Manan Ahmad, "Comparative study of speaker recognition methods: DTW, GMM and SVM," 2007.

[9] Fred Richardson, Michael Brandstein, Jennifer Melot, and Douglas A Reynolds, "Speaker Recognition Using Real vs Synthetic Parallel Data for DNN Channel Compensation.," in *INTERSPEECH*, 2016, pp. 2796–2800.

[10] Alicia Lozano-Diez, Anna Silnova, Pavel Matejka, Ondrej Glembek, Oldrich Plchot, Jan Pešán, Lukáš Burget, and Joaquin Gonzalez-Rodriguez, "Analysis and optimization of bottleneck features for speaker recognition," in *Proceedings of Odyssey*, 2016, vol. 2016, pp. 352–357.

[11] Anil Alexander, Oscar Forth, Alankar Aryal Atreya, and Finnian Kelly, "Vocalise: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features," Odyssey, 2016.

[12] Douglas E Sturim, Pedro A Torres-Carrasquillo, and Joseph P Campbell, "Corpora for the evaluation of robust speaker recognition systems.," in *INTERSPEECH*, 2016, pp. 2776–2780.

[13] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4047–4051.

[14] Mohd Ali Yusnita, Murugesa Pandiyan Paulraj, Sazali Yaacob, Shahriman Abu Bakar, and A Saidatul, "Malaysian english accents identification using lpc and formant analysis," in *Control System, Computing and Engineering (ICCSCE), 2011 IEEE International Conference on*. IEEE, 2011, pp. 472–476.

[15] Themos Stafylakis, Md Jahangir Alam, and Patrick Kenny, "Text-dependent speaker recognition with random digit strings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1194–1203, 2016.

[16] Vincent Hughes, Paul Foulkes, and Sophie Wood, "Formant dynamics and durations of um improve the performance of automatic speaker recognition systems," in *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA). University of Western Sydney, Australia.* York, 2016.

[17] Gheorghe Pop, Dragos Draghicescu, and Dragos Burileanu, "On forensic speaker recognition case pre-assessment," in *Speech Technology and Human-Computer Dialogue (SpeD), 2013 7th Conference on*. IEEE, 2013, pp. 1–8.

[18] Rajeev Aggarwal, Jai Karan Singh, Vijay Kumar Gupta, Sanjay Rathore, Mukesh Tiwari, and Anubhuti Khare, "Noise reduction of speech signal using wavelet transform with modified universal threshold," *International Journal of Computer Applications*, vol. 20, no. 5, pp. 14–19, 2011.

[19] Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61–81, 2016.

[20] Bageshree V Sathe-Pathak and Ashish R Panat, "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 4, pp. 296–299, 2012.

[21] Thomas Rückstieß, Christian Osendorfer, and P Patrick van der Smagt, "Sequential feature selection for classification.," in *Australasian conference on artificial intelligence*. Springer, 2011, pp. 132–141.