TEXT-INDEPENDENT FORENSIC SPEAKER IDENTIFICATION USING TELEPHONE SPEECH

J.S. Bouten and A.P.A. Broeders
Netherlands Forensic Institute, Rijswijk, the Netherlands

ABSTRACT

On several occasions we have been asked by police officials whether it is feasible for investigative purposes to automatically identify a speaker using speech from intercepted telephone conversations. Since forensic phoneticians are typically not involved in speaker **identification** but in speaker **verification** we tend to reformulate this question in : Is it feasible for investigative purposes to automatically **verify** a speaker's identity using a database built from interceped telephone conversations?

This paper presents an experiment in which a closed set text independent speaker identification method is tested on telephone speech. The speech database used was built from telephone speech recorded for investigative purposes.
The distance measure used for the speaker recognition is related to second order statistical tests. It is expressed as a function of the eigenvalues of a test covariance matrix relative to a training covariance matrix. The text independent recognition technique is based on distance measures developed by Bimbot and Mathan in 1995 [1].

1. INTRODUCTION

When intercepting telephone conversations the Netherlands police are often confronted with speech in foreign languages. The problems encountered with suspects speaking foreign languages are fourfold:

1) Police officials have problems in recognizing the voices of the speakers and therefore have problems categorizing and combining the information gathered from the intercepted conversations.
2) Using interpreters for the identification task in general will result in a high cost factor.
3) For some languages the number of reliable interpreters is limited and since interpreters have been known to be intimidated by criminals police sometimes have problems trusting some interpreters.
4) Since interpreters have not been trained in identifying people their abilities or talents to recognize a person are sometimes limited and some simply refuse to do a job they are not qualified to do.

2. MOTIVATION

The 'Gerechtelijk Laboratorium' (National Forensic Institute), which is part of the Ministry of Justice, has 3 tasks:

1 To supply expert testimony in court cases
2 To advise the law enforcement community on scientific matters (mostly concerning evidence used in court cases)
3 To keep in touch with the latest developments in those fields of science that might be of interest to the law enforcement community and investigate whether methods or tools can be developed from those that can be used by them this community.

This third task lead us to take a look at current developments in automatic speaker identification technology. As argued by Broeders [5] in Stockholm at the International Congress of Phonetic Sciences the rationale behind our interest in an automated procedure for speaker identification is that in large scale police investigations a degree of uncertainty may not be problematic and an informed use of automatic procedures (in carefully controlled forensic conditions) may improve the quality of decisions made by police and lead to considerable savings in time and staff expenditure.

3. TEXT (IN)DEPENDENT SPEAKER RECOGNITION

Speaker recognition technology can be subdivided into two fields. Text dependent speaker recognition and text independent speaker recognition. In text dependent speaker recognition an attempt is made to recognize a speaker by comparing a specific utterance with a template of that same utterance stored in a database. Text dependent speaker recognition technology is often used in systems that try to limit access to a service to those people allowed to use it. These systems in most cases prompt the person trying to gain access for a specific utterance. Since it is in the user's interest to cooperate he will be happy to produce the requested utterance.
When intercepting telephone conversations it is of the utmost importance that the persons speaking on the telephone do not know that their speech is monitored. Since it is our experience that it is unlikely that there will be a repetition of some utterance in different intercepted telephone conversations by the same persons only a text independent speaker recognition methodology might be usable since in text independent speaker recognition a speaker is recognized by comparing an utterance with a template derived from some utterance of that same speaker possibly but not necessarily including that utterance.

4. EXPERIMENT AND RESULTS

In the following we will presents an experiment in which a closed set text independent speaker identification method is tested on telephone speech. The speech database used was built from telephone speech recorded for investigative purposes.

The distance measure used for the speaker recognition is related to second order statistical tests. It is expressed as a function of the eigenvalues of a test covariance matrix relative to a training covariance matrix. The text independent recognition technique is based on distance measures developed by Bimbot and Mathan in 1995 [1].

4.1 Signal Analysis

Starting point is a matrix consisting of a sequence a of M m-dimensional analysis vectors of the speech material of a given speaker. The analysis vectors used in the experiments consist of 17 coefficients of a mel scaled power spectrum. This matrix a will be called the reference or training sequence of M vectors. Next the covariance matrix of this matrix is computed. The same procedure is applied to a sequence of analysis vectors for a test utterance b.

The next step is to seek a measure of resemblance between the covariance matrices A and B. For this the product matrix B*A^-1 is computed. The closer the eigenvalues of this product are to 1, the more A and B are alike. An actual distance measure can thus be expressed as a function of the eigenvalues of this product.

4.2 Experiments and results

Bimbot and Mathan tried several distance measures which are based on combinations of the arithmetic, geometric and harmonic means of the eigenvalues. These similarity measures are based in part on earlier work by Gish [3] and Grenier [4]. Bimbot and Mathan showed that these distance measures are an excellent basis for speaker verification. In most of their experiments they used speech with a bandwidth of 8 kHz from the TIMIT database. Their experiments showed, that on 630 speakers using 1 training sentence and 5 test sentences per speaker a best score of 100 % correct recognition was found. The duration of the training and test sentences in that experiment were 15 seconds. By artificially reducing the bandwidth of the speech from 8 kHz to 4 kHz by reducing the number of filter coefficients used from 24 to 17 they tested the effects of the loss of speaker specific characteristics in the 4 to 8 kHz band on the recognition score. Recognition results in general were down especially when using short utterances for training and test material. These results were verified by using the NTIMIT database. The NTIMIT database was build by recording the TIMIT database through a telephone channel. The experiment showed the arithmetic-geometric sphericity measure to give the best results when using telephone speech (630 speakers, 69 %).

To test the distance measure on telephone speech recorded in the Netherlands we conducted an experiment using telephone conversations that were recorded for investigative purposes. A selection of 13 recorded telephone conversations was used. From these samples of speech of 13 speakers was taken. The speech was samples at 8 kHz using 16 bits. For each speaker a training utterance of 10 seconds and 5 test utterances of 3 seconds were taken. Using the arithmetic-geometric sphericity distance metric [1], 63 out of 65 test utterances were correctly identified.

4.3. Discussion

Although the number of speakers is relatively small in comparison to an actual investigation where typically some 50 suspects are involved, we feel confident that using this technique it might well prove to be practically feasible for operational purposes to introduce an automatic speaker verification procedure using a telephone database.

5. FURTHER RESEARCH

In the near future we are planning to do further experiments using the experiences gained so far. This will involve a large scale test under real-world forensic conditions. Especially the generally more adverse conditions, with temporal drift owing to multi-session recordings, and with a mismatch between training and test recording conditions will have to be carefully controlled. The same goes for the use of channel normalisation techniques. The use of some discriminant analysis to improve the performance and the implementation of a rejection strategy for speaker verification will have to be considered.

In the course of the project an attempt will be made to assess the effect on the recognition performance of factors like speech rate and intensity and communicative factors like interlocutor and type of call.

Since the contents of the database play a crucial role in the recognition process and the suspects of a crime are likely to change from case to case a methodology for building a database has to be developed. Furthermore we believe that in any practical situation the selection of speech segments must be done by a native speaker of the target language, preferably by an experienced phonetician or a linguistically informed speech scientist. To find such a person may well prove to be the more difficult part of the exercise.

REFERENCES

[1] Bimbot, F., Mathan L. (1993) 'Text-free Speaker Recognition using an arithmetic-harmonic Sphericity Measure', *Proceedings of eurospeech*, Berlin, pp. 169-172

[2] Bimbot, F., Magrin-Chagnolleau, I & Mathan L. (1995), 'Second-order statistical measures for text-independent speaker identification', *Speech Communication* 17, pp. 177-192

[3] Gish, H. 'Robust discrimination in automatic speaker identification', *IEEE-ICASSP*, 1990.

[4] Grenier, Y. '*Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*', PhD ENST-E-77005, 1977

[5] Broeders, A.P.A. 'The role of speaker recognition techniques in forensic investigations', *Proceedings of the International Conference of Phonetic Sciences 1995*, Vol 3, pp. 154-161