Vocal Forgery in Forensic Sciences

Patrick Perrot[1,2], Mathieu Morel[2], Joseph Razik[2], and Gérard Chollet[2]

[1] Institut de Recherche Criminelle de la Gendarmerie Nationale,
Rosny sous Bois 93110 France
patrick.perrot@gendarmerie.defense.gouv.fr
[2] CNRS-LTCI-Telecom-ParisTech, 37-39 rue Darreau, 75014 Paris France
{chollet,razik,morel}@telecom-paristech.fr

Abstract. This article describes techniques of vocal forgery able to affect automatic speaker recognition system in a forensic context. Vocal forgery covers two main aspects: voice transformation and voice conversion. Concerning voice transformation, this article proposes an automatic analysis of four specific disguised voices in order to detect the forgery and, for voice conversion, different ways to automatically imitate a target voice. Vocal forgery appears as a real and relevant question for forensic expertise. In most cases, criminals who make a terrorist claim or a miscellaneous call, disguise their voices to hide their identity or to take the identity of another person. Disguise is considered in this paper as a deliberate action of the speaker who wants to conceal or falsify his identity. Different techniques exist to transform one's own voice. Some are sophisticated as software manipulation, some others are simpler as using an handkerchief over the mouth. In voice transformation, the presented work is dedicated to the study of disguise used in the most common cases. In voice conversion, different techniques will be presented, compared, and applied on an original example of the French President voice.

Keywords: disguised voices, voice conversion, SVM classifier, identification.

One of the most important issue in the area of forensic speaker recognition is the vocal forgery. How is it possible to detect or to compensate it? What is the influence of disguise on automatic speaker recognition systems? This article tries to give an answer to these questions and raises the question of automatic imitation.

1 Vocal Forgery and Automatic Speaker Recognition in Forensic Sciences

To falsify one's identity or hide it, different possibilities are offered to criminals. They can choose between a transformation of their voice by using simple means like an handkerchief in front of the mouth or speaking with a higher voice, or by using a sophisticated method to imitate the voice of another person in order to compromise him/her.

M. Sorell (Ed.): e-Forensics 2009, LNICST 8, pp. 179–185, 2009.

1.1 Voice Transformation as Forgery

The possibilities for criminals to transform their voice are very numerous. The literature describes different experiments conducted to detect a specific disguise according to a phonetic approach [9][12]. Software manipulation is a very relevant mean to disturb significantly a forensic recognition. The speech is generally dramatically altered. Many parameters can be affected: fundamental frequency, formants, rhythm.

Another disguise which is especially efficient is the whispered voice. This kind of method eliminates the voiced part, the information about fundamental frequency but also alters the intensity. The main drawback for criminals, who use such a disguise, is the difficulty to deliver an audible and intelligible message. Some other techniques are presented in [15][13].

What is interesting to notice is that most general criminals have just used one form of disguise. Masthoff [12]demonstrates that listeners have many difficulties when more than one disguise is used. Impact of disguise in automatic speaker recognition system is presented in [10].

Our work deals with the most common disguises used according to answer of nearly 100 persons to a questionnaire and to the experience of the forensic research institute of the French Gendarmerie. The chosen disguises are, a hand over the mouth,pinched nostrils, high pitch voice, low pitch voice. The impact of these different disguises on speaker recognition performance is presented in [14]. Disguise is a real problem for forensic expert. A preliminary step which consists in detecting if the voice is disguised or not, could be a very useful tool, in order to avoid confusion. The performance of SVM classification is presented in section 2.1 under neutral and degraded conditions.

1.2 Voice Conversion as Forgery

Voice conversion, which consists in producing a sound pronounced by a source speaker to sound like a target speaker, appears as a good way to achieve forgery [11]. A simple way is an imitation of the target by a professional impersonator. This technique of conversion presents the main drawbacks to be difficult to reproduce and thus, is not a universal method. A good mean to compensate this question is to perform an automatic conversion. Different techniques are proposed in the literature. This section describes the main algorithms developed and presents a comparison of two methods. Let us consider a sequence of a spectral vectors pronounced by the source speaker:

$$X = [x1, x2, \ldots, xn] \; . \tag{1}$$

and a sequence corresponding to the pronunciation of the same utterance by the target speaker:

$$Y = [y1, y2, \ldots, yn] \; . \tag{2}$$

Voice conversion is based on the calculation of a conversion function F that minimizes the mean squared error:

$$\varepsilon_{mse} = E\left|\|Y - F(X)\|^2\right| \ . \tag{3}$$

where E is the expectation.

To calculate the conversion function, most representative and relevant techniques, developed in the past decades, are based on Gaussian Mixture Models and related techniques [3][4][2]. Automatic voice conversion is divided into two main steps: training and conversion. First, two speech recordings of the same utterance (one for the source and one for the target) are time aligned by DTW (Dynamic Time Warping), then both signals are analyzed asynchronously by HNM (Harmonic plus Noise Model) as proposed in [3]. MFCC (Mel Frequency Cepstral Coefficient) are extracted from the HNM parameters. The mapping between these two aligned sets of MFCC features (source and target) is based on Gaussian Mixture Models (GMM). The joint density function $P(X, Y)$ as proposed in [4] is estimated with a GMM, i.e. with a probability distribution given by:

$$P(z) = \sum_{q=1}^{Q} \alpha_q \mathcal{N}_q(z, \mu_q, \Sigma_q) \ . \tag{4}$$

where $z = [x, y]$ is the joint spectral vector, Q the number of gaussians and α_q, μ_q, and $\sum_q$ respectively the weight, the mean, and the covariance matrix of the q^{th} gaussian component. The iterative algorithm EM (Expectation Maximization) is used to estimate the parameters of the GMM on all the joint spectral vectors from the training set. Following the training step the transformation step consists in applying this function to a speech (different from the training data set) of the source. HNM analysis is performed on the speech source and MFCC are extracted from the voiced frames. Then, for each spectral vector x of the source, the corresponding vector of the target y is predicted by finding the expected value of y given x in the joint probability.

Another voice conversion technique based on a client dictionary is possible and presented in [14]. This method consists in using a dictionary of a target voice and replacing speech segments of the source voice by their counterpart of the target voice. It is based on the ALISP (Automatic Language Independent Speech Processing) technique. The principle of ALISP 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.

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 [1] 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. The result of the ALISP training is an inventory of client speech segments, divided into 64 classes according to a codebook of 64 symbols. All the speech segments contained in our inventory are represented by their Harmonic plus Noise Model (HNM) parameters. This will allow a smooth concatenative synthesis of new sentences using the stored segments.

The second part of our processing consists in encoding the impostor's voice using the ALISP codebook, and then in performing decoding using synthetic units taken from the segment inventory obtained from client's voice. The different results and the impact of such a conversion on an automatic speaker recognition system is presented in [14]. What is observed is a significant decrease of the performance, that is to say that the target is recognized instead of the source speaker.

2 Experiments and Results

2.1 Identification of Disguised Voices Based on a SVM Classifier

The technique used in this part is a classifier based on VQ (vector quantization) and SVM (support vector machine) dedicated to the identification of disguised voices. SVM is a very efficient discriminant tool in pattern recognition. This method proposes a decision function that only uses a subset of a training database called support vectors. Let us consider a training set

$$A = (x_1, y_1), ...(x_m, y_m) \ . \tag{5}$$

composed by m couples (attribute vectors and labels) with $x_i \in \Re^n$ and $y_i \in \{-1, +1\}$. SVM algorithm consists in projecting x_i vectors in a new space T from a non linear function:

$$\phi : \Re^n \rightarrow T \ . \tag{6}$$

The second point is to find out the optimal boundary or hyperplane (w, b) of the two class in T. The y class of a new sample x is defined by:

$$y = sign(w.\phi(x) + b) \ . \tag{7}$$

The optimal hyperplane is the one which maximizes the distance between itself and the closest label vectors. The principle to use a vector quantization prior to the application of a SVM classification is to increase the robustness of noise by building some centroids representative of the sample distribution. The experiment is realized on a training set of 40 people in each disguise and the test is performed on 20 speakers. The training set consists in a reading of the phonetic balanced text: the north wind and the sun, and the test corpus is composed of 10 sentences. Results are presented on DET curves (plotting false acceptance rate against false rejection rate). Fig 1 represents the identification of the four chosen disguises and the normal voice in neutral conditions that is to say without specific noise.

In order to be more realistic, the experiment is carried out by adding different noises on the test set: white and pink noise (Fig 2 and 3) and babble noise (Fig 4). The idea is to measure the impact of degraded conditions on the performance of the classifier.

Noisy conditions affect significantly the results of classification. That is unfortunately not very surprising because data from normal voice and smooth disguised voices are very linked. We notice that voices with a low pitch seem to be the most difficult to discriminate (low pitch voice and hand over the mouth).

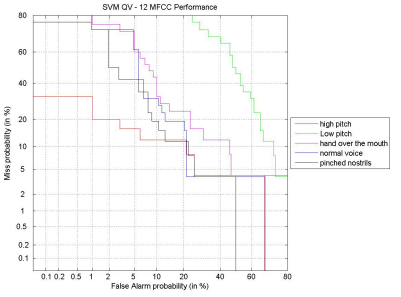

Fig. 1. Neutral conditions

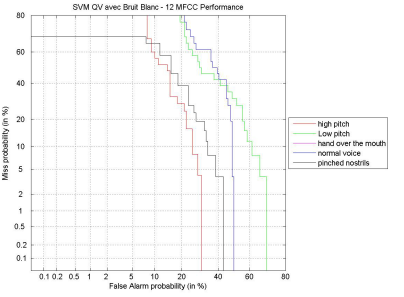

Fig. 2. Degraded conditions: white noise

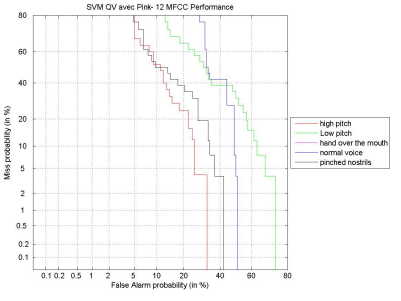

Fig. 3. Degraded conditions: pink noise

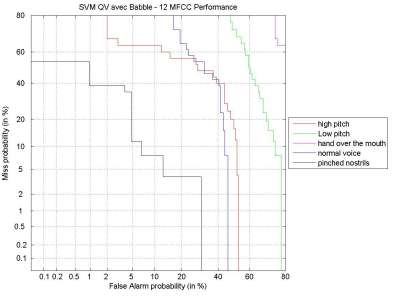

Fig. 4. Degraded conditions: babble noise

2.2 Comparison of Voice Conversion Technique as Forgery

Two different methods of voice conversion have been proposed in section 1.2. A comparison of both methods is performed to measure the level of the conversion quality. The experiment has been based on the voice of the current French President as target speaker in order to be close to a realistic scenario. A 40 minutes speech of the French president was collected and 70 sentences extracted from this discourse are pronounced by a male (the source speaker), for conversion based on GMM in the training task.

A preliminary step consists in aligning both speech segments by DTW and then calculating the conversion function as described in section 1.2. In the case of the conversion based on ALISP, 35 minutes of the discourse have been used to build the dictionary of the target voice. The test has been performed on 10 sentences, different from those of the training set.

Two different comparison methods have been done. The first one is a calculation of the spectral distortion measure between the converted and the target speech (Table 1). After the temporal alignment step of the both signals the distance between MFCC's features is:

$$d = \sum_{t=0}^{n} \sum_{k=1}^{20} (c_k^1(t) - c_k^2(t))^2 \ . \tag{8}$$

where $c_k^1(t)$ and $c_k^2(t)$ are the k^{th} MFCC's coefficient at time t of the converted and the target signal. This measure is normalized by the initial distortion between the source and the target speaker to lie between 0 and 1. 0 means that the converted speech is similar to the target speech and 1 means that the conversion has no effect.

Table 1. Comparison of conversion system

conversion system	GMM	ALISP
spectral distortion	0.77	0.78

The second technique to evaluate the level of the conversion is based on the listening of the result. This kind of technique is subjective and a significant issue is the quality of the speech after conversion. This one is not enough satisfying to allow to recognize significantly the target or the client. A listening of the conversion result will be proposed.

3 Conclusion

Different perspectives of vocal forgery have been presented in this paper. This question appears as a real issue for forensic examination where the risk of confusion between two speakers could have great consequences. Disguised voices considered as a transformation of the voice by simple means alter significantly the performance of automatic speaker recognition. A preliminary detection of disguise before an examination of speaker recognition could be very useful in order to avoid confusion. The presented classification provides interesting results, even if under degraded conditions, the discrimination between disguises is more difficult, especially in the case of low pitch voices. Voice conversion also appears to be an interesting way for impostors to take the voice of a specific person. A comparison of two conversion methods is presented and the results are efficient. This kind of forgery does alter the performance of speaker recognition. An example using the voice of the French president is presented and reveals the threat of such a technique in forensic or terrorism cases.

References

1. Bimbot, F., Chollet, G., Deleglise, P., Montacie, C.: Temporal decomposition and acoustic-phonetic decoding of speech. In: ICASSP, pp. 445–448 (1998)
2. Duxans, H., Bonafonte, A.: Estimation of GMM in voice conversion including unaligned data. In: EUROSPEECH, pp. 861–864 (2003)
3. Stylianou, Y., Cappe, O.: Statistical methods for voice quality transformation. In: EUROSPEECH, pp. 447–450 (1995)
4. Kain, A., Maccon, M.W.: Spectral voice conversion for text to speech synthesis. In: ICASSP, pp. 285–288 (1998)

5. Perrot, P., Aversano, G., Blouet, R., Charbit, M., Chollet, G.: Voice forgery using ALISP: indexation in a client memory. In: ICASSP, pp. 17–20 (2005)
6. Tsuge, S., Shishibori, M., Kita, K., Ren, F., Kuroiwa, S.: Study of Intra-Speakers Speech Variability Over Long and Short Time Periods for Speech Recognition. In: ICASSP (2006)
7. Ortega-Garcia, J., Cruz-Llanas, S., Gonzalez-Rodriguez, J.: Speech variability in automatic speaker recognition systems for forensic purposes In: IEEE 33rd Annual International Carnahan Conference (1999)
8. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C.: Automatic speech recognition and speech variability: A review. Speech Communication 49(10-11), 763–786 (2007)
9. Kunzel, H.J.: Effect of voice disguise on fundamental frequency. Forensic Linguistics, vol. 7 (2000)
10. Kunzel, H., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In: Odyssey (ed.), pp. 153–156 (2004)
11. Genoud, D., Chollet, G.: Voice transformations: Some tools for the imposture of speaker verification systems. Advances in Phonetics. Franz Steiner Verlag (1999)
12. Masthoff, H.: A report on voice disguise experiment. Forensic Linguistics, vol. 3 (1996)
13. Orchard, T., Yarmey, A.: The effect of whispers, voice sample duration, and voice distinctiveness on criminal Speaker Identification. Applied Cognitive Psychology vo 9(3), 249–260 (1995)
14. Perrot, P., Chollet, G.: The question of disguised voices. In: Acoustics 2008, Paris (2008)
15. Reich, A.R., Duke, J.E.: Effect of selective vocal disguise upon speaker identification by listening. Journal of Acoustical Society of America 66, 1023–1028 (1979)