EFFECT OF SPEECH TRANSFORMATION ON IMPOSTOR ACCEPTANCE

Driss Matrouf, Jean-François Bonastre, Corinne Fredouille

LIA, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
{driss.matrouf,jean-francois.bonastre,corinne.fredouille}@univ-avignon.fr

ABSTRACT

This paper investigates the effect of voice transformation on automatic speaker recognition system performance. We focus on increasing the impostor acceptance rate, by modifying the voice of an impostor in order to target a specific speaker. This paper is based on the following idea: in several applications and particularly in forensic situations, it is reasonable to think that some organizations have a knowledge on the speaker recognition method used and could impersonate a given, well known speaker. This paper presents some experiments based on NIST SRE 2005 protocol and a simple impostor voice transformation method. The results show that this simple voice transformation allows a drastic increase of the false acceptance rate, without a degradation of the natural aspect of the voice.

1. INTRODUCTION

Speech is a compelling biometric for several well-known reasons and particularly because it is the only one available modality in a large set of situations. Even if this biometric modality presents lower performance compared - for example - to iris, the progress achieved during the last decades brings the automatic speaker recognition systems at a usable level of performance for commercial applications. During the same period, in the forensic area, judges, lawyers, detectives, and law enforcement agencies have wanted to use forensic voice authentication to investigate a suspect or to confirm a judgment of guilt or innocence [1][2]. Despite the fact that the scientific basis of person authentication by his/her voice has been largely questioned by researchers [3][4][5] and the "need of caution" message sent by the scientific community in [6], forensic speaker recognition methods are widely used, particularly in the context of worldwide terrorism events. Some recent developments show the interest of Bayesian based methods in forensic speaker recognition [7][8][9]. This approach allows to present more precisely the results of a voice identification to a judge. If it is a real progress, it does not solve several problems linked on how the method is evaluated, how hypotheses are defined or how the confidence on the expert is taken into account.
This paper investigates a different point: if you know the identification method used by the expert, if you have a voice excerpt of the target person *X*, is it possible to transform the voice of someone else in order to obtain a positive identification ? Of course, the transformed voice should correspond to a natural voice.

In this paper we investigate this possibility, by using a state-of-the-art GMM based text-independent speaker detection system and a simple Gaussian-Dependent Filtering technique for impostor voice transformation. The objective is close to the voice-forgery approach proposed in [15] even if our goal is only to obtain positive system decisions for impostors (without loosing the natural aspect of the voice) and not to synthesize a voice excerpt close to the target speaker for a human perception point of view.

This paper is organized as follows. Firstly, the speaker recognition framework is presented in section 2. The proposed voice transformation method is presented in section 3. A set of experiments and the corresponding results are presented in section 4 and 5, using NIST 2005 SRE framework. Some conclusions and future work are proposed in section 6.

2. GMM-UBM SPEAKER RECOGNITION APPROACH

GMM-UBM is the predominate approach used in speaker recognition systems, particularly for text-independent task [10]. Given a segment of speech Y and a speaker S, the speaker verification task consists in determining if Y was spoken by S or not. This task is often stated as basic hypothesis test between two hypotheses: Y comes from the hypothesized speaker S ($H0$), and Y is not from the hypothesized speaker S ($H1$). A likelihood ratio (LR) between these two hypotheses is estimated and compared to a decision threshold θ. The LR test is given by:

$$LR(Y, H0, H1) = \frac{p(Y|H0)}{p(Y|H1)} \tag{1}$$

where Y is the observed speech segment, $p(Y|H0)$ is the likelihood function for the hypothesis $H0$ evaluated for Y, $p(Y|H1)$ is the likelihood function for $H1$ and θ is the decision threshold for accepting or rejecting $H0$. If $LR(Y, H0, H1) > \theta$, $H0$ is accepted else $H1$ is accepted.
A model denoted λ_{hyp} represents $H0$, it is learned using an extract of speaker S voice. The model $\lambda_{\overline{hyp}}$ represents the alternative hypothesis, $H1$, and is usually learned using data gathered from a large set of speakers.
The likelihood ratio statistic becomes $\frac{p(Y|\lambda_{hyp})}{p(Y|\lambda_{\overline{hyp}})}$. Often, the logarithm of this statistic is used giving the $logLR$ (LLR):

$$LLR(Y) = log(p(Y|\lambda_{hyp})) - log(p(Y|\lambda_{\overline{hyp}})). \tag{2}$$

In the presented approach, the models are Gaussian Mixture Models which estimate a probability density function by:

$$p(x|\lambda) = \Sigma_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i) \tag{3}$$

where w_i, μ_i and Σ_i are weights, means and covariances associated with the Gaussian components in the mixture. Usually a large number of components in the mixture and diagonal covariance matrices are used.

The model $\lambda_{\overline{hyp}}$ is denoted world model or Universal Background Model (UBM) when the model is environment independent. Its parameters are estimated using the EM algorithm. The speaker model λ_{hyp} parameters are generally obtained by adapting the world model parameters, using the Bayesian adaptation framework. Generally, only mean parameters are adapted and the other parameters remain unchanged [11].

3. SPEECH TRANSFORMATION

Our goal in this paper is to transform speech signal belonging to a speaker in order to increase its likelihood given the GMM corresponding to another speaker. Listening the resulting signal, the effects of the transformation must appear as natural as possible. The principle retained in this paper is to analyze the impostor signal frame by frame and to transform each frame in order to get it closer to the target speaker GMM. Of course, the work should be done in the speaker recognition feature space, generally obtained by a cepstral parameterization followed by a feature normalization process (based on speech/non speech frame detection). The main constraint of this approach is to transform the impostor speech signal when the objective is to move the signal in the targeted automatic speaker recognition (ASR) feature space.

In order to achieve this objective, we use two parallel sets of acoustic models, with a one-to-one mapping between Gaussian components, for a target speaker S. The first one is in the targeted speaker recognition feature space (cepstral plus feature normalization). This model is denoted "automatic speaker recognition" (asr) model or "master model" and is used in order to estimate the a posteriori probabilities of the GMM Gaussian components given each frame. The second one, denoted here "filtering" model (fil), is used for estimating the optimal time-varying filter parameters using the probabilities given by the master model. In this paper, we used LPCC parameterization (instead of LPC) for the filtering model (without feature normalization) as this representation is well suited for GMM modeling. This parallel model based technique increases the independence between the transformation process and the targeted speaker recognition system.

Let Y be the signal to transform. Y is the corresponding set of frames: $Y = \{y_1, \ldots, y_n\}$. Let us consider y, a frame of speaker S' (the impostor) and x its corresponding frame of the speaker S (the targeted speaker). The source-filter model leads to the following relations in the spectral domain:

$$Y(f) = H_y(f)S_y(f) \tag{4}$$

$$X(f) = H_x(f)S_x(f) \tag{5}$$

where Y and X are the spectral representations of y and x. H_y and H_x are the transfer functions corresponding to both x and y; S_x and S_y are the Fourier transforms of the source signals corresponding to x and y. It is important to note that the cepstrum is nothing other than a compact representation of the transfer function H. So, to bring y as close as possible to x - in terms of spectral slope - it is enough to replace H_y with H_x in equation 4:

$$Y'(f) = H_x(f)S_y(f) = \frac{H_x(f)}{H_y(f)}Y(f) \tag{6}$$

We call H_x the target transfer function and H_y the source transfer function. If we decide not to modify the phase of the original signal, the filter to apply to the signal y becomes:

$$H_{yx}(f) = \frac{|H_x(f)|}{|H_y(f)|} \tag{7}$$

In this paper the transfer functions are estimated as follows:

$$H_x(f) = \frac{G_x}{A_x(f)} \tag{8}$$

$$H_y(f) = \frac{G_y}{A_y(f)} \tag{9}$$

where $A_x(f)$ and $A_y(f)$ are the Fourier transforms of the prediction coefficients of the signals x and y, G_x and G_y are the gains of the residual signals s_x and s_y (S_x and S_y are the spectral representation of s_x and s_y). The source, the gain and the prediction coefficients of y are obtained directly from y. The source, the gain and the prediction coefficients of x are obtained from the LPCC coefficients corresponding to the filtering-model Gaussian component having generated the frame y in speaker S model.
This scheme is generalized by using all the components with different *a posteriori* probabilities. The target transfer function is derived from the linear combination of all the filtering GMM means weighted by their *a posteriori* probabilities. The *a posteriori* probabilities are estimated thanks to the ASR GMM.

$$x_{fil} = \Sigma_{i=1}^{M} p(g^i_{asr}|y)\mu^i_{fil} \tag{10}$$

where x_{fil} is the target representation (at the filtering level) of H_x, $p(g^i_{asr}|y)$ the *a posteriori* probability of Gaussian component i given the frame y. μ^i_{fil} is the mean of the Gaussian g^i_{fil} corresponding to the Gaussian g^i_{asr} (with the bijection strategy). The target prediction coefficients are estimated from x_{fil} by using a lpcc-lpc transformation. Figure 1 presents a block diagram of impostor frame transformation.

Synthesis of the transformed signal is done frame by frame independently using the standard overlap-add technique with Hamming windows, where the resulting signal is obtained by adding the resulting window-based signals.

4. EXPERIMENTAL PROTOCOL

In this section, we present the experimental protocols used to demonstrate that the described signal transformation can disturb the speaker recognition system. Experiments are conducted in the context of the NIST SRE 2005 evaluation campaign organized by NIST in April 2005 (see the evaluation plan for more details [12]).

4.1. Experimental corpora

Experimental corpora used in this paper are extracted from the NIST SRE 2005 evaluation campaign. This campaign is focused on the evaluation of the automatic speaker recognition systems on conversational telephone speech (speaker detection task). Two main corpora are available in this context: an evaluation data set issued from

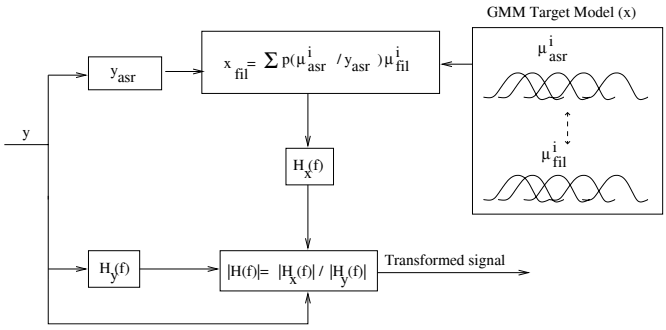

Fig. 1. Transform block diagram for one frame: The target transfer function H_x is estimated by using 2 parallel GMM, with one-to-one component tying. The fist one allows the *a posteriori* probability estimation; and the second one is used for filtering.

the Mixer corpus and a development data set, used for the system development and tuning, issued from the previous evaluation campaigns.
In this work, two main corpora are derived from the official ones:

- the corpus $Eva05$, composed of the male speakers of the official evaluation data set. This corpus, including 1231 client trials and 12317 impostor trials, is used for the testing phase;
- the corpus $Dev05$, composed of male speakers and derived from the official development data set. This corpus is used for the UBM world model training, required for the speaker recognition baseline system and the voice transform process as well as for the T-Norm score normalization required only for the speaker recognition system.

In order to evaluate the behavior of the voice transformation process described in this paper when combined with a state-of-the-art speaker recognition system, similar speaker recognition testing phases are conducted with and without voice transformation on the NIST SRE $Eva05$ corpus. In the voice transformation case, each impostor trial is carried out by comparing the right target model (claimed speaker id) and the transformed impostor test signal. The target trials remain unchanged for both cases.

4.2. Baseline speaker recognition system

The LIA_SpkDet system [13] developed at the LIA lab is used as baseline in this paper. Built from the ALIZE platform [16][14], it was evaluated during the NIST SRE'04 and SRE'05 campaigns, where it obtained about the best performance for a cepstral GMM-UBM system. Both the LIA_SpkDet system and the ALIZE platform are distributed under an open source licence.
The LIA_SpkDet system is based on classical UBM-GMM models and T-Norm approach for likelihood score normalization. For the front-end processing, the signal is characterized by 32 coefficients including 16 linear frequency cepstral coefficients (LFCC) (Filter-bank analysis) and their first derivative coefficients. A frame removal based on a three component GMM energy modeling is computed. A mean and variance normalization process is finally applied on coefficients. The world and target models contain 2048 components and a top ten component selection is used for likelihood computation.

4.3. Voice transformation model specifications

In this work, the world master-GMM is gathered from the baseline system (*c.f.* 4.2). A world filter-GMM is estimated by using the statistics of the last EM iteration of the world master-GMM estimation, in order to obtain the component to component tying between the two models. A similar process is used for estimating the target models: the target speaker master-GMMs are estimated by adapting only means of the master world-GMM and the target filter-GMM are estimated by adapting means of the world filter-GMM, using the statistics of the corresponding target master-GMM. It is important to notice that the speaker recognition world model and the target speaker training files are used for the voice transformation process.

5. RESULTS

The influence of the impostor voice transformation process on the behavior of a baseline speaker detection system is measured through classical DET performance curves - figure 3 - and through impostor and client score distributions - figure 2. In the latter, 3 T-normalized score distributions are provided (T-Norm is applied for all the results presented in this paper), corresponding to: original impostor trials (without applying the impostor voice transformation process), transformed impostor trials and target trials. A drastic move of the impostor distribution is observed when the voice transformation is applied, giving impostor score mean higher than the target speaker score mean. Regarding the DET performance curves, the performance of the baseline speaker detection system drastically decreases when combined with the impostor voice transformation scheme. This demonstrates that the behavior of the baseline system is largely disturbed by the impostor voice transformation.

A spectrogram of an original impostor speech segment and of the corresponding transformed signal are shown in Figure 4. Listening to several examples of transformed signals, we did not notice any distortion and the signal remained natural.

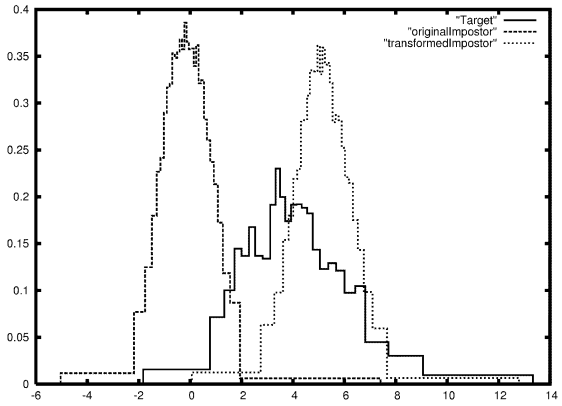

Fig. 2. T-Norm score distributions for original impostor trials (without applying the impostor voice transformation process), transformed impostor trials and target trials.

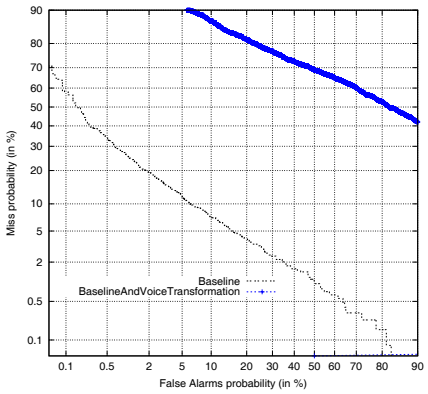

Fig. 3. DET performance curves of the baseline speaker detection system without(thin line) and with applying impostor voice transformation (thick line), NIST $Eva05$ corpus.

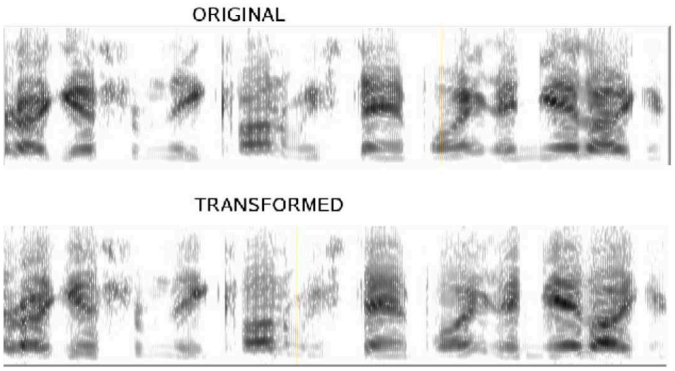

Fig. 4. Example of one impostor signal, spectrograms of original and transformed signals.

6. CONCLUSION AND FUTURE WORK

In this paper we investigate the effect of artificially modified impostor speech on a speaker recognition system. It seems reasonable to think that an organization which wants to attribute a speech segment to a given - well known - speaker has a knowledge of the speaker recognition system used by a specific scientific police department, as well as a general knowledge on the state-of-the-art in speaker recognition. We demonstrate in this paper that, following this hypothesis, it seems relatively easy to transform the voice of someone in order to target a specific-speaker voice, in terms of the automatic speaker recognition system.

In this paper, a complete knowledge of the speaker recognition system was assumed, including the feature extraction process, the modeling method, the world model and the target-speaker training segments. Further experiments will be proposed, for exploring the effect of the different levels of knowledge on the speaker recognition system: knowledge of the method only, knowledge on the feature extraction/normalization process, knowledge on the world model and knowledge on the targeted speaker training examples.

7. REFERENCES

[1] R.H. Bolt, F.S. Cooper, D.M. Green, S.L. Hamlet, J.G. McKnight, J.M. Pickett, O. Tosi, B.D. Underwood, D.L. Hogan, "On the Theory and Practice of Voice Identification", *National Research Council, National Academy of Sciences, Washington, D.C.*, 1979.

[2] O. Tosi, "Voice Identification: Theory and Legal Applications",*University Park Press: Baltimore, Maryland*, 1979.

[3] R.H. Bolt, F.S. Cooper, E.E.Jr. David, P.B. Denes, J.M. Pickett, K.N. Stevens, "Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes",*Journal of the Acoustical Society of America*, 47, 2 (2), 597-612, 1970.

[4] J.F. Nolan, "The Phonetic Bases of Speaker Recognition",*Cambridge University Press: Cambridge*, 1983.

[5] L.J. Boë, "Forensic voice identification in France",*Speech Communication, Elsevier*, Volume 31, Issues 2-3, June 2000, pp. 205-224 (http://dx.doi.org/10.1016/S0167-6393(99)00079-5).

[6] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J.P. Campbell, D.A. Reynolds, I. Magrin-Chagnolleau, "Person Authentication by Voice: A Need for Caution", *Proceeding of Eurospeech 2003*, 2003

[7] C. Champod, D. Meuwly, "The inference of identity in forensic speaker recognition",*Speech Communication*, Vol. 31, 2-3, pp 193-203, 2000

[8] J. González-Rodríguez, J. Ortega, and J.J. Lucena, "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", *Proc. Odyssey2001 Speaker Recognition Workshop*, pp. 135-138, Crete (Greece), 2001

[9] P. Rose, T. Osanai, Y. Kinoshita, "Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold", *Speech Language and the Law*, 2003; 10/2: 179-202.

[10] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, D. A. Reynolds, "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, 2004, Vol.4, pp.430-451

[11] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing (DSP), a review journal Special issue on NIST 1999 speaker recognition workshop*, vol. 10(1-3), pp 19-41, 2000.

[12] NIST Speaker Recognition Evaluation campaigns web site, http://www.nist.gov/speech/tests/spk/index.htm

[13] LIA_SpkDet system web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

[14] J.-F. Bonastre, F. Wils, S. Meignier, "ALIZE, a free toolkit for speaker recognition", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[15] P. Perrot, G. Aversano, R. Blouet, M. Charbit, G. Chollet, "Voice Forgery Using ALISP: Indexation in a Client Memory", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[16] ALIZE project web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/