Voice Disguise and Automatic Detection: Review and Perspectives

Conference Paper · January 2005

DOI: 10.1007/978-3-540-71505-4_7 · Source: DBLP

3 authors:

Patrick Perrot
GENDARMERIE NATIONALE
29 PUBLICATIONS **225** CITATIONS
SEE PROFILE

Guido Aversano
Nuance Communications
19 PUBLICATIONS **462** CITATIONS
SEE PROFILE

Gerard Chollet
French National Centre for Scientific Research
331 PUBLICATIONS **3,497** CITATIONS
SEE PROFILE

Voice disguise and automatic detection
review and perspectives

Patrick Perrot [12] , Guido Aversano [1], Gérard Chollet [1]

[1] CNRS-LTCI-Ecole Nationale Supérieure des Télécommunications (ENST)
75013 Paris, France
[2] Institut de Recherche Criminelle de la Gendarmerie Nationale (IRCGN)
93110, Rosny sous Bois, France
{perrot, aversano, chollet @tsi.enst.fr}

Abstract. This study focuses on the question of voice disguise and its detection. Voice disguise is considered as a deliberate action of the speaker who wants to falsify or to conceal his identity; the problem of voice alteration caused by channel distortion is not presented in this work. A large range of options are open to a speaker to change his voice and to trick a human ear or an automatic system. A voice can be transformed by electronic scrambling or more simply by exploiting intra-speaker variability: modification of pitch, modification of the position of the articulators as lips or tongue which affect the formant frequencies. The proposed work is divided in three parts: the first one is a classification of the different options available for changing one's voice, the second one presents a review of the different techniques in the literature and the third one describes the main indicators proposed in the literature to distinguish a disguised voice from the original voice, and proposes some perspectives based on disordered and emotional speech.

Keywords: classification disguise automatic detection

1 Introduction

In the field of disguised voices, different studies have been carried out on some specific features and some specific kinds of disguise. Voice disguise is the purposeful change of perceived age, gender, identity, personality of a person. It can be realized mechanically by using some particular means to disturb the speech production system, or electronically by changing the sound before it gets to the listener. Several applications are concerned by voice disguise: forensic science, entertainment, speech synthesis, speech coding. In the case of forensic science a study [26] reveals that a voice disguise is used at a level of 52% in the case where the offender uses his voice and supposes that he is recorded. The challenge is to find indicators to detect the type of disguise and if possible the original voice. Research into voice disguise could also provide some possibilities to personalize a voice in the development of modern synthesisers. The question of voice disguise includes the voice transformation, the voice conversion and the alteration of the voice by mechanic means. The principle of disguise consists in modifying the voice of one person to sound differently (mechanic alteration or transformation), or sound like another person (conversion). There are a number of different features to be mapped like voice quality, timing characteristics and pitch register. Voice modality can not be compared to digital fingerprints or DNA in terms of robustness. This is the reason why it is interesting to study this modality in order to understand invariable and variable features of the voice. So, the aim of this paper is, after a classification of voice disguise, to present the different approaches in the literature and last to indicate some directions of research to establish an automatic detection.

2 Classification of voice disguise

The difficulty in classifying the voice disguise is to determine what a normal voice is. Some people speak naturally with a creaky voice, while some others with a hoarse voice. A disguise is applied when there is a deliberate will to transform one's voice to imitate someone or just to change the sound. Distinguish electronic and non electronic alteration appears as a good method of classification.

Table 1. Classification of voice disguise

	Electronic	**Non-electronic**
Voice conversion	Vector quantization LMR (Linear Multivariate Regression) GMM (Gaussian Mixture Model) Indexation in a client memory …	Imitation
Voice transformation	Electronic device Voice changer software	Mechanic alteration prosody alteration

In the first category (electronic), there are two kinds of disguise: conversion and transformation. The first one consists in transforming a source speaker voice to sound like a target speaker voice, and the second one in modifying electronically some specific parameters like the frequency, the speaking rate and so on, in order to change the sound of the voice. There are many free software which offer to modify the register of his own voice [45].

In the second category (non-electronic), we consider as voice conversion the imitation by a professional impersonator and as voice transformation the different possibilities to change one or more parameters of the voice as described below. The impostor can use a mechanic system like a pen in the mouth for instance, or simply use a natural modification of his voice by adopting a foreign accent for example. Some features that could change are presented in table 2 as proposed in [33].

The techniques used for the both categories are detailed with references in 3.

Table 2. Non-electronic voice disguise

Prosody	Deformation	Phonemic
Intonation	Pinched nostrils	Use of dialect
Stress placement	Clenched Jaw	Foreign accent
Segment lengthening or shortening	Use of bite blocks (pipe smoker speech)	Speech defect
Speech tempo	Pulled cheeks	Hyper-nasal (velum lowered throughout
	Object over mouth	
	Objects in mouth	
	Tongue holding	

The different terms used in this table are:
Prosody: indicators of many channels of linguistic and para-linguistic information
Deformation: forced physical changes in the vocal tract
Phonemic: use of unusual allophones

3 State of the art

The question of voice disguise has been studied since the 1970's [16] [18] [25]. The main field of application at the beginning of those studies was intelligence and forensic science. The first works determined some means to change the voice quality and to evaluate the ability to identify the disguise. The aim of this section is to describe these different approaches adapting them to the classification presented above. As stated previously, a deliberated intention is considered as part of the definition of the voice disguise.

3.1 Electronic: voice transformation

Voice transformation is defined as the technique of changing the voice sounds by different means but without intention of imitating another voice. In this section, our purpose focuses on a presentation of change voice techniques based on electronic devices. The interest of those changes is to conceal the identity of the speaker. This could be used by radio or television to mask the identity of a person being interviewed but also in the case of anonymous or miscellaneous calls. However these means are relatively uncommon, representing one to ten percent of voice disguise situations [26]. Today with the development of internet, more and more software is available to change the voice [45]. The main technique, proposed by these software, consists in modifying the pitch register by a simple move of the mean or the pitch contour. It is also possible to add some specific effects to your original voice. These software proposes a modification of your voice pitch or your voice timbre in real time and an unlimited number of new voices. This offers an easy way to disguise your voice in voice chat, PC Phone, online gambling, voice mail, voice message and so on.

The best way to change the voice consists in modifying the prosody by using a modification of the pitch or of the duration of the segments. There are some methods able to modify the time scale or the frequency range independently. The change of the time scale offers the possibility to alter the duration of the signal without changing the frequency properties. The change of the frequency range is the contrary, that is to say modifying the pitch of a sound without changing the duration, and keeping the position of formants.

A technique which gives some interesting results is the TD PSOLA (Time Domain Pitch Synchronous Overlapp and Add) [30] [12]. This method proposes a flexible creation and modification of high quality speech. Using PSOLA, prosodic changes are easily performed. The following figure describes this technique:

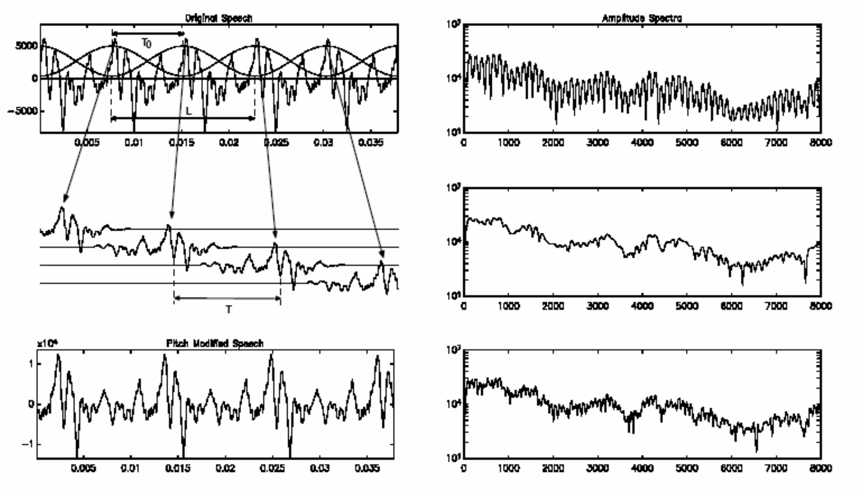

Fig. 1. TD PSOLA technique

3.2 Electronic: voice conversion

There are many techniques which have been developed in voice conversion because of the extended field of applications: speech synthesis, interpreted telephony or very low rate bit speech coding. Voice conversion is the process of transforming the characteristics of speech uttered by a source speaker, such that a listener would believe the speech was pronounced by a target speaker. Different techniques are possible for a voice conversion:

- spectral conversion [1][9][23][24][34][36][37][42]
- indexation in a client memory [32]

Generally the automatic method applied to convert a voice is based on a spectral conversion. From a set of features x = [x1, x2, ...xn] characterising a succession of source speaker speech sounds, and a set of features describing these same sounds but produced by a target speaker, the aim of the conversion is to find a transformation function between both sets and to apply it to a new sentence uttered by the source speaker. A spectral transformation can be performed by finding the conversion function F that minimizes the mean square error: $\varepsilon_{mse} = E[\|y - F(x)\|^2]$, where E is the expectation. The main approach begins by the construction of a training phase (fig.2) in which the sets from source and target voice are first aligned and then used to define the function to map the acoustic space of the source to that of the target (Fig.3). So, a voice conversion system is composed of three steps:

- modelling step: this step is constituted by the parameterisation step which consists in extracting the acoustic features (MFCC: Mel frequency cepstral coefficients) and by the modelling of these features
- transformation step: this step consists in elaborating the mapping between the source and the target voice by minimizing the distance
- Synthesis step: from the transformation function, the acoustic features of source speech are mapped and the perceptual qualities of the synthesised speech are maintained.

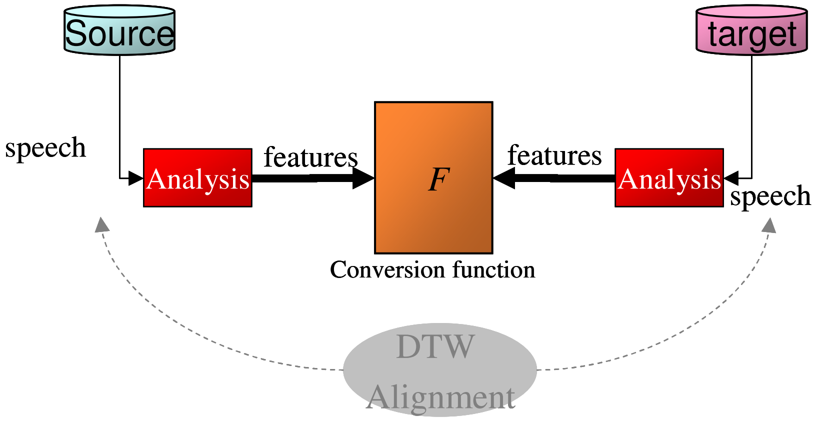

Fig.2 Training step

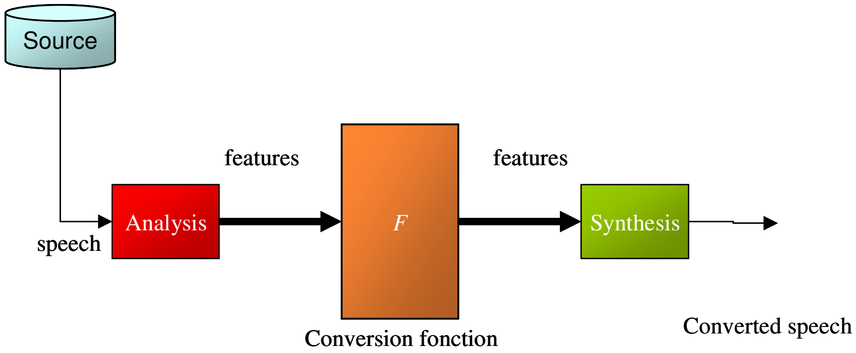

Fig.3 Conversion step

The historical technique has been proposed by [1] who described the means to elaborate a mapping codebook between a source and a target voice. The determination of the codebook is based on vector quantization. The main problem of this method is the question of discontinuities when moving from one codebook to another over time.

Another original technique is presented in [34]. It differs from the previous by proposing a much more natural output by using the PSOLA (Pitch Synchronous Overlap and Add) synthesis and by proposing two methods to learn the spectral mapping: LMR (Linear Multivariate Regression) and DFW (Dynamic Frequency Warping). The principle of the first one is a projection of the acoustic space of the source speaker into the acoustic space of the target speaker and the second one consists in elaborating an optimal non-linear warping of the frequency axis. Both techniques provide some reasonable results even if the LMR (Linear Multivariate Regression) method is better. The main problem is the presence of audible distortions which disturb the result.

Lots of papers on voice conversion are based on GMM (Gaussian Mixture Models) approaches in order to find the statistical relation between the spectral envelopes of two different speakers who pronounced the same sentence. The works proposed in [36] [37] uses a continuous probabilistic model of the source envelope. The source and the target speech were first aligned by DTW (Dynamic Time Warping). Then MFCC (Mel Frequency Cepstral Coefficient) were calculated for each frame of speech, and a vector was produced by the source MFCC followed by the same frame target MFCC. A GMM was fitted to this data, using the EM (Expectation Maximization) algorithm. The aim of the conversion function is to transform the source data set into its counterpart in the data target set. This method increases the voice quality of the conversion by an attenuation of the discontinuities. Based on the same mapping of the spectral envelope as Stylianou and al., Kain proposes in [23] [24] another technique where he predicts the residual from the predicted spectral envelope. This technique provides a higher quality transformation.

Another work different from the previous has been proposed in [32]. The principle of this system 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. Such analysis is not based on a priori linguistic knowledge [10]. Firstly, a collection of speech segments is constituted by segmenting a set of training sentences, all pronounced by the target voice. This step is performed using the temporal decomposition algorithm [5] 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. A set of HMMs (Hidden Markov Models) is then trained on this data, providing a stochastic model for each class. The result of the ALISP (Automatic Language Independent Speech Processing) training is an inventory of client speech segments, divided into 64 classes according to a codebook of 64 symbols. The encode phase consists in replacing the segment of the source voice, recognized by HMM, by their equivalent of the target voice. The results (Fig.4) on an automatic speaker recognition system are proposed below where we notice a significant degradation of the recognition performance.

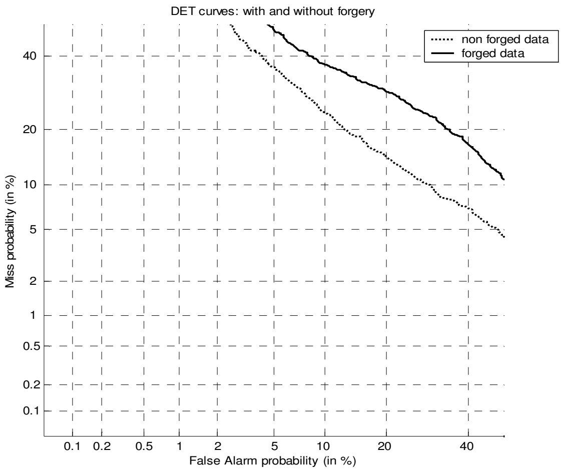

Fig. 4 DET curve: voice forgery

3.3 Non Electronic: voice conversion

Imitation is a well-known case of voice conversion more often used in the field of advertising. Zetterholm [38] and Mejvaldova [28] studied the different techniques used by a professional imitator to impersonate some voices. The principle is based on the impersonation of some specific characteristics of a target voice linked to prosody, pitch register, voice quality, dialect or the speech style. It is impossible for an impersonator to imitate all the voice register of the target speaker, but some specific parameters are enough to disturb the recognition. E. Zetterholm has demonstrated that the impersonator adapted his fundamental frequency and the position of his formants to the target voice. In [6] a study presents the difficulties of an automatic recognition system compared to a case of impersonation. A professional impersonator has imitated two target speakers: Blomberg and al. show the impersonator's adjustments, of the formant positions.

3.4 Non Electronic: voice transformation

This category of voice disguise includes the alteration of voice by using a mechanic system like a pen in the mouth and an handkerchief over the mouth, or by changing the prosody like dialect, accent or pitch register to get a low or high frequency voice modification, in order to trick the identity perception. Modifications in the rate of speech, the use of deliberate pause, the removal of extraneous syllables between words, the removal of frequency vocal fry and the heightening of prosodic variation (intonation and stress) are basically the more common voice transformation.

- Mechanic alteration [13] [43][44]

What we mean by mechanic alteration is the use of ways to disturb an original voice. Those are some ways like a pen in the mouth, parallel to the lips, a handkerchief or a hand over the mouth, or the action of pinching the nose. These types of disguise influence the quality of the voice by reducing or modifying for instance the zero of the transfer function of the vocal tract and actually the speech production system.

- Prosody alteration [19][29][31]

In this section we are interested in the deliberated changes of the voice prosody that is to say the supra segmental characteristics of the signal. Several parameters are able to influence the prosody of an individual such as:

- the physiological characteristics of the vocal tract (linked to the age and the gender)
- the emotion
- the social origin (accent)
- region origin (dialect)
- etc …

In theory a subject could modify all of those parameters to disguise his voice by changing the prosody. A classification of prosody is provided in [4], where three components could be proposed:

Table 3.

Production	Acoustic	Perception
Vocal folds tension	Fundamental Frequency (F0)	Pitch - melody
Subglottic pressure	Intensity	Sone
Air flow	Interval	Rythm – duration

Actually, the easiest prosody features to change a voice are the acoustic parameters and the perception parameters. Some of them can be easily measured.

- Fundamental frequency: estimation of the larynx frequency from a speech signal
- Duration: including speech rate, duration and distribution of the pause, syllable duration
- Intensity: the signal energy during a time interval

4. Research program to detect voice disguise

The question of characterizing different kinds of voice disguise and recognizing the original voice by automatic means has not been a subject of much study.

The main reason of this lack of study could be that the alteration caused by disguise has some important consequences on the voice quality and on the different features that provide a capacity for recognition. Nevertheless some studies have been published in this field, and in this section we are going to examine the different works before presenting some directions of research.

A rare work has been carried out in the field of mechanic disguise [13] by holding a pencil between the teeth, parallel to the lips. They analysed the effect of this disguise on the first three formants of seven oral vowels and detailed the proportional alteration of those formants. They noticed that perceptually the most evident effect was the lowering of the high vowels. The superimposition of settings of lips, jaw, and tongue affected speech segments to differing degrees, depending on the phoneme's susceptibility. In [26], Masthoff deals also with the problem of using a means to block some parameters of the vocal track and he notices that when more than one speech characteristic are simultaneously changed, the identification task is significantly more difficult for listeners. So we can reasonably conclude that this is the same for automatic detection.

In [20] the author proposes a study on three kinds of disguise on reading sentences: raising fundamental frequency, lowering fundamental frequency, denasalization by firmly pinching their nose. By focusing his work on F0, Künzel has showed that there is a direct and constant link between the F0 of a speaker's natural speech behaviour and the kind of disguise he will use. Speakers with higher-than-average F0 tend to increase their F0 levels. This process may involve register changes from modal voice to falsetto. Speakers with lower-than-average F0 prefer to disguise their voices by lowering F0 even more and often end up with permanently creaky voice.

A particular disguise, the whispered voice, has been studied in [31]. In this kind of disguise we can easily understand that the F0 and the pitch will be disturbed and even eliminated. There are also some influences on the available information about vocal intensity and voice quality.

The phenomenon of “creaky voice” that is to say the action of lowering the fundamental frequency has been examined in [21] [29]. S. Moosmüller studied the modifications of the vocal tract in this kind of disguise and the effect on formant frequencies. She noticed a sex difference in the analysis of the second and the third formant. Her work is based on 750 creaky and modal vowels pronounced by 5 female and four male speakers. It appears that creaky vowels uttered by women show a lower second formant as compared to the same vowels produced by the same speaker using modal phonation.

In an interesting work in [20], the authors analyse the effect of three kinds of disguise (falsetto speech, lowered voice pitch and pinched nose) on the performance of an automatic speaker recognition system. The experiment is limited to the estimation of the performance degradation when the suspect is known as being the speaker of the disguised speech. The results are depending on the reference population. If it contains speech data which exhibits the same type of disguise the influence is marginal on the performance. On the contrary, if the reference population is assembled with normal speech only, the effects have important consequences on the performance of the system for the three cases of disguise. Those different works reveal the lack of a global study on voice disguise; global in the sense of the number of disguises, the number of features studied, but also in the sense of technical approaches used.

In order to complete those different works on voice disguise, we have decided to examine the different parameters studied in the case of the pathological voice [17] and the analysis of emotion [2] [15] [35] and to determine if we could apply these works on our study. In addition, we propose to compare the classification of these features to the classification of features used in automatic speaker or speech recognition, the MFCC (Mel Frequency Ceptstral Coefficient) applied on voice disguise.

4.1 Principles of the study

Our work focuses on the following disguises:

- pinched nostril voice
- high pitched voice
- low pitched voice
- a hand over the mouth
- electronic voice transformation
- electronic voice conversion

The first work consists in building a voice disguise database. According to our work we impose ourselves two major constraints. The first one is that the speech must be intelligible, and the second one is that the chosen disguise is commonly used. These conditions are necessary for work oriented towards forensic sciences. The training session and the tests will be carried out using a database developed in our lab. The speech samples will be collected in a controlled environment sampled at a 16 bit resolution. 50 French people will be recorded according to a specific protocol. Each of them will pronounce 10 phonetically balanced sentences, the different vowels of the vocalic triangle, an article from a common newspaper and a phonetically balanced text. There will be five sessions of recording: normal voice, high pitched voice, low pitched voice, hand over mouth voice, pinched nostril voice. Electronic transformation and conversion will be based on their normal voices. The database is divided in two sets: one for training (35 people) one for testing (15 people).

The following figure shows a block diagram describing the process of the automatic disguise detection system.

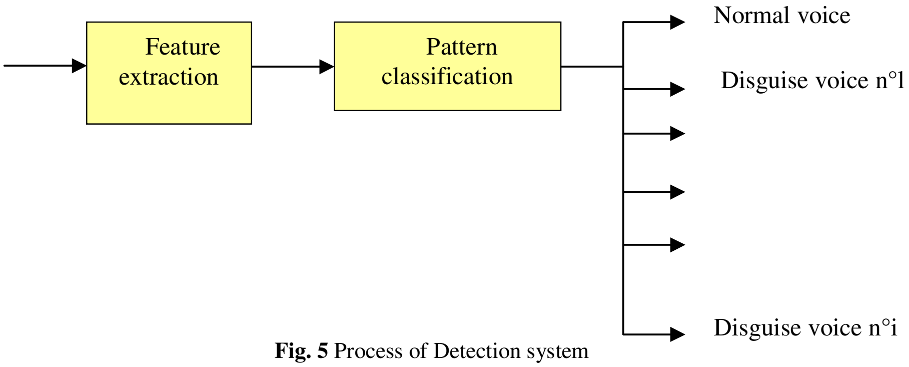

Fig. 5 Process of Detection system

Our aim is to compare two kinds of approaches: the first one is based on the principle of an automatic speaker recognition system by classification of cepstral coefficients: MFCC (Mel frequency Cepstral Coefficient); the second one regards the classification of each kind of disguise based on the prosodic parameters. In this second approach we plan to estimate some boundaries between the disguise (inter disguise) but also to analyse the main components for each kind of disguise (intra disguise).

4.2 Approach based on automatic speaker recognition system by classification of spectral coefficient

The principle of this method is to build a Universal Background Model on the one hand for all disguises in order to detect if a voice is disguised or not and on the other hand for each kind of disguise in order to classify a speech sample in one of the disguises studied. The first work consists in extracting the features. We choose in this phase to use the MFCC. The mel frequency cepstral coefficient (MFCC) is one of the most important features required among various kinds of speech applications. These features represent the spectral contour of the signal and are captured from short time frame of the speech signal. Then a Fourier transform is applied to calculate the magnitude spectrum before quantifying it using a mel spaced filterbank. This last operation transforms the spectrum of each frame in a succession of coefficients which characterize the energy in each frequency bandpass of the mel range. A DCT is applied on the logarithm of those coefficients.

In addition to the MFCC, we calculate the first and the second derivatives of the MFCC in order to take into account dynamic information which introduces the temporal structure of the speech signal like for instance the phenomenon of co-articulation.

The second main step of this approach is the creation of a universal background model by training for all disguise and for each kind of disguise. The using of the first UBM (all disguise) is to discriminate a disguised voice from a normal voice. This UBM will be built on a statistic modelling based on GMM (Gaussian mixture models) on the entire disguised voice corpus. A Gaussian mixture density is a weigthed sum of M component densities:

$$f(x/disgMod)=\sum_{j=1}^{M} g_j N(x,\mu_j,\Sigma_j)$$

where

x is a D dimensional vector resulting from the feature extraction
j = 1 to M are the component densities
g_j are the mixture weigths
disgMod is the disguise model

The different parameters (g_j, μ_j , Σ_j) of the UBM will be estimated by the EM (Expectation – Maximisation) algorithm. This algorithm is a general technique for maximum likelihood estimation (MLE). The first step of the algorithm (E: estimation) consist in estimating the likelihood on all the data from the last iteration and the second step (M: maximisation) consists in maximizing the density function of the first step.

The same principle is used to estimate the specific UBM of each kind of disguise.

The last step of this approach is the verification phase which consists in knowing if an unknown recording comes from a normal voice of from a disguised voice and what kind of disguise by calculation of a likelihood ratio as follows.

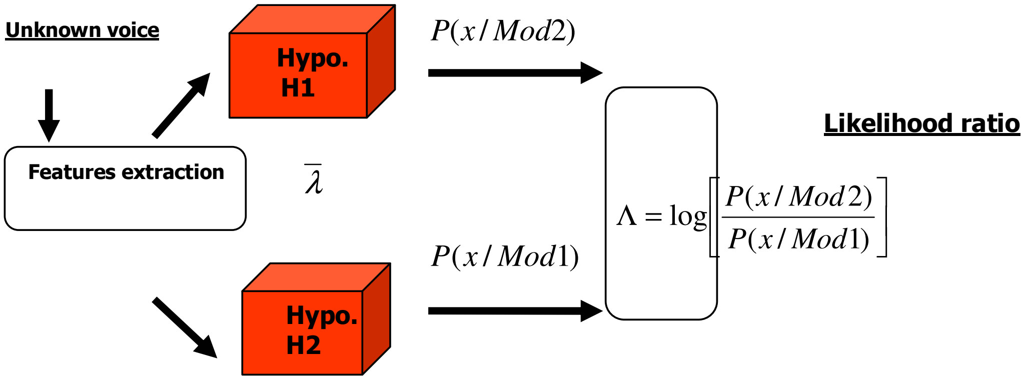

Fig. 6 Verification task: test phase

H1 : the unknown voice is a normal voice
H2 : the unknown voice is a disguised voice
Mod1: normal voice UBM
Mod2: disguised voice UBM

4.3 Approach based on analysis and classification of prosodic features

The second objective consists in detecting the disguise by studying some specific prosody parameters. Perceptually, it is easy to notice a difference between an original voice and a voice transformed by such a technique, but the question is what the main parameters that have been changed are. We can expect that each kind of disguise will have some specific parameters that will change.

We are using PRAAT [8] for a statistical study on

<u>F0 features:</u>

- min, max, mean, std,
- Jitter: this parameter measures the instability of frequency in each cycle:

$$\frac{\frac{1}{n-1}\sum_{i=n-1}^{1}\left|F0_{i+1}-F0_{i}\right|}{F0_{mean}}$$

<u>Energy features:</u>

$$E(x)=\int\left|x(t)\right|^{2}dt \rightarrow \text{idB} = 20\ \log 10(E(x))$$

$$i\text{dB} = 20\ \log 10(E(x)/\text{Eo}).$$

- mean: Eo
- energy proportion in 5 Frequency bandwidths (0 - 1kHz - 2kHz – 3kHz – 4kHz - 5KHz).
- The shimmer: this parameter measures the instability of amplitude in each cycle.

$$\frac{\frac{1}{n-1}\sum_{i=n-1}^{1}\left|E_{i+1}-E_{i}\right|}{E_{à}}$$

<u>Rhythm features:</u>

- Speaking rate based on the calculation of number of phoneme per second
- Voiced/unvoiced region
- Ration of speech to pause time

<u>Formant features for each vowels of the vocalic triangle:</u>

- F1: characteristic from the mouth opening
- F2: characteristic from the tongue position
- F3: characteristic from the lips contour

After this extraction phase, the aim is to organize and classify the different features in order to find boundaries inter disguise.

Inter disguise

Contrary to the first approach, where we used the MFCC which have the property of being uncorrelated, a first work in this approach consists in extracting the main information in the global distribution of the coefficient before classification. Different methods of data analysis and classification will be used in order to select the main features which influence a disguise and reduce the dimension via a PCA (Principal Component Analysis) for instance. The use of PCA allows the number of variables in a multivariate data set to be reduced, whilst retaining as much as possible of the variation present in the data set. This reduction is achieved by taking p variables X_1, X_2,…, X_p and finding the combinations of these to produce principal components (PCs) PC_1, PC_2,…, PC_p, which are uncorrelated. These PCs are also termed eigenvectors. PCs are ordered so that PC_1 exhibits the greatest amount of the variation, PC_2 exhibits the second greatest amount of the variation, PC_3 exhibits the third greatest amount of the variation, and so on. The aim is to be able to describe the original number of variables (X variables) as a smaller number of new variables (PCs).

We will also apply a LDA (Linear Discriminant Analysis) in order to increase the class distribution by optimization of the ratio between the intra class the inter class. After this step of data description and organization, we will structure our database by applying supervised classification methods based on training.

The classification step planned is based on GMM clustering as previously explained. The principle is to build a GMM model for each kind of model based on a training phase and to calculate a distance which will be the likelihood function between a test vector and the model.

Intra disguise

From the same set of prosodic features, a study in order to evaluate the main characteristics of each of disguise has been planned. The idea is to extract the more significant components for each disguise and the influence of some specific features. The algorithm used is a PCA.

And last, we investigated a measurement of the vocalic triangle [39][40][41] moving between a normal voice and a disguise n°k voice. The study based on a limited corpus of voice disguised provides interesting results as shown by the following figure.

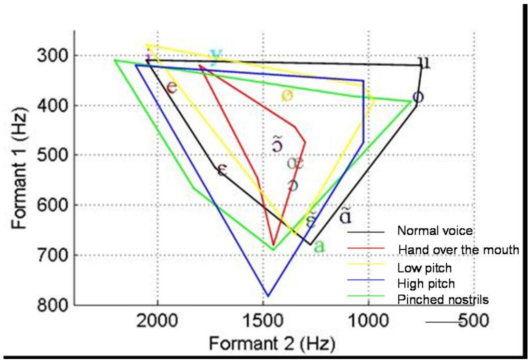

Fig. 7. Vocalic triangle

This vocalic triangle is based on a corpus of twenty people who pronouced the different french vowels in five disguise voice (included normal voice) .

The aim of the PCA and the comparison of vocalic triangle analysis are to find some specific clues in each kind of disguise.

Our last objective will consist in:

- Fusing both approaches in order to increase voice disguise recognition
- Synthesising a disguised voice from the original voice of the suspect and to evaluate the distance between this voice and the unknown voice.
- Or applying a transformation adapted to the type of disguise determined previously, to the disguised voice and comparing it to an undisguised voice of the suspect.

Such a goal will be certainly possible in the case of simple disguise, like a lower pitch for instance, but if the impostor uses different kinds of disguise, the task will be more complicated.

*   *

*

To conclude, this paper presents the problem of voice disguise under different aspects. In the first part, we propose a classification where we distinguish different possibilities of disguise depending on the means employed. The question of disguise is considered under the aspect of a deliberate action in order to falsify identity. All the problems of voice transformation caused by channel distortion are not studied. A review of different works on some particular disguises is detailed in the second part in order to understand the difficulties encountered in characterizing disguises. An important problem could be the using of different disguises in the same time. But the main studies in the forensic field reveal that in the most case the impostor has just used a specific disguise. Lastly in a third section, we first present the relative importance of specific features in different kinds of disguises and, secondly we propose some directions of research to evaluate the impact of disguise on an automatic speaker recognition system and to determine the disguise by finding out some indicators linked to study of emotional or pathological voice.

References

1. M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, “Voice conversion through vector quantization,” *Proc. ICASSP 88,* New-York, 1988

2. N. Amir, “Classifying emotions in speech: a comparison of methods” in Proceedings EUROSPEECH 2001, Scandinavia

3. G. Baudoin, J. Cernocky, F. El Chami, M. Charbit, G. Chollet, D. Petrovska-Delacretaz. “Advances in Very Low Bit Rate Speech Coding using Recognition and Synthesis Techniques,” *Proc. Of the 5th Text, Speech and Dialog workshop, TSD 2002*, Brno, Czech Republic, pp. 269-276, 2002

4. F. Beaugendre, “Modèle de l’intonation pour la synthèse”.1995 de la parole", in "Fondements et perspectives en traitement automatique de la parole", Aupelf-Uref (ed.).

5. F. Bimbot, G. Chollet, P. Deleglise, C. Montacié, “Temporal Decomposition and Acoustic-phonetic Decoding of Speech,” *Proc. ICASSP 88*, New-York, pp. 445-448, 1988

6. M. Blomberg, Daniel Elenius, E. Zetterholm, “Speaker verification scores and acoustics analysis of a professional impersonator,” *Proc. FONETIK 2004*

7. R. Blouet, C. Mokbel, G. Chollet, “BECARS: a free software for speaker recognition,” *ODYSSEY 2004*, Toledo, 2004

8. P. Boersma, D. Weenink, “PRAAT: doing phonetics by computer. http://www.praat.org”

9. O. Cappe, Y. Stylianou, E Moulines, “Statistical methods for voice quality transformation,” *Proc. of EUROSPEECH 95*, Madrid, 1995

10. G. Chollet, J. Cernocky, A. Constantinescu, S. Deligne, F . Bimbot, “Toward ALISP: a proposal for Automatic Language Independent Speech Processing,” *Computational Models of Speech Processing*, NATO ASI Series, 1997

11. V. Delvaux, T. Metens, A. Soquet. “French nasal vowels: articulary and acoustic properties”, Proc. Of the 7th ICSLP, Denver, 1,53-56,2002

12. T. Dutoit, “High quality text to speech synthesis: a comparison of four candidates algorithms”, Proc. ICASSP 1994 vol.1 pp 565-568, Adelaïde, Australie.

13. Figueiredo Ricardo Molina de, Britto, Helena de Souza “A report on the acoustic effects of one type of disguise”, Forensic Linguistics, 1996, 3, 1, 168-175

14. D. Genoud, G. Chollet, “Voice transformations: some tools for the imposture of speaker verification systems,” *Advances in Phonetics*, A. Braun (ed.), Franz Steiner Verlag, Stuttgart, 1999

15. D. Gibbon, U. Gut, “Measuring speech rhythm”, Proc. Eurospeech 2001, Scandinavia

16. W. Endres, W. Balbach, G. Flösser, “Voice spectrograms as a function of age, voice disguise and voice imitation”, Journal of the Acosutical Society of America, 49:1842-8, 1971

17. L. Gu, J.G. Harris, R. Shrivastav, C. Sapienza, “Disordered speech evaluation using objective quality measures”, Proc. ICASSP 2005, Philadelphie

18. M. Hall, “Spectrographic analysis of interspeaker and intraspeaker variability of professional mimicry”, MA dissertation, Michigan State University. 1975

19. Hermann J. Künzel “Effects of voice disguise on fundamental frequency”, Forensic linguistics 7:149-179, 2000

20. H. Künzel, J.Gonzalez-Rodriguez, J. Ortega-Garcia, “Effect of voice disguise on the performance of a forensic automatic speaker recognition system”, Proc. Odyssey 2004

21. A. Hirson, Duckworth M, “Glottal fry and voice disguise: a case study in forensic phonetics”, Journal of Biomedical Enginering vol 15:193-200, 1993

22. D.Jiang, W. Zhang, L. Shen, L. Cai, “Prosody analysis and modelling for emotional speech synthesis”, Proc. ICASSP 2005, Philadelphie

23. A. Kain, M. W. Macon, “Spectral voice conversion for text to speech synthesis,” *Proc. ICASSP 98,* New-York, 1998

24. A. Kain, M. W. Macon, “Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction,” *Proc. ICASSP 01,* Salt Lake City, 2001
25. R.C. Lummis, A.E Rosenberg, “Test of an automatic speaker verification method with intensively trained professional mimics”, Journal of Acoustical Society of America, vol 9, number 1, 1972
26. H. Masthoff. « A report on voice disguise experiment, Forensic Linguistics”, 3(1) :160-167. 1996
27. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, “The DET curve in assessment of detection task performance,” *Proc. EUROSPEECH 97*, Rhodes, Greece, pp. 1895-1898, 1997.
28. J. Melvaldova, “Caractéristiques temporelle de la parole imitée”, Proceedings JEP (Journées d’Etudes sur la Parole) 2004
29. S. Moosmüller, “The influence of creaky voice on formant frequency changes”, The International Journal of Speech, Language and the Law, vol8, n°1, 2001
30. E. Moulines, F. Charpentier, “Pitch synchronous waveform processing techniques for text to speech synthesis using diphone”. Speech comm. vol.9 p 453-497
31. T. Ochard, A. Yarmey, “The effects of whispers, voice sample duration and voice distinctiveness on criminal Speaker Identification”, Appl. Cogn. Psychol., 31:249-260
32. P. Perrot, G. Aversano, R. Blouet, M. Charbit, G. Chollet, « voice forgery using ALISP » Proc. ICASSP 2005, Philadelphie
33. R. Rodman, Speaker Recognition of disguised voices: a program for research, consortium on Speech Technology Conference on Speaker by man and machine: direction for forensic applications, Ankara, Turkey, COST 250, 1998
34. H. Valbret, E. Moulines, J.P. Tubach, “Voice trans-formation using PSOLA technique” *Proc. ICASSP 92*, San Francisco, 1992
35. I. Shafran, M. Mohri, “A comparison of classifiers for detecting emotion from speech”, Proc. ICASSP 2005, Philadelphie
36. Y. Stylianou, O. Cappe, “A system for voice conversion based on probabilistic classification and a harmonic plus noise model,” *Proc ICASSP 98*, Seattle, WA, pp. 281-284, 1998
37. Y. Stylianou, O. Cappe, E. Moulines “Continuous probalistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, 6(2):131-142, March 1998.
38. Zetterholm, E. Voice Imitation. A phonetic study of perceptual illusions and acoustic success. Dissertation. Department of Linguistics and Phonetics, Lund University. 2003
39. Rostolland D., 1982a, "Acoustic features of shouted voice", Acustica 50, pp. 118-125.
40. Rostolland D., 1982b, "Phonetic structure of shouted voice", Acustica 51, pp. 80-89.
41. Rostolland D., 1985, "Intelligibility of shouted voice", Acoustica 57, pp. 103-121
42. Abboud B. Bredin H, Aversano G., Chollet G. “Audio visual forgery in identity verification” Workshop on Nonlinear Speech Processing, Heraklion, Crete, 20-23 Sept. 2005
43. B.S. Atal, “Automatic speaker recognition based on pitch contours”, Journal of Acoustical Society of America, 52:1687-1697, 1972.
44. J. Zalewski, W. Maljewski, H. Hollien, “Cross correlation between Long-term speech Spectra as a criterion for speaker identification, Acoustica 34:20-24, 1975
45. http://www.zdnet.fr/telecharger/windows/fiche/0,39021313,11009007s,00.htm