IDIAP analysis of the latest bin Laden tape

Background information

Among other activities in speech processing, computer vision and machine learning, IDIAP has been involved in automatic speaker verification/authentication for about 10 years, where they incorporate state-of-the-art statistical approaches, as are also used in many other laboratories.

After the release of the latest bin Laden tape, IDIAP was often approached by journalists asking for general information about automatic speaker verification systems: how they worked, and how well they perform. Our answer has always been that these systems perform relatively well in well-controlled environments, such as banking and telephony applications, the main focus of the researchers working in this area. These systems are based on the collection and modeling of many utterances spoken by numerous people, speaking the targeted language, as well as a few utterances, pronounced in clean environments, from each person whose voice print will later have to be identified. In these well-controlled environments, correct verification performance in the range of 95-98% is often reported. However, when working in uncontrolled environments with degraded quality, and/or when there are insufficient training utterances (which is typically the case in forensic applications), this performance level can drop dramatically, making it impossible to draw conclusions with strong certainty.

More recently, IDIAP was also approached by the French TV station France 2, asking again for the same kind of information. In this framework, it was also agreed that IDIAP would illustrate the way state-of-the-art speaker verification systems work by processing the bin Laden recordings that were available at France 2, including the unauthenticated bin Laden recording broadcasted by Quatar Al Jazeera on November 12, 2002. US experts who have heard or processed the tape usually support the conclusion by US law enforcement officials that it probably is bin Laden speaking. However, it is also usually agreed that the latest tape will likely never be fully authenticated because its poor quality defies complete analysis by the best voice/linguist experts or the most sophisticated voice print technology. Although IDIAP fully agrees with these statements, it was decided, mainly motivated by pure scientific curiosity, to go ahead with the experiment and see what conclusion our state-of-the-art speaker authentication system would reach.

Experiment

A few days ago, France 2 thus kindly provided IDIAP with about **1h30 of audio/video recordings,** recorded through CNN or Al Jazeera, and including about **one hour of authenticated** (through video) **recordings from bin Laden** and about **30 minutes voice signal from his associates** or other persons speaking Arabic.

As usually done in the scientific community, and to allow for a fair and unbiased evaluation, these recordings were split into an independent « training set » (i.e., not used to evaluate the system) used to build a statistical model of bin Laden's voice (see technical detail below), and a « test set » used to evaluate/predict the performance of the system. Usually, the training data should also include a large number of speakers (in the targeted language) to build a so-called « world model » used as a reference of « non-bin Laden examples », and several examples of bin Laden voice. Given the very limited number of « non-bin Laden » Arabic speakers, we decided not to train a new « world model » and to rely on a pre-trained English model.

Although this is a first weakness of our approach, results below will show that it didn't significantly limit the performance of the system on the data in hand.

Before presenting the system with the latest unauthenticated recording, we extracted **44 recordings** of audio data from the above broadcast material (thus covering about 1h30 of audio data) test data from the above broadcasted material. These recordings included:

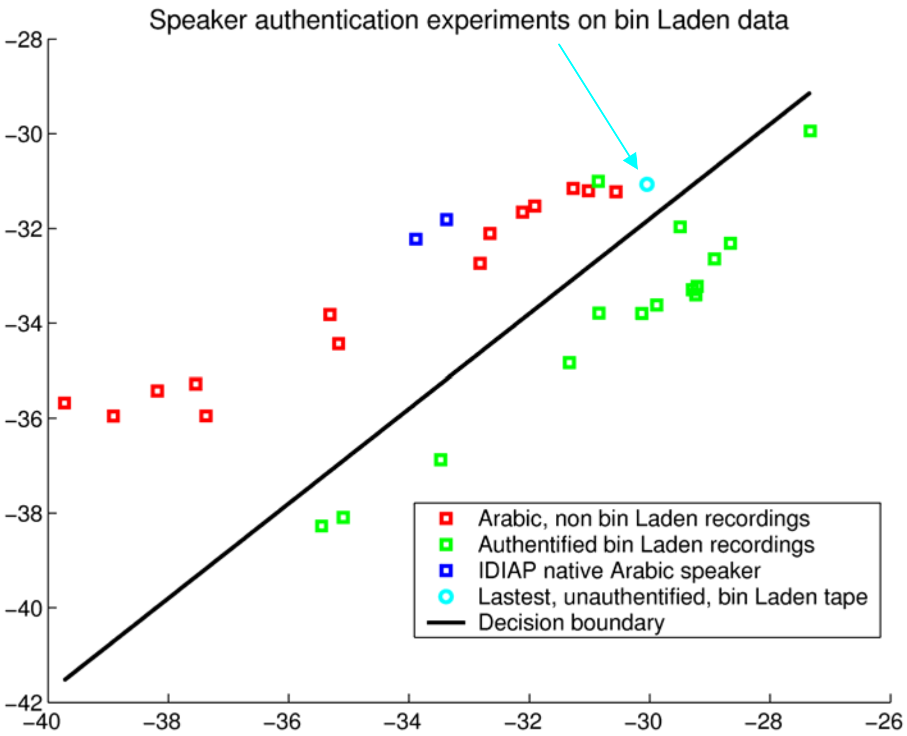
- **30 recordings authenticated as from bin Laden**, which were split in two sets :
  - **15 to train the model**, referred to as "**train set**": not represented on the model below
  - **15 to evaluate the model**: these appear as **green squares** in the figure below.
- **14 recordings authenticated as from other Arabic speakers** and also used to evaluate the model (to evaluate "impostor accesses"): these appear as **red squares** in the figure below.

The last two sets of recordings, referred to as "**test set**" thus contained **29 recordings**, including 15 recordings from bin Laden and 14 from other Arabic speakers. The quality of these recordings (in the training set, as well as in the test set) was highly variable, ranging from good quality to mediocre and very poor quality. In an informal blind test performed with colleagues at IDIAP (who are not experts in linguistics), including one native Arabic speaker, it was often very difficult to achieve a high human classification rate.

After having optimized our statistical model (briefly discussed below), and keeping in mind the limited amount of data, we used the resulting automatic user authentication system to generate one data point (associated with an identity decision) for each test recording, thus resulting in 29 data points in the two-dimensional space represented in the figure below. According to the figure plot below, all the points above the decision threshold (the line in black, optimized to maximize the decision margin) would be classified as « non bin Laden », while the recordings falling below the decision threshold would be classified as containing the voice print of bin Laden. Thus, it can be seen on the figure below that **all of the 29 recordings, but one (the green square above the line), were properly classified**.

Based on this experiment, an unreasonable conclusion (often drawn by some journalists) is that the resulting system is reliable at 97% since it made one error over 29 examples. However, when dealing with statistics, drawing this kind of conclusion when dealing with so few examples is simply too premature, and often wrong. In theory, we would need an infinite number of examples to draw any definite conclusion, but the more examples we have, the more reliable the conclusions would be. However, to further test our model, we also recorded two utterances (represented by blue squares in the figure plot below) from a native Arabic speaker at IDIAP repeating and mimicking one of bin Laden's recording. It can be seen from the plot that these two additional recordings were also properly rejected as « non bin Laden », finally resulting in one error over 31 examples.

Thus, relying on the relatively good performance of the system on this (too limited!) test data, we then presented our system with the latest, unauthenticated bin Laden recording. The resulting point is represented by a light blue circle pointed by an arrow. According to the system, this last recording would thus not be attributed to bin Laden. However, on top of the limitations already discussed above, it can be seen that this point also falls very close to the decision threshold, which further decreases the confidence we can have in this result.



Conclusion

The work reported here was mainly motivated by pure scientific curiosity, also aiming at showing the possibilities and limitations offered by automatic speaker authentication system in non-optimal conditions (typically, noisy environments and limited amount of recordings). While this study does not permit us to draw any definite (statistically significant) conclusions, it nonetheless shows that there is serious room for doubt, and that it is also difficult to agree with some US officials saying that it is 100% sure that it is bin Laden. When addressing a problem with a scientific perspective (as opposed to a political approach), one has to be ready to also accept the uncertainty of the results. Even if the confidence of these results can be boosted by exploiting multiple automatic systems and multiple human expert opinions, it will never be possible to authenticate the latest bin Laden tape with 100% assurance.

A few words about the statistical user authentication technology used

The speaker authentication technology used here is a state-of-the-art approach based on (text independent) statistical modeling of the spectral characteristics of each voice by a set of multi-Gaussian densities in an acoustic space, optimized to extract at best the voice print characteristics while being independent of the transfer channel characteristics. Of course, the state-of-the-art acoustic features currently used do not perfectly achieve this ideal goal! Typically, the speaker's model is optimized by adapting a « world model » (also represented by a large set of multi-Gaussian parameters) to the targeted speaker.

During authentication of a new utterance, acoustic features are extracted, and the probability that the model (*customer model*) of the claimed identity could have generated this utterance is calculated and compared to the probability of the « *world model* » (corresponding to « anybody else »). If the probability of the customer model is greater than the probability of the world model, the system authenticates the speaker, otherwise it rejects it. These two probabilities correspond to the two dimensional space represented in the figure above, and the decision threshold is represented by the line (also optimized on a separate training data) separating at best the two classes.

For the study of the latest recording attributed to bin Laden, it was not possible to build an Arabic « world model » given the limited amount of recordings. We thus used an English world model available at IDIAP. Of course, the bin Laden model was optimized by using many recordings, of different quality, attributed to bin Laden. The decision threshold was also optimized by using bin Laden's recordings (customer accesses), as well as all the recordings of the other Arabic speakers available (impostor accesses).

A few words about IDIAP

Currently numbering about 60 scientists, IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence) is a semi-private, non-profit, research institute located in Martigny, Switzerland, and carrying research and development in the fields of automatic speech and speaker recognition, computer vision, and machine learning. Involved in numerous national and international research projects, IDIAP is also the Leading House of a National Research Center of Competence (NCCR) in « Interactive Multimodal Information Management ».