The Use of Audio Fingerprints for Authentication of Speakers on Speech Operated Interfaces

Abraham Woubie, Tom Bäckström and Pablo Pérez Zarazaga

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
abraham.zewoudie@aalto.fi, tom.backstrom@aalto.fi, pablo.perezzarazaga@aalto.fi

Abstract

In a multi-speaker and multi-device environment, we need acoustic fingerprint information for authentication between devices. Thus, in these kinds of environments, it is crucial to continuously check the authenticity of speakers and devices within a short duration since different kinds of people could join or leave the environment. In this work, we propose the provision of different levels of authentication to different speakers in a multi-speaker multi-device environment using acoustic audio fingerprint information. Firstly, the audio fingerprints are extracted continuously every few seconds. Then, the extracted fingerprints are passed to a speaker recognition module which checks if the fingerprint is enrolled for that particular environment or not. Finally, the proper level of authentication is provided for each speaker. Our experimental results on Voxceleb-1 dataset show that acoustic fingerprints can be successfully used for authentication purposes in a multi-speaker multi-device environment.

Index Terms: audio fingerprints, GAN, privacy, speech interface, speaker recognition

1. Introduction

Recently, speech interfaces have become the preferred user interface of smart devices. From Siri to Alexa, speech interfaces are embedded in various devices that people use daily including mobile phones, tablets, smartwatches and smart speakers. Speech interfaces help different kinds of people such as hearing impaired, blind and individuals with neuro disabilities. People usually interact with these devices using short duration of speech. Thus, the integration of a speaker recognition system that identifies speakers from short-duration utterances is crucial in speech interfaces. For example, a voice assistant in a home scenario needs rapid identification of speakers since a person entering the home immediately joins the present conversation and the corresponding trusted relationship. At the extreme, a short greeting at the door could be of the order of 2 seconds. Therefore, we need to recognize speakers within 2 seconds because in a scenario where people can move between rooms, we need to quickly adapt to new persons in the room. Hence, we need to continuously check the authenticity of speakers within a short duration as different speakers could join the environment.

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals. It is divided into speaker identification and speaker verification. Speaker identification determines which registered speaker provides a given utterance from amongst a set of known speakers [1]. Speaker verification accepts or rejects the identity claim of a speaker [2]. One of the essential parts of speaker recognition modules is feature extraction. Feature extraction retrieves relevant information from the acoustic signal. Therefore, the extracted features should have large between-speaker variability and small within-speaker variability. Mel Frequency Cepstral Coefficients (MFCC) are the most widely used acoustic features for speaker recognition [3].

Recently, different approaches have been developed to improve the performance of speaker recognition systems. The most popular ones are based on i-vectors [4] and x-vectors [5,6]. Usually, i-vectors are normally extracted from long duration of speech segments since the Maximum A Posteriori (MAP) point estimate of the total factors (i-vectors) for segments of short duration cannot be estimated reliably. However, in a multi-device and multi-speaker environment, extracting speech segments from longer duration is not ideal since different devices and speakers could join or leave the environment at any time. Thus, the state-of-the-art i-vector techniques can not be used in a multi-device and multi-speaker environment to identify speakers.

Most of the state-of-the-art of speaker recognition systems use either the whole utterance to identify speakers [7–10] or segment of speech with a duration of at least two seconds for good recognition performance [11–13]. Although the state-of-the-art speaker recognition systems use medium and long speech segments (i.e., at least 4 seconds), speech interfaces usually use very short-duration commands which have duration of less than 4 seconds.

Thus, the main contribution of this work is that we propose the extraction of audio fingerprints from few seconds of speech (i.e., 2, 2.5 and 3 seconds) and provision of different levels of authentications to different speakers in a multi-device and multi-speaker environment. The extracted fingerprint is passed to the speaker recognition module. The speaker recognition module checks the authenticity of the fingerprint for that particular environment and provides the allowed level of authentication. The extraction and authentication process is repeatedly carried out every few seconds.

The main applications of the proposed work can be carried out in different scenarios such as:

- **Home Scenario**: Let us assume that each family member has his/her own phone and the family members are in the same room. During their meeting/discussion, a visitor may join the family. What kind of access shall we give to the family members and the visitor? Thus, a proper access management could be applied to identify speakers and provide different levels of access for the family members and visitors.
- **Business Applications**: Consider telecommunications applications at the office; what kind of access should be given for the different collaborators in different business applications such as a teleconference?
- **Customer Service**: A customer may need to provide personal identification (i.e., passport/national id) to get access to a service. However, the customer may not want
to provide his personal identification since it can be photocopied and forged easily. Thus, the customer could provide his voice sample and audio fingerprints are extracted from the voice samples, and his/her request to the service is accepted/rejected based on his fingerprint.

The rest of this paper is organized as follows. The next section gives an overview of acoustic fingerprints. Section 3 describes the architecture of the proposed system. The experimental results and conclusions are presented in Section 4 and Section 5, respectively.

2. Acoustic Fingerprints

An audio fingerprint is a compact content-based signature that summarizes an audio recording [14]. Audio fingerprint has recently attracted attention since it allows the identification of audio irrespective of its format and without the need of meta-data or watermark embedding. An ideal audio fingerprint accurately identifies a music/speech regardless of the level of compression and distortion or interference in the transmission channel [14]. Another benefit of an acoustic fingerprint is it is computationally efficient. Efficiency is critical in a real application both in the calculation of the fingerprint of the unknown audio. The computational cost is related to the size of the fingerprint. Since the size of audio fingerprints is low dimensional (i.e., 32 by 32), the training of deep learning models using acoustic fingerprint is computationally feasible.

The extraction of audio fingerprints is normally based on spectrogram which is an approximate decomposition of the signal over time and frequency. It is created by taking a short window of time of the signal, and then performing a Fourier transform that decomposes that window over its frequencies. By repeatedly performing this calculation for subsequent windows of time, we find the frequency composition of the audio as time progresses.

In this work, we use two different types of acoustic fingerprints: quantized and non-quantized. The acoustic fingerprint method selected for this work applies a decorrelating transform to the time and frequency axes of the signal spectrogram and then quantizes the decorrelated values [15]. The corresponding audio segment is divided into 32 time frames and their frequency components are grouped into 32 energy bands, resulting in a $32 \times 32$ energy spectrogram of the audio signal. The values of this matrix are then decorrelated using a 2D-DCT over the time and frequency axes of the spectrogram. The real values of the 2D-DCT transformed spectrogram are flattened into an array of features which represent the non-quantized fingerprint.

The decorrelated values from the 2D-DCT contain most of their information in the lower time-frequency bands. Thus, the resulting values are quantized taking advantage of the distribution of the information. In this case, the quantization method is based on the mutual information between matching audio samples so that the fingerprint maximizes the amount of information contained in audio samples from the same scenario [15]. The quantized components are then represented as an array of binary values that will be used as the quantized fingerprint.

3. Proposed System

The proposed system that authenticates speakers based on their acoustic fingerprints is shown in Figure 1. The proposed system has two modules, i.e., training and evaluation. The system is trained and evaluated using two different types of acoustic fingerprints (i.e., quantized and non-quantized) to assess the impact of quantization. While the size of the quantized acoustic fingerprint is 32 by 16, the non-quantized has a size of 32 by 32. We trained the following types of networks in the training phase: discriminator, adversary and mechanism. Firstly, we randomly select sample acoustic fingerprints from the training set. Then, we generate sample noises where each sample noise has a size of vector 20. The noises are concatenated with the fingerprint to distort the fingerprint information. Then, the adversary is trained using the concatenated fingerprint and noise. The discriminator is trained using two types of inputs: the original fingerprint from the training data, and the concatenated fingerprint and noise. Finally, the mechanism is trained using three inputs: the original fingerprints, the original fingerprints concatenated with the noise, and speaker labels. Since the main purpose of the mechanism is to reduce the distortion between the two fingerprints, it uses speaker label information. While the adversary is trained to minimize the log-loss with respect to this posterior estimate, the mechanism network is trained to attain the objective of minimizing distortion by maximizing the adversarial loss.

For the evaluation phase, given an unseen test utterance, an acoustic fingerprint is extracted from the test utterance. A sample noise with size of vector 20 is also generated. Then, the unseen fingerprint and noise are concatenated. Afterwards, the combined fingerprint is compared against the critic model. For example, if the environment is a home scenario, the question becomes “Does the unseen acoustic fingerprint belong to one of the persons living in the house?” or “Is the speaker a visitor?”. Finally, the proper level of authentication will be given for the speaker. Note that since different speakers could join or leave the environment, we need to continuously check the authentication process. Thus, we continuously extract acoustic fingerprints from short duration (i.e., 2, 2.5 and 3 seconds) of speech in the evaluation phase and check if the fingerprint is enrolled for a particular environment.

![Figure 1: The proposed system that authenticates speakers based on their acoustic fingerprints.](image)
We handle both the quantized and non-quantized versions as vectors of 512 and 1024, respectively. We consider the fingerprint to be both the useful and the observed data, i.e., \( W = Y \), the label as the sensitive attribute \( X \), and the mechanism release as a fingerprint \( Z \). We measure the distortion between the original and released fingerprint \( Y, Z \) as:

\[
d(Y, Z) = -\frac{1}{512} \sum_{i=1}^{512} Y[i] \log(Z[i]) + (1 - Y[i]) \log(1 - Z[i]),
\]

which, for a fixed \( Y \), corresponds to measuring the average KL-divergence between corresponding pixels that are each treated as a Bernoulli distribution. Thus, the utility objective is to minimize the fingerprint distortion (i.e., acoustic fingerprints from the same speaker have high similarities irrespective of the session and channel variabilities).

4. Experiments

4.1. Experimental Setup

We used Keras [16] as the neural network environment to accomplish the objectives of this work. We used Python for extracting the audio fingerprints. Both the mechanism and adversary networks used two hidden layers with 1000 nodes each and fully-connected links between all layers. The hidden layers used tanh as the activation function. The mechanism input layer uses 512 + 20 nodes for the audio fingerprints concatenated with 20 random uniform seed noise values. The mechanism output layer uses 512 nodes with the sigmoid activation function to directly produce an audio fingerprint. The discriminator network architecture uses a single hidden layer with 500 nodes, and has an output layer with one node that uses the sigmoid activation function. The network is trained for 100 epochs using a batch-size of 64.

We have compared the performance of the two types of audio fingerprints (i.e., quantized and non-quantized). In addition, we have also analyzed the performance of these two fingerprint types by using different duration of speeches (2, 2.5 and 3 seconds). While the quantized audio fingerprint has a size of 32 by 16, the non-quantized one has 32 by 32. The non-quantized fingerprint is normalized using mean variance normalization.

The proposed system is evaluated on VoxCeleb1 dataset which consists of 148642 training and 4874 test audios. We selected 5000 audio files from this training set to train the proposed model and selected 1000 test files to evaluate the performance of the trained mechanisms. The total number of speakers in the training and test sets are 40 and 10, respectively.

The performance of the proposed system is evaluated using two metrics: (i) the Equal Error Rate (EER) which is the rate at which both acceptance and rejection errors are equal; and (ii) the cost function

\[
C_{det} = C_{miss} X P_{miss} X P_{tar} + C_{fa} X P_{fa} X (1 - P_{tar})
\]

where we assume a prior target probability \( P_{tar} \) of 0.01 and equal weights of 1.0 between misses \( C_{miss} \) and false alarms \( C_{fa} \). Both metrics are commonly used for evaluating identity verification systems.

4.2. Experimental Results

Figure 2 shows the results of quantized versions of acoustic fingerprints using different duration of speeches. The results of the figure demonstrate that while the use of 3 seconds of acoustics fingerprint provides the lowest EER (i.e., 7.33%), the use of 2 seconds provides the highest EER (i.e., 10.82%).

Similarly, we have also compared the performance of the non-quantized version of the acoustic fingerprint. The results of Figure 2 show that the use of 3 seconds of acoustics fingerprint provides the lowest EER (i.e., 10.8%). The extraction of acoustic fingerprint from 2 seconds of duration provides the highest EER (i.e., 13.23%).

In addition to EER, we have also compared the minimum detection cost function (minDCF) values of different durations of seconds both for the quantized and non-quantized versions of audio fingerprints. As it is shown in Table 1 and 2, the use of 3 seconds provides the lowest minDCF values both the quantized
and non-quantized audio fingerprints. The extraction of acoustic fingerprint from 2 seconds provides the highest minDCF values for the two types of audio fingerprints.

Table 1: EER and minDCF of quantized fingerprints for different duration of audio fingerprints.

<table>
<thead>
<tr>
<th>Duration</th>
<th>EER</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 seconds</td>
<td>10.82%</td>
<td>0.72</td>
</tr>
<tr>
<td>2.5 seconds</td>
<td>8.33%</td>
<td>0.65</td>
</tr>
<tr>
<td>3 seconds</td>
<td>7.33%</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Thus, the results reported demonstrate that acoustic fingerprints provide useful and complementary speaker information. The experimental results also show that the quantized version of audio fingerprints provide both better EER and minDCF values compared to the non-quantized versions of the audio fingerprints. The experimental results also demonstrate that both the minDCF and EER values decrease when the duration of audio fingerprints increase both for the quantized and non-quantized versions of audio fingerprints.

Table 2: EER and minDCF of non quantized fingerprints for different duration of audio fingerprints.

<table>
<thead>
<tr>
<th>Duration</th>
<th>EER</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 seconds</td>
<td>13.23%</td>
<td>0.86</td>
</tr>
<tr>
<td>2.5 seconds</td>
<td>12.88%</td>
<td>0.83</td>
</tr>
<tr>
<td>3 seconds</td>
<td>10.8%</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Although the EER values of quantized version of audio fingerprints are less than the state-of-the-art results on short-duration segments [12, 13], we need to take into consideration that both the computational cost of extracting acoustic fingerprints and training of GAN model using acoustic fingerprint is very low. We have compared the running time complexity of our proposed technique with a speaker verification system that uses spectrogram with CNN architecture. While the audio-fingerprint experiment with GAN technique with a batch size of 100 and 100 epochs take 10 minutes on Quadro P2000 GPU, the use of CNN architecture using spectrogram with the same batch size and same number of epochs on the same device takes approximately 60 minutes. This shows that the proposed system can easily enroll new speakers for home or other types of scenarios. Thus, the results reported in this work can be taken as a good start point for authentication of speakers on speech operated devices.

5. Conclusions

In multi-device and multi-speaker environments, we need fingerprint information for authentication between devices. In this work, we propose the provision of different levels of authentication to different speakers in these environments using acoustic fingerprint. Thus, we first train speaker models for a selected number of speakers using their audio fingerprints. Then, given an unseen test utterance, an audio fingerprint is extracted every few seconds (i.e., 2, 2.5 or 3). Afterwards, the extracted fingerprint is compared against trained speaker models. Finally, the speaker is provided the allowed level of authentication based on whether the speaker is enrolled for that environment or not.

The future work could focus on reducing the Equal Error Rate (EER) for audio fingerprints extracted from short duration segments. In addition to this, the future work could focus on preserving the privacy of speakers in a multi-device and multi-speaker environment.

6. Acknowledgment

This work has been supported by the Jane and Aatos Erkko foundation funding under contract 700795 AUTHSPKR.

7. References