VOICE FORGERY USING ALISP: INDEXATION IN A CLIENT MEMORY

Patrick Perrot[1, 2]*, Guido Aversano*[1]*, Raphaël Blouet*[1]*, Maurice Charbit*[1]*, Gérard Chollet*[1]

[1] Ecole Nationale Supérieure des Télécommunications, dpt TSI
46, rue Barrault 75634 Paris
France

[2] Institut de Recherche Criminelle de la Gendarmerie Nationale, dpt SIP
(member of European Network of Forensic Science Institutes)
1, boulevard Théophile Sueur
93111 Rosny sous bois cedex
France

ABSTRACT

This article deals with a technique of voice forgery using the ALISP (Automatic Language Independent Speech Processing) approach. Such a technique allows to transform the voice of an arbitrary person (the impostor), forging the identity of another person (the client). Our goal is to demonstrate that an automatic speaker recognition system could be seriously threatened by a transformation of this kind. For this purpose, we use a speaker verification system to calculate the likelihood that the forged voice belongs to the genuine client.

Experiments on NIST 2004 evaluation data show that the equal error rate for the verification task is significantly increased by our voice transformation.

1. INTRODUCTION

Biometric identity verification systems are today increasingly sophisticated. Multimodal systems (combining many different features, like voice, face, fingerprints, iris, signature, etc.) provide interesting results. However, we must take into consideration the risk of forgery. The reliability of a biometric system is its ability to cope with different forgery scenarios. Voice imposture constitutes a good example of threat for security systems.

In this paper a novel technique to realize text-independent speech transformation is presented. This method will be tested in an identity verification forgery scenario.

Firstly, it is important to define how we conceive a vocal forgery, distinguishing between speech modification and speech conversion. A speech modification would consist in changing some characteristics of the voice, such as pitch, timing, etc., without the intention of matching it with the voice of another individual. On the contrary, a voice conversion (or transformation) occurs when there is the explicit will to convert the voice of a person X (impostor) in the one of another person Y (client).

The kind of forgery we focus on can be realized in two ways:

- by a professional impersonator
- by automatic transformation of voice.

The first method has been already studied in literature [4]. Indeed, these studies evidence that the impersonation consists in imitating some specific characteristics of the client voice which are sufficient to disturb a human ear. These characteristics are pitch register, voice quality, dialect, prosody, and speech style. An impersonator cannot imitate all the aspects of the voice, but he can anyway succeed in the impersonation. However it is not possible to establish a general ranking order of important features. So, even if a human ear could be tricked by those kinds of impersonation, the literature reports that an automatic speaker verification system normally recognizes the forgery.

The second method, i.e. automatic voice transformation is what is studied in this paper. Different automatic voice conversion techniques will be succinctly presented, then a new original technique will be detailed, based on the ALISP (Automatic Language Independent Speech Processing) approach [7]. The proposed system, unlike the other presented methods, allows text-independent voice conversion.

Forgery experimental results are obtained by the BECARS speaker recognition system that is used to calculate and compare identity verification scores on non-forged and forged data. Experiments were carried out on speech data from 2004 NIST Speaker Recognition Evaluation.

2. VOICE TRANSFORMATION

To realize an automatic forgery as explained above, different techniques are possible, also according to the available quantity for the client voice.

If only a limited amount of client data is available, it could be interesting to consider spectral conversion techniques, which give some interesting results according to previous studies. Consider a sequence of spectral vectors pronounced by the impostor, $X = [x_1, x_2, \ldots, x_n]$, and a sequence composed by the same words, pronounced by the client, $Y = [y_1, y_2, \ldots, y_n]$.

A spectral transformation can be performed by finding the conversion function F that minimizes the mean square error: $\varepsilon_{mse} = E[||y - F(x)||^2]$, where E is the expectation.

It is necessary to calculate a conversion function which maps the features of the impostor to the features of the client. Different methods have been studied to calculate the conversion function, e.g. vector quantization using a mapping codebook [1], Gaussian Mixture Models (GMM) with least squares estimation, GMM with joint density estimation [6, 9, 10] or dynamic frequency warping [12].

This approach works if the impostor pronounces the same sentence or word(s) as the client; it is a text dependent method.

This paper proposes another method exploitable by impostors to reproduce the voice of an authorized client. In particular, speech segments obtained from client recordings can be used to synthesize new sentences that the client has never pronounced. Below it will be explained how a very-low bit-rate speech coding system can be adapted to serve forgery purposes, transforming any input speech into client's voice.

3. THE ALISP SYSTEM

In order to automatically transform the voice, we use an ALISP-based voice encoder [2]. The principle of this system is to encode speech by recognition and synthesis in terms of basic acoustic units that can be derived by an automatic analysis of the signal. Such analysis is not based on a priori linguistic knowledge.

Firstly, a collection of speech segments is constituted by segmenting a set of training sentences, all pronounced by the target client voice. This step is performed using the temporal decomposition algorithm [3] on MFCC speech features.

Segments resulting from temporal decomposition are then organized by vector quantization into 64 different classes. The training data is thus automatically labelled, using symbols that correspond to the above classes.

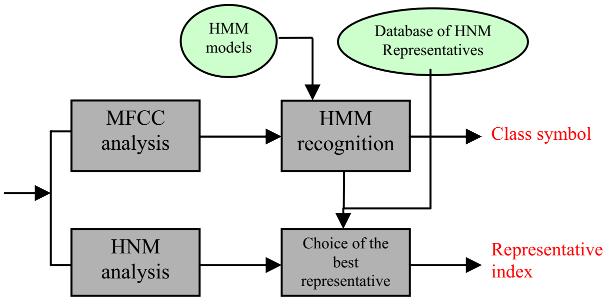

Figure 1. The ALISP encoding process.

A set of HMMs (Hidden Markov Models) is then trained on this data, providing a stochastic model for each ALISP class. Iterative re-estimation of model parameters is performed, which also produces a new segmentation of the training corpus. The result of the ALISP training is an inventory of client speech segments, divided into 64 classes according to a codebook of 64 symbols. The chosen number of classes ($64 = 2^6$) is comparable with the number of phonetic categories. A set of 64 HMMs trained to recognize the codebook units is also obtained.

All the speech segments contained in our inventory are represented by their Harmonic plus Noise Model (HNM) parameters [13]. This will allow a smooth concatenative synthesis of new sentences using the stored segments.

During the training, statistics on client's prosody are also collected and stored. In the current experiments, only the mean and the standard deviation of client's *F0* are retained.

The second part of our processing consists in encoding the impostor's voice using the above ALISP codebook, and then in performing decoding using synthesis units taken from the segment inventory obtained from client's voice.

Figure 1 shows how ALISP encoding works. After the extraction of MFCC features, the encoder uses previously trained HMM models to recognize the sequence of ALISP symbols in impostor's speech segment. For each recognized symbol, the system searches for a good representative unit in the client's segment inventory. This search is limited to the class corresponding to that symbol. The best representative is chosen by comparing the spectral envelope of source (impostor) speech with those of client segments. This comparison is performed by Dynamic Time Warping (DTW) on HNM parameters.

From a sentence pronounced by an impostor, the ALISP encoding produces a sequence of class symbols and within-class indexes of representative units.

The new voice-transformed sentence can now be synthesized by concatenating client speech segments, which are taken from the available inventory according to the above encoding information. The adopted HNM synthesis mechanism allows modifying the length and the prosody of the segments, according to the dynamics of the input voice.

Anyway, to reproduce client's prosody behavior we transform impostor's *F0* trajectory, adapting it to the statistics calculated on client's speech.

4. THE REFERENCE VERIFICATION SYSTEM: BECARS

To evaluate the efficacy of our forgery the BECARS automatic speaker recognition system has been chosen as reference. BECARS is one of the best available verification systems, according to the last NIST evaluation campaign [5]. It consists in several open-source tools that allow to train and run Gaussian Mixture Models (GMM) for speaker recognition and verification tasks. The main feature of this software is the possibility of using several adaptation techniques including *Maximum A Posteriori* (MAP) and *Maximum Likelihood Linear Regression* (MLLR).

The verification task consists in detecting whether a speech sentence has been uttered by the claimed identity or by an impostor. The decision is based on a score, given by the likelihood ratio between a client model and a world model.

The adopted method for score computation is based on the use of a hierarchical Gaussian clusterization technique that is described in details in [5]. Our verification is performed on 20 MFCCs with the associated dynamic features. Moreover cepstral mean subtraction and feature warping are used to normalize the parameters.

5. EVALUATION AND DATABASE EXPERIMENTAL PROTOCOL

Our experiments were carried out on data from the NIST 2004 evaluation. As it is described in the NIST evaluation plan, these data are mostly some excerpts of conversational telephone speech in English. However we also found in it other languages different from English and speech recorded by different microphones. In addition, data are provided without a prior removal of silence intervals.

Silences constitute a big part of available data and this causes problems for modeling and verification. Thus, removal of silence, based on signal energy, is performed. The threshold for speech activity detection is calculated by modeling silence and speech with two Gaussian functions. The chosen threshold value corresponds to the intersection of the two distributions. If we find at least 20 adjacent frames (100ms) that are all under the threshold, that segment is eliminated.

For verification tests we adopted the NIST 2004 "1side-1side" core evaluation protocol that consists in adapting client GMMs on 5 minutes (including silence) of client's speech and verifying speaker identity on 5 minutes (including silence) test sentences. The world model has been built using NIST 2000–2001 data.

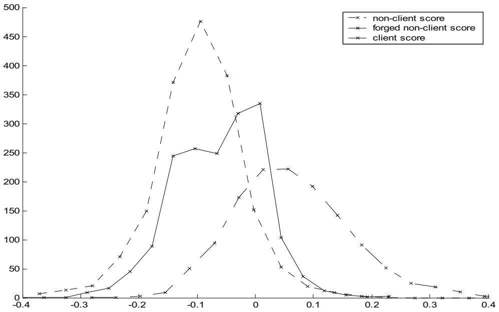

Figure 2. Score frequency distributions for impostor, transformed impostor and client speech.

On the other hand, ALISP training is performed on all the available client data (with an average of about 50 minutes per client), without silence removal.

The NIST evaluation protocol also includes the list of tests to perform. A test consists in verifying the claimed identity of a speech sentence. We can distinguish between client accesses and impostor accesses, according to whether or not the claimed identity corresponds to the true identity of the speaker.

For the purposes of this paper, we limited our experiments to 1320 client accesses and 1729 impostor accesses taken from the NIST list.

6. EXPERIMENTS AND RESULTS

Our reference verification performance is obtained by evaluating the verification score of both clients and impostors, without applying any voice transformation.

Afterwards, non-client speech sentences are transformed, trying to forge the voice of the corresponding claimed identity. Having applied the voice transformation, new scores are calculated for impostor accesses.

Figure 2 shows the score frequency distributions found respectively for impostors, transformed impostors and clients.

We can observe that client and impostor distributions have a regular Gaussian shape, while the transformed impostor distribution presents two peaks in correspondence with the peaks of the previous two distributions.

It appears that more than one half of the impostor scores were shifted towards typical client values by our voice transformation. Anyway, we can find in the above figure that there is a clear separation between successful and unsuccessful transformations. Unsuccessful ones are probably due to a poor training of the ALISP system, caused by the great amount of silence that characterizes

NIST data. It is possible to trace Detection Error Tradeoff (DET) curves [11] using the obtained scores. Figure 3 shows the two curves corresponding to original and forged data.

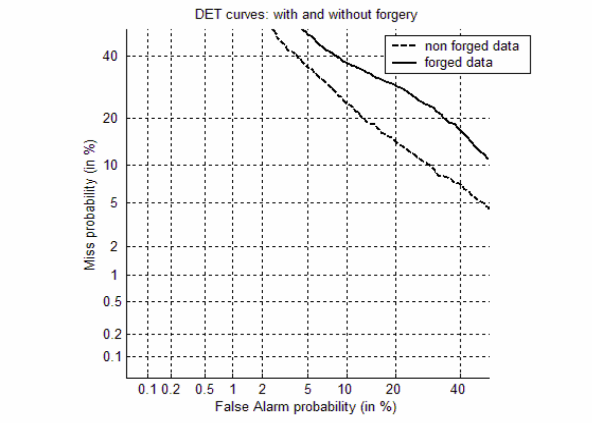

Figure 3. DET curves for verification tests on original and forged speech data.

The equal error rate (EER) found for verification tests on non-forged data was 16%. After voice transformation, the equal error rate was increased to 26%. Note that the statistical uncertainty for the above results is 2%, corresponding to a 95% confidence interval.

7. CONCLUSION AND PERSPECTIVES

The voice transformation method presented in this paper has been realized by simply adapting ALISP techniques originally thought for speech coding. Results obtained on the NIST 2004 evaluation database show that GMM-based speaker verification is not enough robust to this kind of forgery. Future improvements to the proposed technique are expected to lead to an even higher verification error rate. In fact, many aspects of the current system can be ameliorated. Future works will include a revision of the ALISP training procedures, to account for the extreme heterogeneity of NIST speech data. Improvements to the HNM analysis/synthesis mechanism and prosody modeling are also planned.

Better voice conversions could be also obtained by setting up a preliminary functional transformation between impostor's and client's acoustic spaces. This will allow a more accurate recognition of ALISP units in impostor speech, by the HMMs trained on client data.

A final possible perspective is the identification of clues for an automatic detection of forgery in speaker verification systems.

8. AKNOWLEDGEMENTS

This work was supported by the IST SecurePhone project. The BECARS system results from collaboration between ENST and the University of Balamand. The development of the ALISP encoder was partially supported by the RNT-SYMPATEX project.

9. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP 88,* New-York, 1988

[2] G. Baudoin, J. Cernocky, F. El Chami, M. Charbit, G. Chollet, D. Petrovska-Delacretaz. "Advances in Very Low Bit Rate Speech Coding using Recognition and Synthesis Techniques," *Proc. Of the 5th Text, Speech and Dialog workshop, TSD 2002*, Brno, Czech Republic, pp. 269-276, 2002

[3] F. Bimbot, G. Chollet, P. Deleglise, C. Montacié, "Temporal Decomposition and Acoustic-phonetic Decoding of Speech," *Proc. ICASSP 88*, New-York, pp. 445-448, 1988

[4] M. Blomberg, Daniel Elenius, E. Zetterholm, "Speaker verification scores and acoustics analysis of a professional impersonator," *Proc. FONETIK 2004*

[5] R. Blouet, C. Mokbel, G. Chollet, "BECARS: a free software for speaker recognition," *ODYSSEY 2004*, Toledo, 2004

[6] O. Cappe, Y. Stylianou, E Moulines, "Statistical methods for voice quality transformation," *Proc. of EUROSPEECH 95*, Madrid, Spain, 1995

[7] G. Chollet, J. Cernocky, A. Constantinescu, S. Deligne, F . Bimbot, "Toward ALISP: a proposal for Automatic Language Independent Speech Processing," *Computational Models of Speech Processing*, NATO ASI Series, 1997

[8] D. Genoud, G. Chollet, "Voice transformations: some tools for the imposture of speaker verification systems," *Advances in Phonetics*, A. Braun (ed.), Franz Steiner Verlag, Stuttgart, 1999

[9] A. Kain, M. W. Macon, "Spectral voice conversion for text to speech synthesis," *Proc. ICASSP 98,* New-York, 1998

[10] A. Kain, M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," *Proc. ICASSP 01,* Salt Lake City, 2001

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET curve in assessment of detection task performance," *Proc. EUROSPEECH 97*, Rhodes, Greece, pp. 1895-1898, 1997

[12] H. Valbret, E. Moulines, J.P. Tubach, "Voice transformation using PSOLA technique," *Proc. ICASSP 92*, San Francisco, 1992

[13] Y. Stylianou, O. Cappe, "A system for voice conversion based on probabilistic classification and a harmonic plus noise model," *Proc ICASSP 98*, Seattle, WA, pp. 281-284, 1998