Artificial impostor voice transformation effects on false acceptance rates

Jean-François Bonastre, Driss Matrouf, Corinne Fredouille

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

10.21437/Interspeech.2007-173

Abstract

This paper investigates the effect of a transfer function-based 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 follows previous works where we demonstrate that, if someone has a knowledge on the speaker recognition method used, it is possible to impersonate a given speaker, in the view of this speaker recognition method. In this paper we extend the previous work by relaxing the needed knowledge on the targeted speaker recognition system. The results show that the voice transformation allows a drastic increase of the false acceptance rate, without damaging the natural perception of the voice, and without needing a large knowledge on the targeted speaker recognition system.

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. Nevertheless, several uncontrolled variability factors remain a main drawback and have a drastic influence on system performance, difficult to predict. The mismatch between recording sessions (including environment, noise, microphone and transmission channel) is the most highlighted of these factors in the literature, maybe not because it is the most influent but certainly because it is one of the most frequent.

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.

In [8][9], we proposed a simple artificial impostor voice transformation process. The main objective was to verify if it is possible to cheat a system, when knowledge on this system is available. It seems reasonable to think that an organization which wants to attribute a speech segment to a given – well known – speaker has knowledge of the kind of 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 demonstrated in these works 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. The objective of these works was close to the voice-forgery approach proposed in [14] even if our goal was only to obtain positive system decisions for impostors (without loosing the natural perception of the voice) and not to synthesize a voice excerpt close to the target speaker for a human point of view.

Moreover, in these preliminary works, a large knowledge on the speaker recognition system was assumed, including the feature extraction process, the modeling process and the world model. The purpose of this paper is not to propose some novelty concerning the transformation process (it reuses the same transformation function) but both to withdraw a large part of the previously presented constraints and to validate the effectiveness of the voice transformation on a validation, corpora.

This paper is organized as follows. Firstly, the voice transformation method is briefly presented in section 2. A set of preliminary experiments are presented in section 3 using NIST 2005 SRE framework. The new validation of the approach is proposed in section 4, using NIST 2006 SRE framework. Some conclusions and future work are proposed in section 5.

2. Speech transformation

This section presents a brief description of the voice transformation method presented in details in [8]. The aim of this method is to transform speech signal belonging to a speaker in order to increase its likelihood given a targeted speaker. For the forensic-oriented part of the paper, listening to the resulting signal, the effects of the transformation must appear as natural as possible.

The principle retained for the voice transformation is to analyze the input voice signal following a source transfer function model and to replace the transfer function of the test data with the transfer function of the targeted speaker. It is done by analyzing the impostor signal frame by frame: on each impostor frame, a target transfer function is estimated using a GMM model of the targeted speaker – it corresponds in fact to a weighted arithmetic mean of the components of this model –, and the transformed signal frame is synthesized using the target transfer function. The final output signal is obtained thanks to a classical overlap-add technique.

The proposed approach synthesizes a new signal as close as possible to the targeted speaker voice, in the behavior of automatic speaker recognition. In order to achieve this goal, the estimation of the transfer functions used during the transformation process should take into account the state-of-the-art in speaker recognition. Particularly, the feature extraction process, usually based on a cepstral parameterization followed by normalization/selection steps (mean removal and variance normalization,

speech/non-speech frame selection) has demonstrated its importance and should be taken into account. This rises the main difficulty related to this approach: the transformation parameters should be estimated using speaker recognition oriented feature extraction process while it is mandatory to keep the ability to resynthesize a signal using the transformed parameters. This is usually not possible when feature normalization is applied.
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 a typical speaker recognition feature space (cepstral plus feature normalization). This model is denoted "automatic speaker recognition" (asr) model or "master model" and is used to estimate the *a posteriori* probabilities of the GMM Gaussian components given each frame. The second one, denoted here "filtering" model (fil), is used to estimate the optimal time-varying filter parameters using the probabilities given by the master model.
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:

$$Y(f) = H_y(f)S_y(f) \quad (1)$$
$$X(f) = H_x(f)S_x(f) \quad (2)$$

where Y and X are the spectral representations of y and x. H_y and H_x are the transfer functions corresponding to both y and x; S_x and S_y are the Fourier transforms of the source signals corresponding to x and y. We call H_x the target transfer function and H_y the source transfer function. If the phase of the original signal is not modified, the filter to apply to the signal y becomes:

$$H_{yx}(f) = \frac{|H_x(f)|}{|H_y(f)|} \quad (3)$$

In this paper the transfer functions are estimated as in [8] and as follows:

$$H_x(f) = \frac{G_x}{A_x(f)} \quad (4)$$
$$H_y(f) = \frac{G_y}{A_y(f)} \quad (5)$$

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 target transfer function is derived from the linear combination of all the filtering GMM means weighted by their *a posteriori* probabilities (estimated using the *master* model).
Synthesis of the transformed signal is done frame by frame using the standard overlap-add technique with Hamming windows, where the resulting signal is obtained by adding the resulting window-based signals.
Figure 1 presents a block diagram of impostor frame transformation.

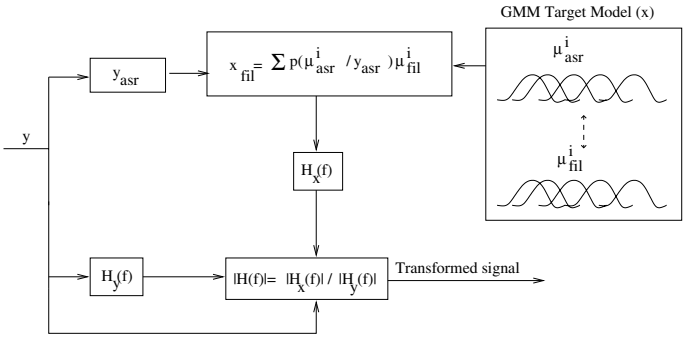

Figure 1: Transform block diagram for one frame: The target transfer function H_x is estimated by using 2 parallel GMMs, with one-to-one component tying. The first one allows the *a posteriori* probability estimation; and the second one is used for filtering.

3. Experiment using development database

This section presents some preliminary experiments using the proposed voice transformation system (already presented in [8]).
Experiments are conducted in the context of the NIST SRE 2005 evaluation campaign [11]. Two corpora are derived from the NIST05 official ones:

- the corpus $Eva05$, composed of 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 (NIST SRE 2002-2004 data). 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. Three different experiments are proposed in this section:

- baseline: no voice transformation is applied
- experiment 1: 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. The transformation uses directly the corresponding speaker training file. Experiment 1 represents the most favorable case to cheat the system.
- experiment 2: same process as for experiment 1 except that a different file of the targeted speaker is randomly selected for the voice transformation [1]. Experiment 2 represents a more realist scenario to cheat a speaker recognition system, where an example of the targeted speaker voice is available but different from the targeted speaker training record.

In all the experiments implying the voice transformation, identical world models are used for both the speaker recognition system and the voice transformation process.

The LIA_SpkDet system [12] developed at the LIA laboratory is used as baseline in this paper. Built from the ALIZE platform [15][13], the LIA_SpkDet system is based on classical UBM-GMM models and T-Norm approach for likelihood

[1] Except for a very small number of speakers for these only the training file is available in the corpus. We verified that the results are similar when the experiment is done with or without these speakers and the corresponding trials.

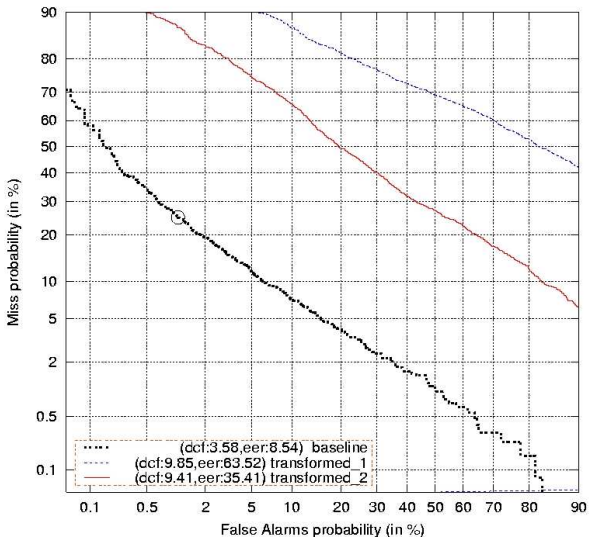

Figure 2: DET for the baseline system, the voice-transformed impostors using the speaker training file (exp. 1) and the voice-transformed impostors using a different file (exp. 2).

score normalization [7]. 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.

In this section, the world master-GMM is gathered from the baseline system. 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 to estimate 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.

Figure 2 presents the det curves of the baseline system (no voice transformation), the experiment 1 – where the transformation process uses the targeted speaker training file – and the experiment 2 – where a different file from the training one is used for the voice transformation. All the results relate to tnormed scores. The performance of the system drastically decreases when a voice transformation is used for the impostor files, for both experiments. Obviously, using a complete knowledge of the targeted speaker (experiment 1: using the training file of the targeted speaker for the voice transformation) gives higher impostor scores than using a limited knowledge (Experiment 2: using the knowledge of the targeted speaker ID but not the corresponding training file).

In order to highlight the differences in the impostor acceptance rate, we propose in table 1 the miss probability and false alarm rates for a threshold corresponding to the optimal dcf reached by the baseline system.
Listening to several examples of transformed signals, we did not notice any distortion and the signal remained natural.

	False A. (%)	Miss P. (%)
Baseline	0.88	27.45
Exp 1	96.55	27.45
Exp 2	49.72	27.45

Table 1: False Alarm and Miss Probability using a priori threshold (fixed using baseline), for the baseline experiment (no transformation), the experiment 1 (impostor trials transformed using the targeted speaker training file), and the experiment 2 (impostor trial transformed using a different file of the targeted speaker).

4. Validation experiments

In the previous section, we reported some development results. Even if the performance of the impostor voice transformation technique was clearly demonstrated, this transformation used a large knowledge on the targeted speaker recognition system. Particularly, the feature extraction process and the world model were identical for the voice transformation and for the speaker recognition system. Moreover, the system was developed on the NIST 2005 database and needed to be validate on a new database.
This section proposes a new set of experiments based upon NIST 2006 database, restricted to the male subset, 1conv-1conv condition (like done in the previous section)[2]. The voice transformation system remains unchanged, it uses the 32 coefficient-based feature extraction process and the master world model (estimated on DEV05) defined in the previous section. The speaker recognition system is also the LIA_SpkDet system but with a new setup, defined for NIST06 SRE campaign:

- the UBM is now trained on a part of the fisher corpus;
- a cohort of 160 -target- male speakers of NIST SRE 2004 database has been used for Tnorm;
- for the front-end processing, the signal is now characterized by 50 coefficients including 19 linear frequency cepstral coefficients (LFCC) issued from a filter-bank analysis, their first derivative coefficients, 11 of their second derivative coefficients and the delta energy;
- the energy-based frame removal is computed *after* the voice transformation process, on each transformed file;
- the experiments are conducted on NIST SRE 2006 database. It provides 22131 tests, including 1570 target tests. All the impostor trials are transformed using the proposed transformation function.

Figure 3 presents the DET curve of the 2006 baseline system without and with applying the impostor voice transformation system, following the experiment 2[3] defined in the previous section (the train segments are not used for the transformation function).
Even if the parameterisation step, the UBM training dataset and the Tnorm cohorts are not shared by the ASR system and the voice transformation step, the results remain very clear. The voice transformation is able to increase very significantly the false acceptation rate. The EER increases drastically from

[2]Even if the development database and the validation database are different in terms of speakers and recordings, they are both issued from the same NIST recording protocol (same kind of data, same scenario, same recording engine)

[3]Like in the previous section, for some speakers only the target file is present in the database. We verified that the influence of this problem is negligible.

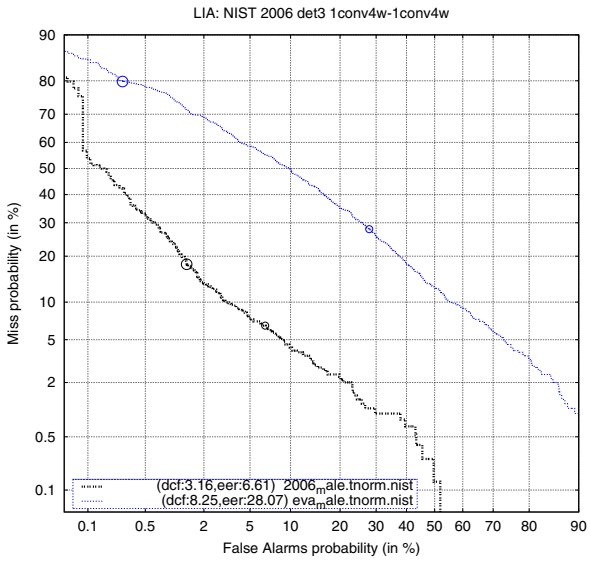

Figure 3: DET of the validation experiment (NIST SRE 2006), using the 06 baseline and the voice-transformed impostors (without using the speaker training file (exp. 2))

6,61% until about 28%.
It is difficult to validate the natural perception of the transformed files by an automatic process, only a perceptual evaluation based on an human jury could assess scientifically this point, which is quite expensive in terms of man power. Moreover, by listening ourself several files (randomly picked), we are very confident on the results of a such experiment. The files remain very natural, except a small increase in term of noise, certainly easy to withdraw by a smoothing of the transformation function parameters.

5. Conclusion and Future Work

In our first works concerning the impostor voice transformation, we investigated the effect of artificially modified speech on a speaker recognition system. We demonstrated that it is quite easy to "cheat" an automatic speaker recognition system, by transforming the voice of someone to be close to the voice of a targeted speaker, in the view of the system. Nevertheless, in these previous works, a large knowledge of the speaker recognition system was assumed, including the feature extraction process, the modeling method and the world model.
In this paper, we proposed a new set of experiments. A new UBM training set and the feature extraction process were used for the speaker recognition system while the voice transformation uses the previous setup. The very good results obtained when using this configuration demonstrated that a strong knowledge on the speaker recognition system is not mandatory to cheat it by transforming the impostor voices.
Even if we argue that only an accurate knowledge on the state of the art is necessary, we have to demonstrate this statement by using different speaker recognition technologies. This point will be addressed in future works. Firstly, some experiments using the latest technologies like Latent Factor Analysis and NAP are currently running. Secondly, the transformed data will be publicly available thanks to NIST as well as the voice transformation system.The latest will certainly authorize some external researchers to test their own speaker recognition technology in this framework.
Finally, the robustness of a speaker recognition system against such voice transformation technology is clearly a key point for the future and we will focus our research efforts on it. Particularly, we hope that the dynamic of speech is more difficult to transform and we want to explore the benefit of the AES approach proposed in [16].

6. 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

[7] 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

[8] D. Matrouf, J.-F. Bonastre and C. Fredouille, "Effect of voice transformation on impostor acceptance", *Proc. of ICASSP 2006*, Toulouse, France, 2006

[9] J.-F. Bonastre, D. Matrouf and C. Fredouille, "Transfer function-based voice transformation for speaker recognition", *Proc. of IEEE Odyssey Workshop*, Puerto-Rico, USA, 2006

[10] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing (DSP)*, vol. 10(1-3), pp 19-41, 2000.

[11] http://www.nist.gov/speech/tests/spk/index.htm

[12] http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

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

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

[15] http://www.lia.univ-avignon.fr/heberges/ALIZE/

[16] Nicolas Scheffer and J.-F. Bonastre, "A multiclass framework for Speaker Verification within an Acoustic Event Sequence system", *Proceedings of ICSLP*, Pittsburgh (USA), 2006, pp 501-504