coustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

the question of disguised voice

Patrick Perrot and Gerard Chollet

Telecom Paris Tech, 46 rue Barrault, 75013 Paris, France
perrot@tsi.enst.fr

Many applications including bank, multimedia, biometrics, need the verification of speaker identity. The current performance of speaker recognition can be considered as sufficient in many fields, but in forensic sciences, caution must be a priority due to the lack of robustness of the systems. Nevertheless the problem of identification is essential in forensic sciences. In most criminal cases, offenders try to disguise their voice before sending an anonymous or miscellaneous call. This is the reason why it is important to study the possibilities of voice disguise before trying to identify a speaker. The purpose of this paper is to present the applications of statistical algorithms in order to detect and identify four specific disguises. The choice of the disguises is based on the most common ones used by offenders.

1 Introduction

Due to the wide range of commercial and law enforcement applications, major breakthroughs and initiatives in the past twenty years have propelled biometrics and specifically voice recognition technology into the spotlight. Voices have been used recently to verify the identity of persons, including security systems and criminal identifications. Forensic speaker recognition used to be performed by phoneticians but nowadays increasing interest is placed on automatic statistical techniques. However, there is a large gap between commercial and forensic applications. Actually, nowadays cheap and versatile systems make it possible to easily and quickly identify a speaker but the level of performance and the robustness of the system are not estimated. The question of disguise is not a real problem in the case of commercial applications because the will of spoofing the system is not in the user mind in most cases.

On the contrary, voice applications in a forensic context suffer from the question of disguise, especially in the case of speaker recognition. Few systems take into account this problem. So, the study proposed in this paper presents the results of three statistical algorithms (K-nearest neighbours, Gaussian Mixture Model, and Support Vector Machine) and the use of different features. This work focuses on the classification of four disguises: hand over the mouth, low, high pitch and pinched nostrils voice.

After a brief state of the art on the question of voice disguise, the different algorithms will be described before proposing the result of the classification on specific features MFCC (Mel Frequency Cepstral Coefficient), MFCC and derivatives).

2 State of the art

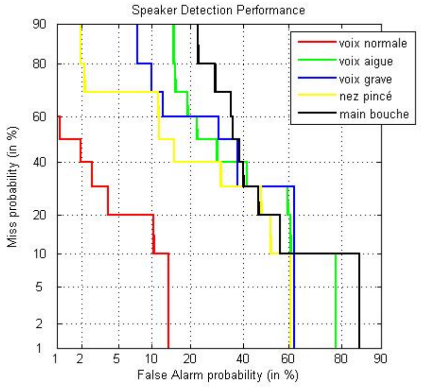
Voice disguise is a deliberate action of a speaker who wants to falsify or to conceal his/her identity. The problem of voice alteration caused by channel distortion is not presented in this work. Lots of possibilities are offered to a speaker to change his own voice and to forger a human ear or an automatic system. He could transform his voice by electronic scrambling or more simply by exploiting the intra-speaker variability: modification of his own pitch, modification of the position of the articulators like lips or tongue which affect the formant frequencies. So, the question of voice disguise includes the voice transformation, the voice conversion and the alteration of the voice by mechanic means. In this study, we limited our applications to non electronic voice transformations, that is to say, a modification based on simple means corresponding to that used in the cases of offences.

Research on voice disguise started in the 1970s with phoneticians like Künzel, Koester, and it is really over the past 10 years that researchers have tried to develop automatic systems to detect the disguise. This question of voice disguise in forensic sciences was not very developed in the literature, certainly because of the difficulties to distinguish a normal voice from a disguised voice in criminal applications. Nevertheless, the increase of voice use in multimedia applications and the current performance of speaker recognition systems offer a new interest for voice disguise. Hollien revealed that in the case of several diguised voices (except whisper and foreign accent) the identification performance of a machine was a little better than chance [8]. In [9], Masthoff establishes a report on the way used by speakers to disguise their voice. His results come from an experiment on 20 German speakers. He notices that the preferred forms of disguise appeared to involve changes in phonation and either one or two techniques. It also results that the disguise depends on the form of the experiment. Künzel proposes in [2] a very complete study on the link between fundamental frequency and voice disguise. He reveals that it is possible to link the F0 in an undisguised mode of a speaker with his disguised F0. Torstensson and al. [11] provide information on the imitation of a foreign accent as disguise. This means has a serious impact on the individual's ability to recognize a speaker. In [5] a detailed way to analyze and to identify voice disguise in an automatic way are described.

3 Speaker identification and disguised voice

The first step of our work is to establish the impact of disguised voice on the performance of automatic speaker identification [2][3][4]. Four specific disguises have been chosen according to their use in criminal cases: hand over the mouth, pinched nostril, high pitch and low pitch. The principle of the automatic identification approach is divided in two parts. Previously for each speaker, 12 MFCC (Mel Frequency cepstral coefficients) and their derivatives have been extracted from a 20ms frame (10ms overlapped) in each speech segment after a silence removal step. The first part consists in a training session, which aim is to build different models for each speaker. This session consists in modeling the speech features of each speaker by the use of GMM. GMMs are widely used in statistical models in many pattern recognition applications. The principle is to approximate any probability density function from a sufficient number of components. The second part is the test which evaluates a distance between the query voice and the different models. The chosen distance in our system is a likelihood ratio and the maximum of this value determines the good speaker.

In order to measure the influence of disguise on our automatic system, different speech segments in disguised voices have been used. The impact on the performance is represented by figure 1. What we notice is a very significant degradation of the performance and this result can be compared to the conclusion of Hollien (“a little bit better than chance”).



.Fig.1 Disguised voice degradation

So, this DET curve reveals the need to be able to decide if a voice is disguised or not before planning an automatic speaker identification.

4 Experiment and Results

Three different experiments based on three different sets of features and three different classification methods have been carried out in order to detect disguised voices.

4.1 Corpus description

Two kinds of speech text have been chosen. The first one A is used for training. It is composed by 30 speaker audio files and three different corpus are built from the phonetic balanced text: “the north wind and the sun”(in French):

A1: 5mn of different speakers in four different disguises. A1 is a general model for disguise

A2: 5mn of normal voices from different speakers

A3: 5mn for each kind of disguise from different speakers.

The second one B is composed by 25 speakers and is used for test. This corpus is based on 10 phonetic French balanced sentences and has a duration between 15 and 20 seconds for each disguise (included normal voice)

The recordings are direct and the test dataset did not participate in the training process.

4.2 Feature extraction

Different sets of feature have been extracted from speech in order to evaluate the relevance of these specific features. 12 MFCC, and 12 MFCC + 12 derivatives.
A first approach is dedicated to 12 MFCC. MFCC, well known as the most common features in the case of speaker as well as speech recognition, are used. These coefficient vectors are computed on a 20 ms window with 10 ms shift. These features are derived from the outputs of a bank filter placed in a mel frequency scale. The filters are typically in triangular shapes, and are operated in the frequency domain. A second approach includes the derivatives of the MFCC in order to take into account the dynamic of the speech. So, a 24-dimension vector is computed.

4.3 Applications on disguised voice detection

In order to evaluate the best way to detect automatically disguised voice, three different classifications have been used on the previously described features.

- k-nearest-neighbors
- GMM (Gaussian Mixture Model)
- VQ (vector quantization) and SVM (Support Vector Machine)

The *k*-Nearest Neighbors (*k*-NN) classification rule is a technique for non-parametric supervised pattern classification. Given the training knowledge of *N* prototype patterns (vectors of dimension *D*) and their correct classification into several classes, it assigns an unclassified pattern to the class that is most heavily represented among its *k* nearest neighbors in the pattern space. The first comparative analysis focuses on 12 MFCC. In the experiment, after different tests for the k value, 20-nearest neighbors have been chosen .

Voice Type	Normal	Disguised
Normal	**62%**	38%
Disguised	22%	**78%**

Table 1 K-nearest-neighbors disguised voice detection

This method is efficient to detect a disguise but the risk to confuse a normal voice with a disguised voice is too important. In addition, a significant drawback of this algorithm is its very important time computing.

Another interesting and well known method in speech or speaker recognition is the use of GMM [6]. The principle is to build a GMM for disguised voices and another one for normal voices. A GMM is basically composed of a superposition of K Gaussian densities. Each density k is weighted with a mixture coefficient $c_{k.}$

$$p(x/m)=\sum_{k=1}^{K} c_{km} N(x,\mu_{km},\Sigma_{km})$$

The mixture coefficient obeys for each model m=1…M the probabilistic constraint: $\sum_{k=1}^{K} c_{km} = 1$

During the recognition phase the scores log(p(x/m)) are accumulated for the sequence X = $\{x_1, x_2, \ldots . x_p\}$

$$S(X/m) = \sum_{j=1}^{P} \log(p(x_j / m))$$

and the model is chosen according to the highest likelihood ratio score.

$$m = \arg\max_m S(x/m)$$

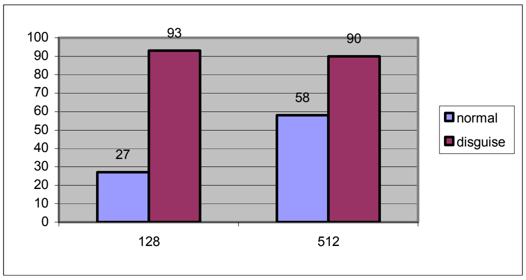
The result obtained are proposed in Table n°2:

Type of voice	Normal	Disguised
Normal	**15%**	85%
Disguised	6%	**94%**

Table 2 GMM (1024) disguised voice detection

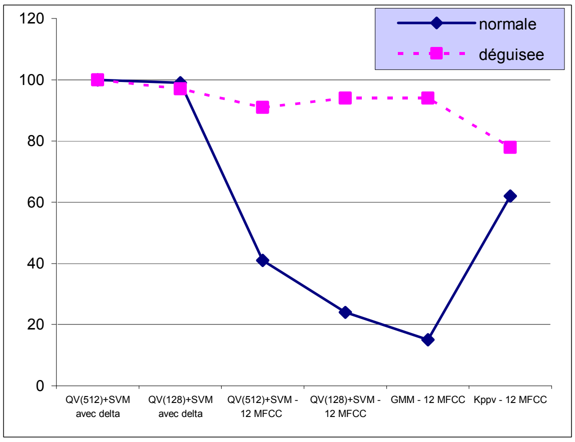
By applying a GMM classification the level of recognition of disguised voices is very high but the risk of confusing a normal with a disguised voice is also very high.

The last classification method used is based on a vector quantization followed by the application of SVM (Support Vector Machine) discrimination. The aim is to build a fast and efficient SVM classifier of data. The Vector Quantization (VQ) is used to simplify the training set. The principle is to represent the vector of each class by specific representatives. SVM is a binary classification method based on a supervised training. Specific kernels are used to optimize the data discrimination. The idea is to find a classifier able to discriminate the data and to optimize the interclass distance. The results proposed in figure 2 are based on 128 and 512 centroïds.



,Fig.2 VQ+SVM classification on MFCC + derivatives

These results are very positive even if it will be necessary to increase the test dataset in order to get a more significant number. What is interesting is that these results are confirmed by using 128 centroïds instead of 512 because 96% of normal voices are recognized as normal and 97% of disguised voices are recognized as disguised. So this set of features composed by 12 MFCC and their first derivatives and this classification technique appear to be thoroughly adapted to the disguised voice detection. The figure n°3 summarizes the different results of classification:



.Fig.3 Detection of disguised voices

4.4 Applications on disguised voice identification

The aim of identification is to be able to say which kind of disguise is used among the four studied disguises. Two different supervised classifiers have been analysed: GMM and VQ+SVM.

4.4.1. Identification based on a GMM classifier

The idea of this method is to build a specific model for each kind of disguise based on GMM. Each test speech segment is compared to the different models and the decision is taken according to the maximum likelihood ratio. This is the same principle as in 4.3. The advantage of this method is to be able to measure the level of identification (position n°1, n°2 and so on), according to the number of disguises. In our experiment four disguises have been analyzed based on 12 MFCC. The figure n°4 reveals the level of identification for each kind of disguise on 25 tests.

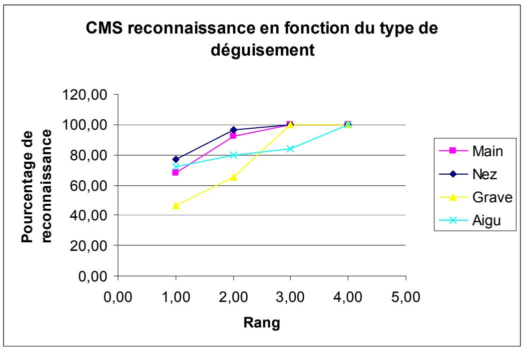


Fig 4: *Cumulative matching score in %*

4.4.2 Identification based on a SVM classifier

The principle of this classification is to be able to discriminate one disguise against all. For instance, to measure the identification level of high pitch voice, a similarity distance is calculated between test high pitch voice features against a model of high pitch voice versus a model of all disguise voices (except high pitch voice).

The classifier is based on VQ and SVM [7]. The identification process has been carried out from 512 centroids for the quantization vector step. The Figure n°5 represents a DET curve [10] on 25 clients and 100 impostors. DET curve is a good way to measure the performance of a system in term of false acceptance rate and false rejection rate.

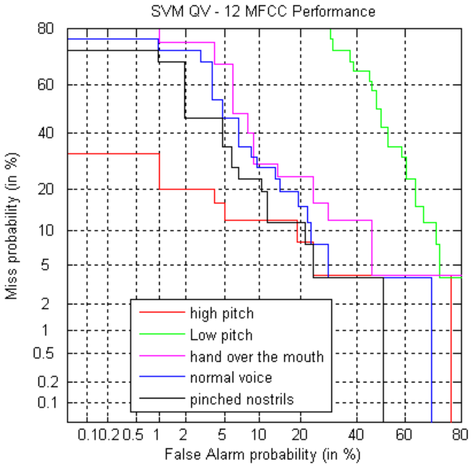


Fig 5: SVM classifer: DET curve on normal conditions

A same process of evaluation has been realized after adding a babble noise in order to be closer to forensic conditions. The following figure presents the results:

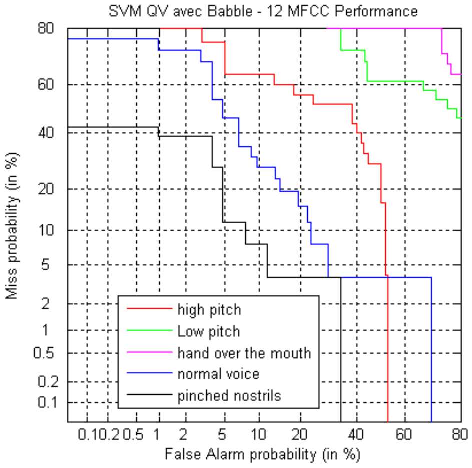


Fig 6: SVM classifer: DET curve on babble conditions

Identification of some specific disguises (high and low pitch voice, and hand over the mouth) is largely degraded by adding a babble noise

5 Conclusion

In forensic science cases, experts are more and more involved on speaker recognition question. One of the main problems is to be able to detect and to identify a disguise in order to avoid false automatic speaker identification. Actually, offenders try to forger their identity by using disguise techniques to avoid being recognized. Disguise has a important impact on the performance of an automatic recognition system. We present a series of experiments and results based on different features and different classification algorithms. The idea is to find a solution to detect and if possible to identify what kind of disguise has been used. The experiment results show that MFCC + their derivatives and QV+SVM classification provide interesting results in a case of detection, that is to say in a case where the question is to be able to say if a normal voice is normal or if a disguised voice is disguised. On the question of identification the results are unbalanced. The SVM classifier has a real problem with the hand over the mouth detection contrary to the GMM classifier that presents a correct level of identification. In the case of the other disguises, the QV+SVM classifier presents some good results. What is planned for future is to measure the robustness of this kind of detection and identification on other noise environments and the influence of the number of centroids.

References

[1] L.J. Boë – Ben Laden et le mythe de l'empreinte vocale In *revue du vivant n°1 –*

[2] H. Künzel, J. Gonzalez-Rodriguez, J. Ortega-Garcia. « Effect of voice disguise on the performance of a forensic automatic speaker recognition system", *Proceedings Speaker Odyssey 2004*

[3] Mats Blomberg, Daniel Elenius, Elisabeth Zetterholm Speaker verification scores and acoustic analysis of a professional impersonator *Proccedings FONETIK 2004*

[4] S.Kajarekar, H. Bratt, E. Shriber, and R. Leaon. A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition", *in Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop*

[5] P. Perrot, G. Aversano, G. Chollet. Voice disguise and automatic detection: review and perspective – *Progress in Non linear Speech Processing, Stylianou Y. Et al (eds) LNCS4391, Springler 2007*

[6] D. Reynolds. Speaker Identification and Verification using Gaussian Mixture Models – *Speech Communication, vol.45, pp.139-152, 2005*

[7] Vladimir Vapnik. The Nature of Statistical Learning Theory . In Springer-*Verlag, 1995.*

[8] H. Hollien. Forensic Voice Identification, *ed. AP – 2001*

[9] H. Masthoff, A report on voice disguise experiment, *Forensic Linguistics – 160-167 - 1996*

[10] A. Martin, G. Doddington, T Karam, MO.rdowski, M. Przybocki, The DET curve in assessement of detection task performance, *Proc. Eurospeech 97*, Greece

[11] N. Torstensson , E.J. Eriksson, K.P.H. Sullivan, Mimicked accents - Do speakers have similars cognitive prototypes ? *Proc. of Australian International Conference on Speech Science and Technology 2004*