

## Detection and Recognition of voice disguise

Patrick Perrot<sup>1-2</sup>, Céline Preteux<sup>1-3</sup>, Sophie Vasseur<sup>1</sup>, Gérard Chollet<sup>2</sup>

<sup>1</sup> Institut de Recherche Criminelle de la Gendarmerie Nationale, Rosny sous bois, France

<sup>2</sup> CNRS- Ecole Nationale Supérieure des Télécommunications, Paris, France

<sup>3</sup> Ecole de l'Air, Salon de Provence, France

[{patrick.perrot|sophie.vasseur}@gendarmerie.defense.gouv.fr](mailto:{patrick.perrot|sophie.vasseur}@gendarmerie.defense.gouv.fr)

[{perrot|chollet}@enst.fr](mailto:{perrot|chollet}@enst.fr)

In the field of forensic sciences, finding out a solution to discriminate a normal voice from a disguised voice could provide interesting clues to investigators. Nevertheless, this task is not obvious because of the definition of a “normal voice”, of the natural variability and of the different disguise possibilities. Speaker recognition is unavoidable in the field of forensic identification techniques. Different methods exist, some based on phonetic approaches, some others based on automatic algorithms. Before analysing a voice it could be interesting to evaluate if the voice is disguised or not. This paper presents a discriminative study on four specific voice disguises. The aim is to establish if a recorded voice is disguised or not and what kind of disguise is used. The choice of disguise was based on the answers of 70 persons to the question: which disguise would you use to modify your voice?

The four main disguises are: a hand over the mouth, a high pitch, a low pitch, and a pinched nostrils voice. First, a formant analysis is presented on the four disguises and compared to a normal voice, and secondly an automatic classification is realized. The obtained results provide interesting clues of discrimination.

### Elaboration of the database

The database has been elaborated from a set of 20 people for the formant analysis and 30 people for the automatic approach. Different French vowels are pronounced by different individuals as well as ten phonetically balanced sentences and a phonetically balanced text “The North Wind and the Sun”. For the automatic approach a training corpus of 5 minutes for each kind of disguise is extracted from the text, and a 20 seconds (per speaker) test corpus is constituted from a set of different speakers who pronounced the sentences.

### Formant analysis

The first and the second formant are extracted for each kind of disguise by the Praat software. The vocalic triangle reveals interesting results in order to discriminate the different disguise as presented below for instance:

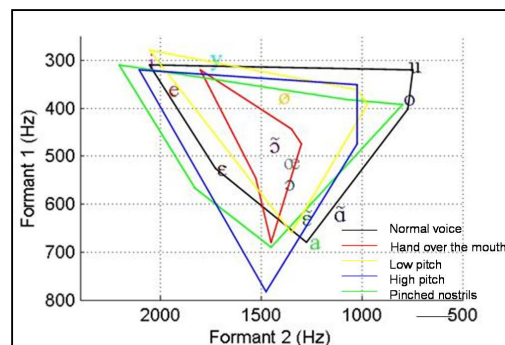


Figure 1: vocalic triangle

## Automatic approach

In order to discriminate the different disguise an automatic approach is studied and compared to a perceptual study. A perceptual test was conducted in order to evaluate the ability of the human perception to evaluate if a voice is disguised or not and if it is possible to determine the choice of the disguise used. The following table illustrated the results of this test on the 20 sentences.

**Table 1.** Recognition rate on Perceptual test

Recognition rate on sentence	
normal voice	1 <sup>st</sup> / 89 % (2 <sup>nd</sup> /pinched nostrils. 7,5%)
hand over the mouth	1 <sup>st</sup> / 82 % (2 <sup>nd</sup> /pinched nostrils. 9%)
pinched nostrils	1 <sup>st</sup> / 73 % (2 <sup>nd</sup> normal voice:19%)
high pitch voice	1 <sup>st</sup> / 87 % (2 <sup>nd</sup> .normal voice:7%)
low pitch voice	1 <sup>st</sup> / 68 % (2 <sup>nd</sup> .normal voice:17%)

A training corpus of 5 mn speech is used for each disguise. First, 12 MFCC (Mel Frequency Ceptral Coefficient) are extracted every 20 ms, after a silence removal step on each record. The classification phase is based on the knn (k-nearest neighbors) algorithm with 20 neighbors. This algorithm is applied to classify a set of 20 second speech from the test corpus.

The nearest neighbor algorithm is based on minimum distance from the query instance to the training samples to determine the 20-nearest neighbors. The Euclidian distance is used in our experiment. Then, we gather the 20-nearest neighbors and the majority of these nearest neighbors determines the prediction of the query instance.

The distinction between disguised and non disguised voice is illustrated by the following table:

**Table 2:** disguised/non disguised recognition

	Normal voice (train corpus)	Disguised voice (train corpus)
Normal voice (test corpus))	85%	15%
Disguised voice (test corpus)	29%	71%

And for each kind of disguise the results are:

**Table 3:** global recognition

Recognition rate on sentence	
normal voice	1er / 85% (2 <sup>nd</sup> /pinched nostrils: 8%)
hand over the mouth	2nd / 33 % (1 <sup>st</sup> normal: 55% )
pinched nostrils	1er / 92% (2. normal voice: 8%)
high pitch voice	1er / 77 % (2. pinched nostrils 15%)
low pitch voice	0% (1.normal voice: 61%)

## Conclusion and Perspectives

The question of voice disguise detection appears as fundamental in forensic applications. Different kinds of approaches provide significant results of discrimination. A complementary study based on formant and automatic analysis could be fused to increase the recognition rate. It could also be interesting to take into account others features like LPCC, delta LPCC and MFCC, F0 and to compare different techniques of classification. This is the next step of this study.

## **References**

- Boersma P., D. Weenink, "PRAAT: doing phonetics by computer. <http://www.praat.org>
- Küntzel H.J, "Effect of voice disguise on speaking fundamental", *Forensic Linguistics*, 2000
- Masthoff, H "A report on voice disguise experiment", *Forensic Linguistics*, 3(1):160-167. 1996
- Perrot, P. Aversano, G. Chollet G. "Voice disguise and automatic detection: review and perspectives" in: *Progress in Nonlinear Speech Processing*. Stylianou Y. et al (eds), LNCS 4391, Springer, 2007