The Quarterly Magazine for Digital Forensics Practitioners

Issue 37 • November 2018

DIGITAL FORENSICS Magazine

Forensics Europe Expo
DFM Forensics Conference 2019
Call for Papers

Secure Payment Systems

Exploring the role of Digital Forensics within National and International Payment Systems

PLUS

- Information Leakage in the IoT
- Voice Biometry and its Use in Forensics
- An Introduction to Open Source Intelligence
- **From the Lab:** Graph Theory for DF Investigations

9 772042 061004 37
Issue 37 / £14.99
TR Media

Voice Biometry and its use in Forensics

Petr Schwarz provides us with an understanding of Voice Biometry.

Sound is a vibration that propagates as an audible wave of pressure through air. The sound has a source that is often a vibrating solid object. When the frequency and time evolvement of the wave are observed, the source may often be identified. It may have unique spectral or temporal characteristics. In the case of humans, we speak about the human voice. The sound is produced when the air goes from the lungs through the glottis. Then it is further modulated, the wave is reflected, attenuated, and new frequencies are created, when the air goes through the human vocal tract. It means through the larynx, oral cavity, transformed by the tongue, gums, teeth, and lips. Each person has a different shape and configuration of these organs. For example, women or children have a smaller body, shorter larynx, and, therefore, their voice has a higher fundamental frequency. The anatomy may vary with the human race (place of living) too, for example, American vs. Asian people. Then people move their vocal organs during speech. These movements are also unique. Some of them are learnt during childhood (dialect, the pronunciation of some words, pronunciation defects), and some of them may not be changed. Therefore, current voice biometry techniques investigate both the frequency representation of speech and its temporal evolvement.

Technology - Representation of Voice

The audible waves of acoustic pressure are measured using microphones, sampled, and then digitalized. The result is a waveform, a series of acoustic pressure measurements evolving in time. The waveform is further converted into a spectrogram, where spectra are computed for each small segment, usually 10 to 30ms long. The spectrogram is an initial representation for current state-of-the-art automatic voice biometry systems.

The Compression of Speaker Information into Speaker Models (or Voiceprints)

The variable length time-frequency representation of a voice is further compressed to a fixed-size record. This record describes the shapes, configuration, and possible movements of vocal organs. It is usually a fixed-size set of floating point numbers. The right set of numbers (or in other words, the best transformation from spectrogram to mathematical model) is found using statistical or artificial intelligence approaches on a large set of recordings. This set may have recordings from thousands of speakers coming from different continents, nationalities, and speaking different languages. The recording conditions (acoustic channel, microphone, telephone, inside, outside etc.) are important too. The voice may be degraded by its transmission through air, telecommunication channels, and some noise is also captured with the voice by the recording devices. The higher variability in acoustic channels covered in the set, the better and more robust the voice biometry system that may be developed. ▷

The audible waves of acoustic pressure are measured using microphones, sampled, and then digitalized.


0 10 1 0
0 10 1 0
0 10 1 0

iVectors and DNN-based Systems

The primary issue for current state-of-the-art systems is how to estimate the transformation from a time-frequency representation of speech to the speaker mathematical model. There are two main approaches widely used today, namely:

- iVectors
- DNN-based systems

iVectors

The iVector approach was first used in production in 2010, and it soon started to be popular. It uses a Universal Gaussian Mixture Background Model (UBM) trained on a large number of speakers, and a technique for the selection of the most important UBM parameters (a few hundred). After the system is trained on sets of known speakers (the development phase), it can process new recordings. The UBM is adapted to each recording and only the most important parameters are kept. These parameters are called an iVector. The iVector encodes the total variability in the input spectrogram (Figure 3); it does not matter if it is speaker variability or channel/noise variability. Therefore, other processing for removing the channel/noise variability from iVector may follow. The whole process is shown in Figure 4.

Deep Neural Networks (DNN) - Based Systems

The disadvantage of the iVector-based system is that it focuses on the total variability in a signal. It describes the entire vocal tract and its movements as precisely as possible. But not all the shapes and movements are equally important to distinguish among speakers. For forensics and investigation use cases, the goal is not to have the best overall picture of the human vocal tract, but rather to find places distinguishing the speakers the most. Here, the Deep Neural Networks have the advantage. DNNs can replace a part of the schema above, for example, they can be used to post-process a spectrogram and extract more detailed information, or they may replace the whole extraction schema.

A fully DNN system uses two deep neural networks. The first neural network is trained to classify short (fixed duration) pieces of the spectrogram to a speaker across a large number of speakers. The final speaker classification is not used when processing

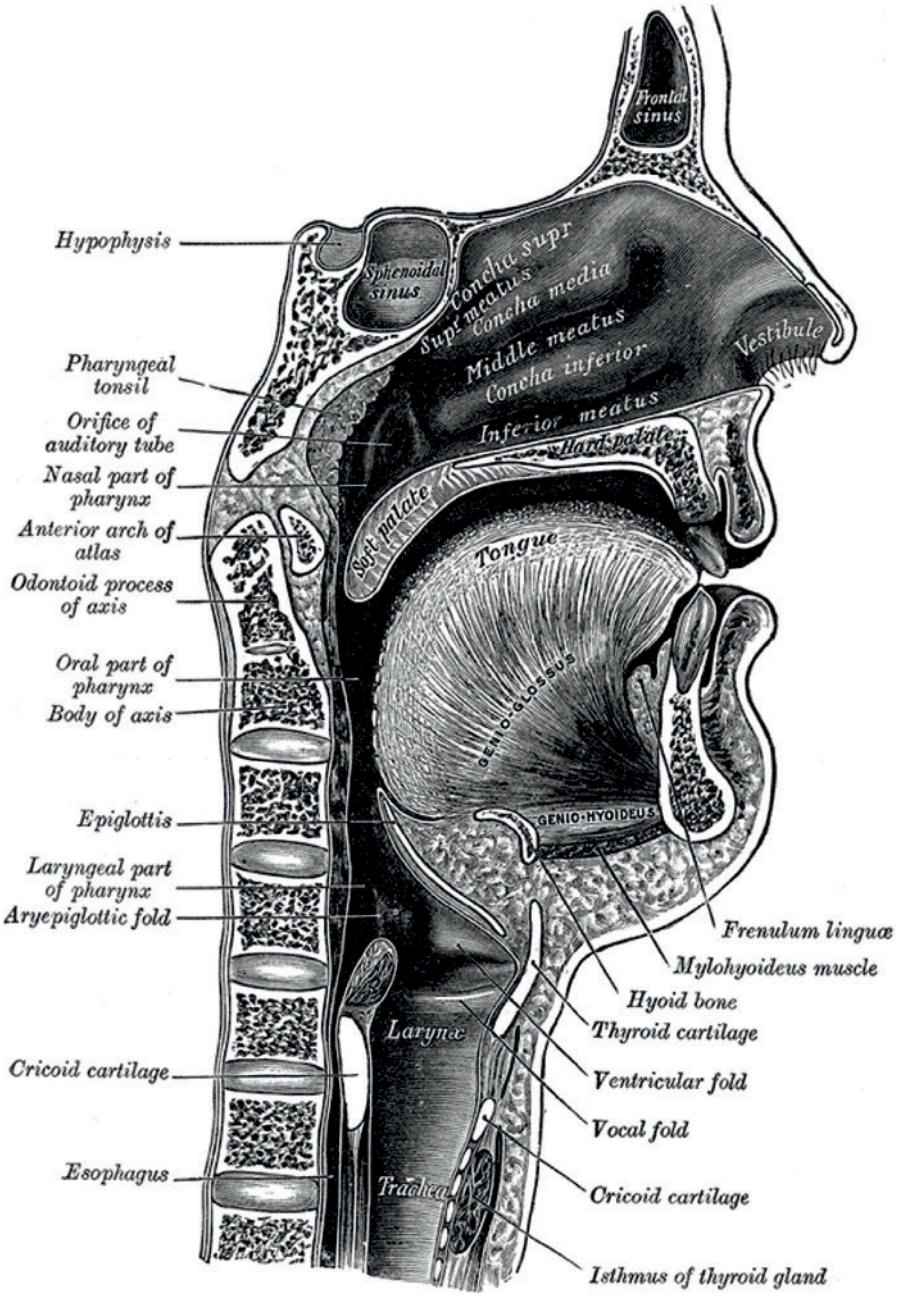

Figure 1. A Diagram of the Human Vocal Tract (Source: Wikipedia)

The wave is reflected, attenuated, and new frequencies are created, when the air goes through the human vocal tract.

new recordings, but rather some final layers of the neural network are removed, and an internal well-compressed speaker representation available inside the network (a few hundred floating point values) is used for future processing. This representation can encode any speaker, not only those seen during neural network training. The information extracted from each short spectrogram piece is then averaged (a collection of statistics) over the whole audio file and sent to the second neural network. The second network is also trained to classify among speakers across a large number of speakers, and again, some final layers are removed and some internal well-compressed speaker information available inside the network is used as the speaker model (or voiceprint) when processing new recordings. The whole process is shown in Figure 5.

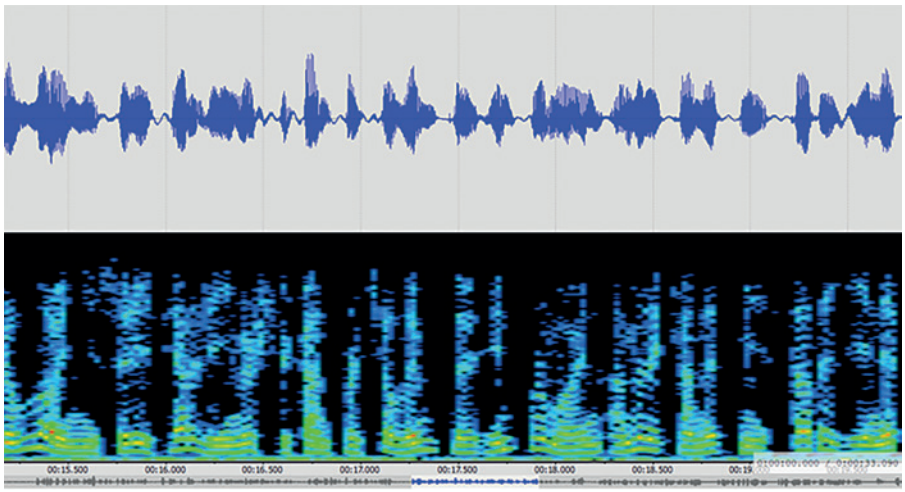

Figure 2. Waveform & Spectogram – A Screenshot from Phonexia Voice Inspector

Using DNN to create a speaker model brings multiple benefits. DNN-based systems can have both higher accuracy and achieve higher processing speed, together with lower RAM requirements. For example, Phonexia, in the latest generation of its voice biometrics, released an exclusively DNN-based system called Deep Embeddings™. The technology is able to create voiceprints twice as fast, be 2.4 times more accurate, and have a memory consumption that is just a quarter of the previous Phonexia voice biometric engine, which was already one of the fastest and most accurate on the market. This evolution opens up the possibility to use voice biometrics technologies on devices with no permanent connection to the Internet, and on smaller and cheaper devices in general.

Speaker Scoring

The iVector system or DNN-based system can be used to estimate a precise speaker model for each recording. When we have two models, these models can be compared to each other to get a similarity measure. The most common technique for the comparison of speaker models is Probabilistic Linear Discriminate Analysis (PLDA). PLDA is trained on a larger set of speaker models to distinguish what is the wanted (speaker) variability and what is unwanted (usually acoustic channel distortion or noise). Then a likelihood ratio between two hypotheses is evaluated. The first hypothesis says: "the two compared models came from the same speaker". The second hypothesis says: "the two compared models are from different speakers".

$$LR = \frac{H1\ (\text{speaker } 1 = \text{speaker } 2)}{H0\ (\text{speaker } 1 \neq \text{speaker } 2)}$$

If the likelihood ratio is higher than 1, the system thinks that it is the same speaker. If lower than 1, the system thinks that it is not the same speaker. If it is close to 1, the system is not sure.

Figure 4. Schema of an iVector Extractor (Source: Phonexia)

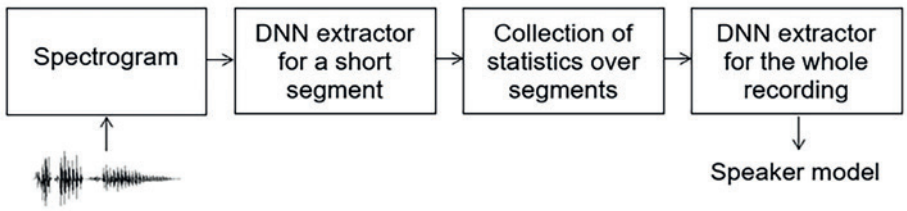

Figure 5. Schema of a DNN Extractor (Source: Phonexia)

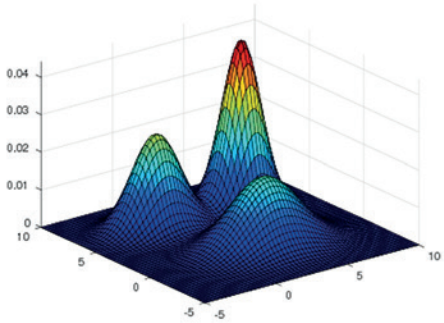

Figure 3 The Gaussian Mixture Model (Source: Phonexia)

Forensic Use

Current state-of-the-art voice biometry systems are very accurate but also very complex. There are complex algorithms, large training audio sets (thousands of speakers), and complicated training procedures. This means it is hard to explain details about the system to an expert in the field, and it is almost impossible to present it in court to a judge or jury in a limited time. Therefore, the common practice is to build a second voice biometry system, a very simple one. The system is trained on a relatively small and well-controlled data set. Likelihood ratios (or scores in general) from any industrial voice biometry system can be taken as evidence into a forensic voice biometry system. The better the first system is, the more conclusive the results from forensic system are, but even if the first system is poor, the forensic system should give valid answers. It may even be that there is no strong support for either hypothesis, the two recordings being from the same speaker, and the two recordings being from a different speaker. ▷

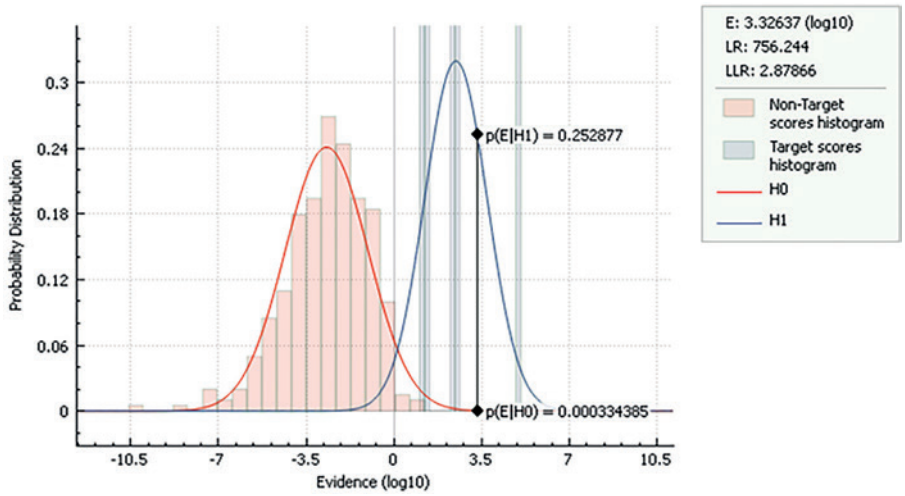

Figure 6. Forensic model – A screenshot from Phonexia Voice Inspector (Source: Phonexia)

REFERENCES

1. "Human voice", Wikipedia, https://en.wikipedia.org/wiki/Human_voice

2. N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Oellet, "Front-end factor analysis for speaker verification," in IEEE Transactions on Audio, Speech, and Language Processes, 2011, Vol. 19, pp. 788–798

3. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur , "X-vectors: Robust DNN Embeddings for Speaker Recognition", ICASSP 2018

4. A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen and T. Niemi, "Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition", European Network of Forensic Science Institutes, 2015

It may be costly to collect the population sets, so it is possible to use some less targeted sets at the beginning during the initial investigation, and then to apply a well-designed (detailed) population set to get stronger proof for presentation to court.

The easiest model is two Gaussians, one Gaussian is modelling speaker evidence, and the second one is modelling population evidence. Three sets of recordings are needed to build such a model:

1. The questioned recording that is investigated
2. Set of suspect recordings
3. Set of population recordings

Suspect recordings are a few recordings collected from the suspect during investigation. The population recordings should be from speakers as close as possible to the suspect, from the same gender, race, nationality, age, speaking the same language, etc. It may be costly to collect the population sets, so it is possible to use some less targeted sets at the beginning during the initial investigation, and then to apply a well-designed (detailed) population set to get stronger proof for presentation to court. For example, if the population set speakers are speaking English and the suspect speaks German, the objection of defence may be that the system recognizes language. It is up to the forensic expert to disprove all such objections using well-designed population sets.

One such model is shown in Figure 6. For the H1 hypothesis (speaker 1 = speaker 2), a suspect speaker distribution is created (the blue line). The entire suspect recordings are compared to each other using an industrial voice biometry system, the evidence for all recording pairs is obtained, and the Gaussian parameters (mean and variance) are calculated. The disadvantage is that several suspect recordings are needed to get a good estimate. For the H0 hypothesis (speaker 1 ≠ speaker 2), a population distribution is created (the red line), the population recordings are compared to the suspect recordings, and again, the Gaussian parameters are calculated.

After that, the questioned recording is compared to a suspect recording. Using the evidence, likelihoods for both hypotheses can be directly read from the Gaussian curves, and a likelihood ratio calculated. The likelihood ratio is presented to the court. If multiple suspect recordings are used, one global likelihood ratio may be obtained by multiplication of per-file likelihood ratios.

With this simple forensic model, the forensic expert is the author of the whole system, has everything under their control, knows all the details about data sets, can answer all possible questions, and can fully focus on the forensic case. •

Petr Schwarz, PhD *is the CTO and co-founder of Phonexia. He helped to establish the well-known research group Speech@FIT at Brno University of Technology, Czech Republic, worked as a researcher at Oregon Graduate Institute in Portland, OR, USA, and founded Phonexia in 2006. He participated in the development of multiple speaker recognition and language identification systems evaluated by the United States National Institute of Standards and Technology. Petr was also a team member on several Johns Hopkins University summer research workshops in the field of human language processing, and he is also the co-author of several open source software projects. He has also worked on several European, USA, and Czech research projects, and is the author or co-author of dozens of impactful research articles.*