Can a Professional Imitator Fool a GMM-Based Speaker Verification System?

Johnny Mariéthoz [1]   Samy Bengio [2]

IDIAP–RR 05-61

January 11, 2006

[1] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, marietho@idiap.ch

[2] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, bengio@idiap.ch

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-61

Can a Professional Imitator Fool a GMM-Based Speaker Verification System?

Johnny Mariéthoz Samy Bengio

January 11, 2006

Abstract. This paper presents an attempt at assessing empirically how a state-of-the-art text-independent speaker verification system behaves when confronted to imposting attempts from a professional imitator who perfectly knows how to imitate in particular the clients he tried to impost. Empirical evidence show that, fortunately, current speaker verification systems are indeed robust to such attempts, even when humans are not able to discriminate between true and imposting accesses (a website with some examples is provided to convince the reader). Furthermore, we show that the knowledge of the lexical content of the access significantly helps the imitator, although fortunately not enough to fool the system. This study thus represents a first step in assessing a speaker verification system against true, informed, impostors.

Contents

1 Introduction

Person authentication systems are in general designed in order to let genuine clients access a given service while forbidding it to impostors. In order to design robust person authentication systems, most state-of-the-art solutions are based on training models using data collected from true clients. Unfortunately, in order to control how such a system is robust to impostor attacks, one would theoretically need true impostors trying to enter the system. Such information is of course rarely available, in particular for the domain of speaker verification, discussed in this paper. Hence, most state-of-the-art solutions assume that the accesses of other clients can be used to simulate impostor accesses.

One question thus still remains open: how would a professional impostor perform against state-of-the-art speaker verification systems? While such professional impostors are not available, similar information could be gather from professional imitators, as they are trained to imitate the voice of well-known public personalities, in such a way that most human beings are fooled when listening to the mimicked voice instead of the true one.

The focus of this paper is to analyze the performance of a professional imitator simulating the voice of well-known public personalities for which the imitator is a specialist of, while trying to impost a GMM based state-of-the-art text-independent speaker verification system.

This would help in trying to answer several important questions, such as: Are the current state-of-the-art systems robust to real impostors? Are imitators better impostors than average people? Are imitators better impostors on certain clients than others? Does prior knowledge help imitators to fool the system?

Only a few prior works have been found in the literature on that topic. For instance, in [2], imitators are asked to impost clients of the YOHO database. Unfortunately, the imitators are not real professional imitators, and they never tried to impost people they really knew how to impost. In [7], while the authors used real professional imitators, they tried to impost people they didn't know before the experiment. Furthermore, the experiment was done on a text-dependent speaker verification system, using HMM based techniques.

The outline of the paper is as follows: in Section 2, the experimental protocol is described; Section 3 provides a succinct description of the baseline text-independent speaker verification system that was used; Section 4 provides results of the experiments, as well as the underlying analysis. Finally, Section 5 concludes the paper.

2 Experimental Protocol

The starting point of the present experiment is the availability of a professional imitator, *Yann Lambiel*[1], who specializes in imitating Swiss public personalities. Together with his help, we selected three such public personalities which he felt was best able to imitate, and which were available for the experiment. These personalities are *Pascal Couchepin*, Swiss federal minister, *Daniel Brélaz*, mayor of Lausanne, and *Christian Constantin*, head of the Sion Football Club. On top of Yann Lambiel, and in order to assess the relative performance of a professional imitator, we also asked two more persons to try to imitate the 3 chosen personalities: an amateur imitator, and a normal person, not particularly skilled at imitation.

Each of the 3 personalities selected 3 different sentences: an everyday common sentence, a personal typical expression, and a proverb. They were asked to pronounce each sentence 3 times to train their personal model, and between 5 to 20 more times for the test phase.

The imitator went through 3 different scenarios: first he tried to impost the personalities without any knowledge of the pronounced sentences, apart from the category (everyday sentence, typical expression, proverb); then he was revealed the text of the three sentences; and finally he had the opportunity to listen to the actual sentences pronounced by each personality.

[1] http://www1.rsr.ch/lapremiere/la_soupe/new/lambiel.html

Finally, the experimental protocol includes 3 impostors: the professional imitator, an amateur imitator who only tried to impost Mr. Constantin, and a naive imitator, who was simply one of the authors.

3 Baseline System

The state-of-the-art text-independent speaker verification system used in this paper is based on a statistical framework [5, 4]: for each access, we compare the likelihoods of the access being generated by a client model and by a non-client model. These models are implemented as diagonal covariance Gaussian Mixture Models (GMMs). The non-client model is the same for all clients (and often called a *world model* or *universal background model*), and is trained in order to maximize the likelihood of a large population of client accesses using the Expectation-Maximization algorithm. The client model is then adapted from the world model using a Bayesian MAP adaptation technique [1].

The world model was trained over a quite limited corpus of only 20 french speaking male speakers, each pronouncing 3 citation sentences found on the web.

All the sentences were sampled ad 8kHz with a 16bit coding scheme. They were then preprocessed and transformed into 16 so-called LFCC features [6] and their first derivative, as well as the log of the energy, yielding a total of 33 features for each 10ms of raw signal. Finally, a state-of-the-art speech/silence detector similar to [3] was used to get rid of the silence parts of the signal. Note that using LFCC features meant that we did not make use of any prosodic information. Furthermore, as the experimental conditions were controlled, we did not use any score normalization procedure.

All hyper-parameters of the system were tuned previously on a separate task. This tuning step yielded the following setting: The world model was composed of 200 Gaussians, and trained to maximize the likelihood while constraining the variances of each Gaussian to be no lower than 60% of the global variance in order to control the capacity of the model. Then, each client model was adapted from the world model using MAP, with the adaptation factor (which governs how much the client model is influenced by the world model parameters) set to 20%. In fact, only the means of the Gaussians were adapted [4], while the variances and weights were copied from the world model.

Note that since the sentences of the clients were known, we could have used text-dependent models such as Hidden Markov Models, but these require much more data to train the original world model, hence this solution was not used. Instead, we used the more convenient Gaussian Mixture Model, which is normally used in a text-independent framework, but can also be used with success for text-dependent tasks.

Finally, in order to take the final decision of accepting or rejecting an access, a threshold was selected to be the same for all clients, to show that it is not necessary to tune this threshold separately for each client.

4 Results and Analysis

In this section, we provide graphical evidence of the outcome of the experiments. The first and most important result, depicted in Figure 1, shows all the scores of the personalities (clients) and the professional imitator (trying to impost the personalities). Each dot in the graph represents an access. When the dot is either a (blue) filled triangle, square, or round, it comes from one of the personalities, while if it is a (red) non-filled symbol, it corresponds to the professional imitator trying to impost the corresponding personality. Finally, the black line corresponds to the threshold.

As can be seen, with the exception of one access from Mr. Brélaz, which was wrongly considered as coming from an impostor, all other accesses were correctly classified, which means that the imitator was not able to impost any of the personalities. Furthermore, it is worth noting that the incorrectly classified access was in fact a miss-pronunciation from Mr. Brélaz, basically containing several hesitations. Note that this graph includes all accesses from all conditions explained in the protocol.

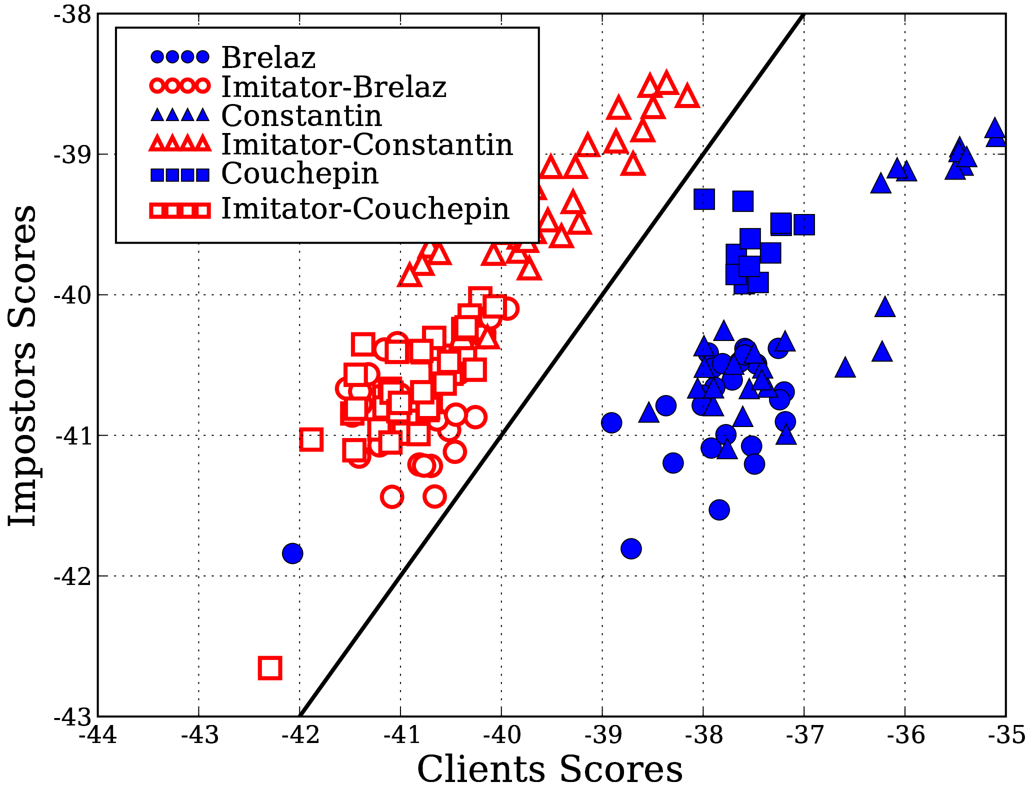

Figure 1: Performance of the professional imitator

This is quite re-assuring as it has often been questioned whether an imitator could impost clients over a speaker verification system. The answer is, according to this experiment, simply no. In the following, we analyze in more details the results of the experiment.

4.1 Is it useful to be a good imitator?

The first question we try to answer concerns the importance of being a good imitator or not. For this, we compare the imposting performance of the professional imitator (Figure 1) with that of an amateur imitator (Figure 2) and a naive imitator (Figure 3). Once again, in each figure, we used the same nomenclature between true accesses (blue filled symbols) and impostor accesses (red empty symbols).

As can be seen, none of the amateur imitator and the naive imitator were able to impost any of the personalities they tried to impost. Furthermore, their imposting performance was worse than that of the professional imitator, showing that it does help to know how to imitate the person one wants to impost, but not enough to fool the system.

4.2 Is it useful to have some knowledge of the pronounced sentences?

In the next series of experiments, we verify whether some knowledge of the content of the sentence pronounced by the clients could be of any help to a professional imitator trying to impost the clients.

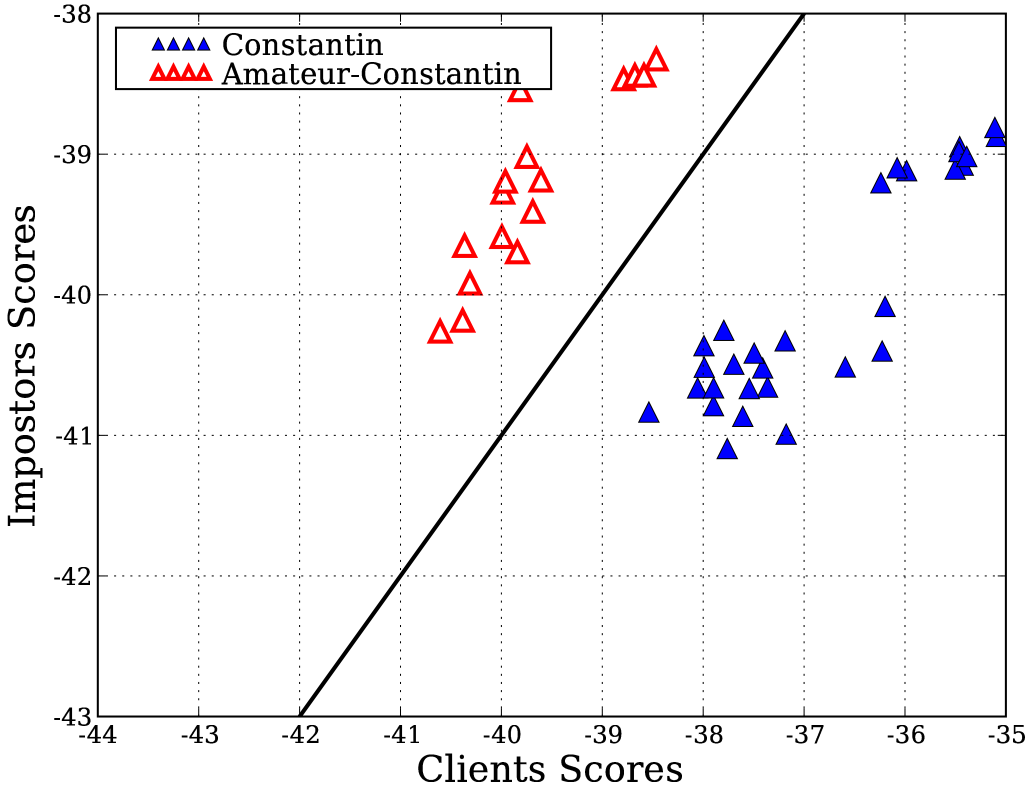

Figure 2: Performance of the amateur imitator

We first present, in Figure 4, the performance of the professional imitator when having no knowledge of the content of the sentences chosen by the clients, apart from the category (everyday sentence, personal citation, proverb). As it can be seen, the system easily separates client and impostor accesses.

In Figure 5, we then present the performance of the professional imitator when he knew the lexical content of the sentences chosen by the clients; in other words, he could have access to a written version of all the sentences, but not a true audio version of them. Comparing Figures 4 and 5, one can clearly see an improvement of the imitator's performance (several impostor accesses are nearer the separating hyperplane), showing that the knowledge did help him significantly, but not enough to fool the system.

Finally, Figure 6 shows the performance of the professional imitator when he had access to a true audio sequence of the sentences pronounced by the clients he wanted to impost. Comparing Figures 5 and 6, it is difficult to see any significant improvement, so while it helps to have access to the true audio sequence, it appears to not be a significant help with respect to the knowledge of the lexical content of the sentences.

5 Conclusion

In this paper, we tried to empirically address the following questions: *Are the current state-of-the-art systems robust to real impostors?* while we could not answer directly to this question, for a lack of true impostors, we used a professional imitator instead, and empirically showed that our speaker

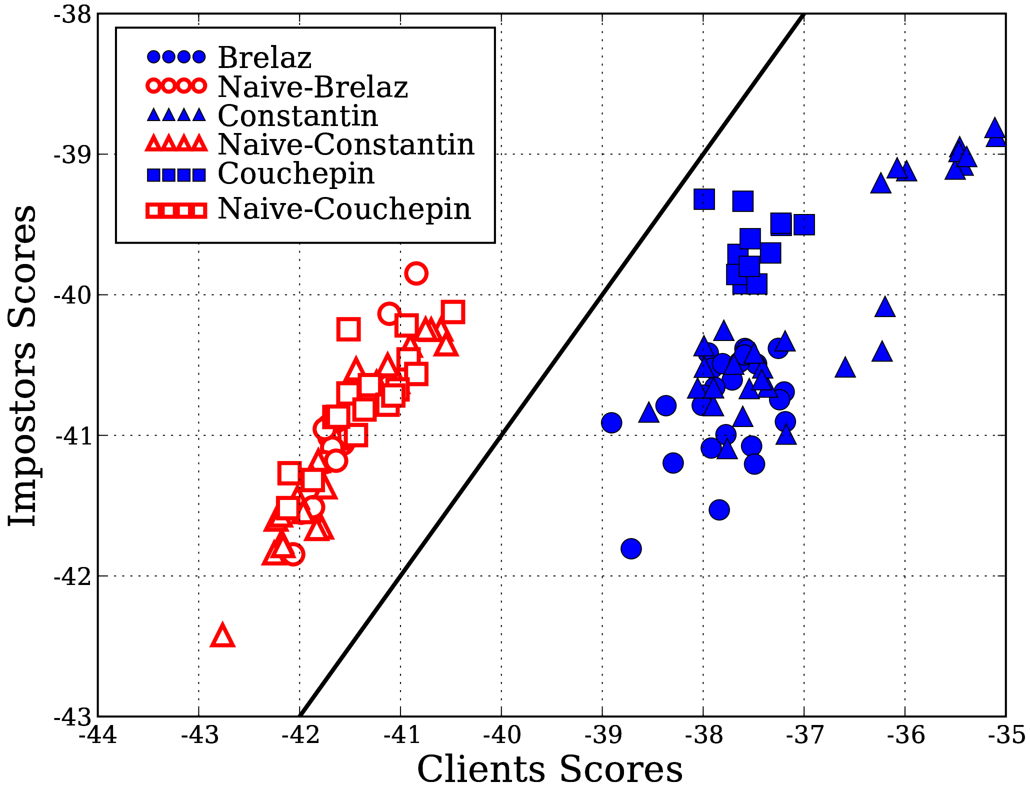

Figure 3: Performance of a naive imitator, which represents the average person

verification system was robust to his imposting attempts in various conditions. *Are imitators better impostors than average people?* Once again, empirical evidences show that this is true. *Are imitators better impostors on certain clients than others?*. Looking at Figure 1, one can see that the professional imitator was significantly better at imposting Mr Brélaz than the two other personalities, hence the answer here is yes, once again. *Does prior knowledge help imitators to fool the system?* Yes, but most importantly the lexical content seems important, and not necessarily the full audio content of the sentences. This might be due to the fact that we did not use any prosodic information in the models, as explained in Section 3.

While the study presented here was performed in controlled conditions, we also invited all the personalities and imitators for a live performance in front of a crowd[2], and even in those uncontrolled conditions, the imitator was never able to impost the system.

Finally, in order to better convince the reader of the difficulty of the task of discriminating between true accesses and the imitator's accesses, we prepared a public website[3] containing several audio clips from the experiment.

[2] In the context of the Swiss 2005' *Science et Cité* event, `http://www.science-et-cite.ch/projekte/festival/fr.aspx`.

[3] http://www.idiap.ch/~marietho/imitations.

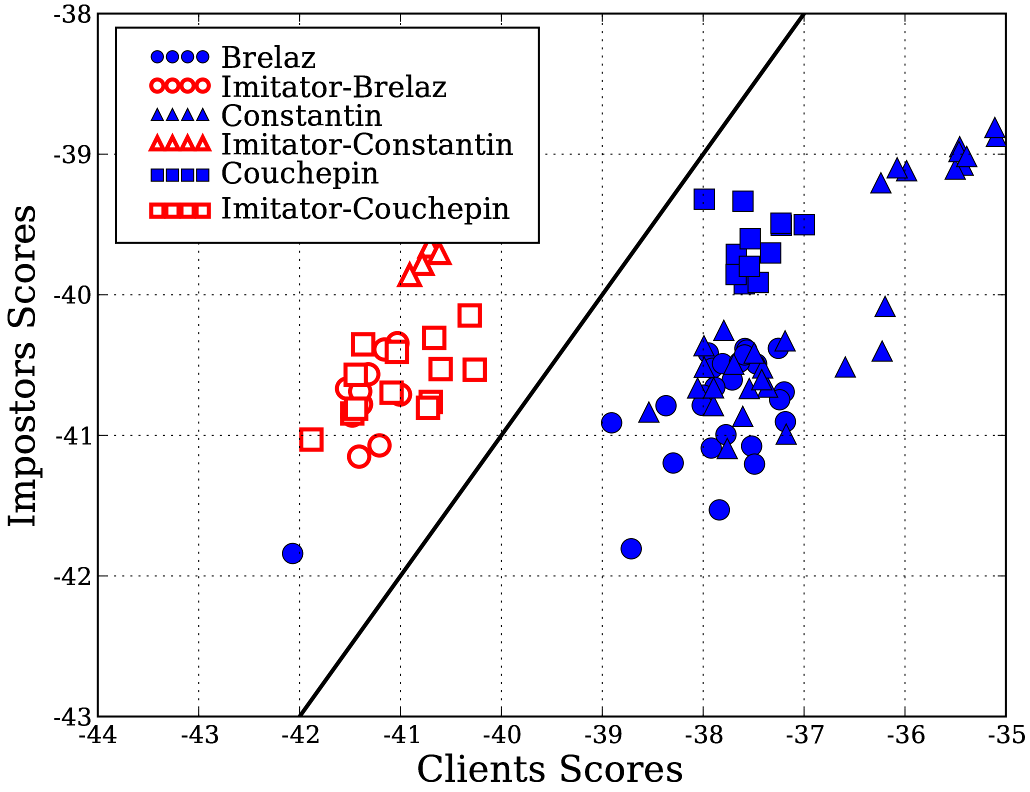

Figure 4: Performance of the professional imitator without any prior knowledge on the content of the client sentences

6 Acknoledgment

References

[1] J. L. Gauvain and C.-H. Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains. *IEEE Tran. Speech Audio Processing*, 2:290–298, 1994.

[2] Y. W. Lau, M. Wagner, and D. Tran. Vulnerability of speaker verification to voice mimicking. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004.

[3] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet. Overview of the 2000-2001 ELISA consortium research activities. In *2001 A Speaker Odyssey*, pages 67–72, 2001.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3), 2000.

[5] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions On Speech and Audio Processing*, 3(1), 1995.

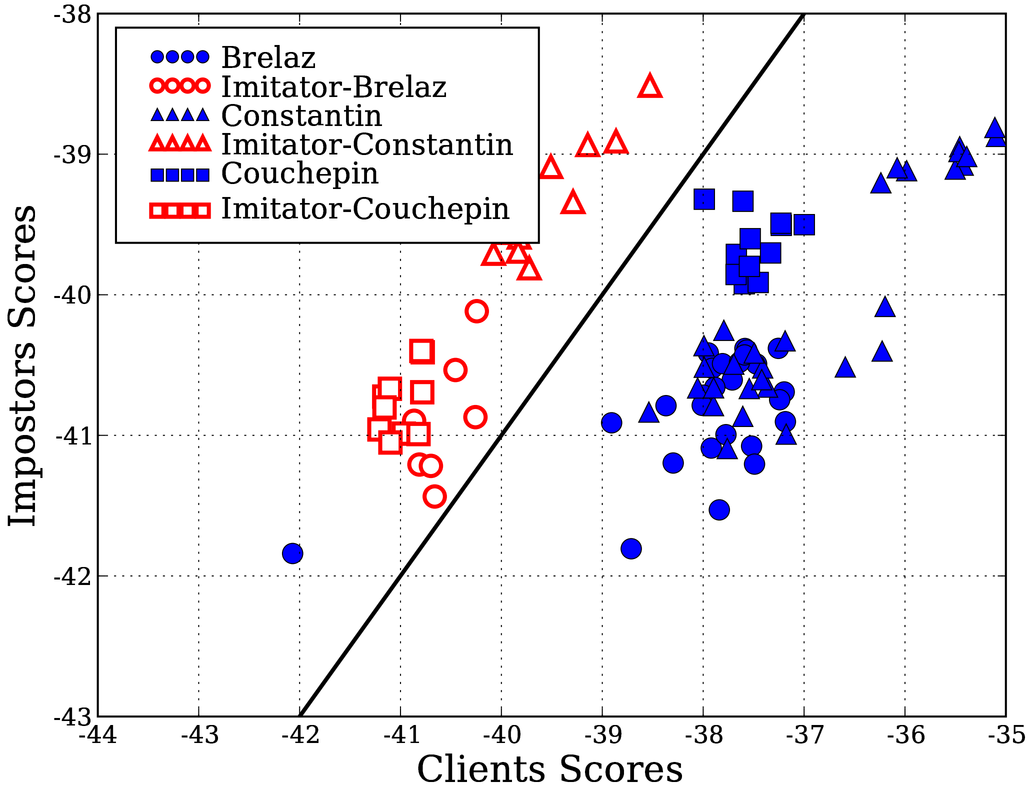

Figure 5: Performance of the professional imitator knowing the lexical content of the client sentences

[6] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B.-H. Juang. A vector quantization approach to speaker recognition. In *Proceedings of the IEEE ICASSP*, pages 387–390, 1985.

[7] E. Zetterholm, M. Blomberg, and D. Elenius. A comparison between human perception and a speaker verification system score of a voice imitation. In *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, 2004.

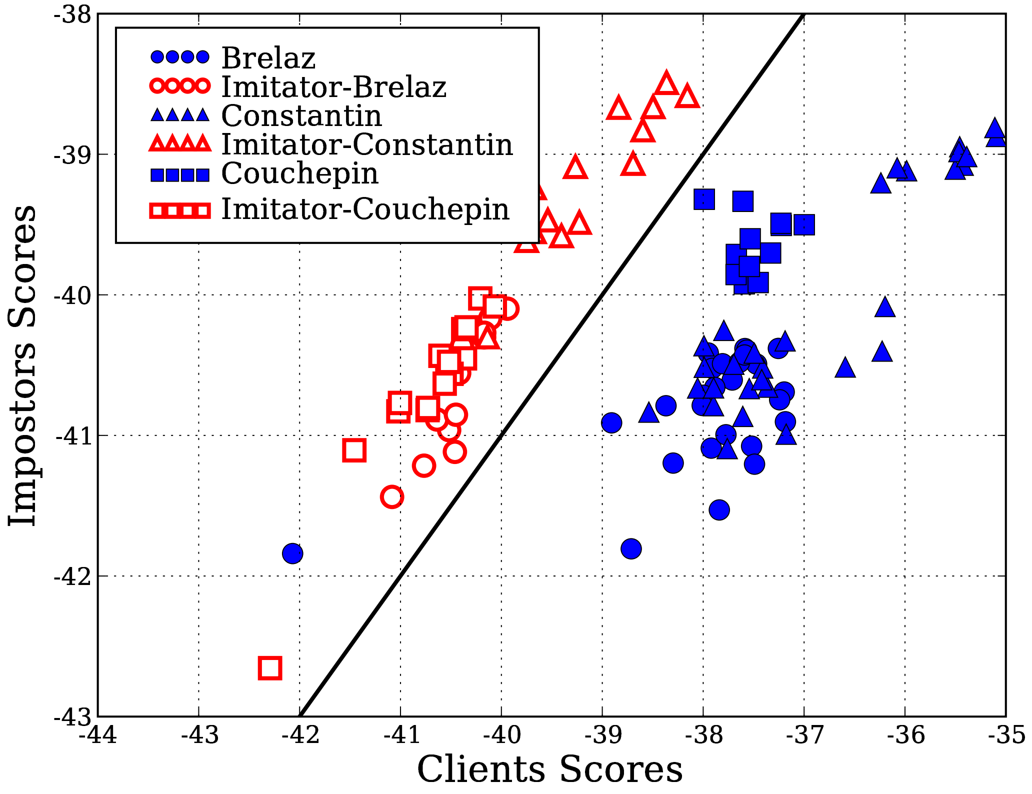

Figure 6: Performance of the professional imitator knowing the audio content of the client sentences