SPEECH PRE-PROCESSING AGAINST INTENTIONAL IMPOSTURE IN SPEAKER RECOGNITION

*Dominique Genoud+, Gérard Chollet**

+IDIAP, CP 592 CH-1920 Martigny Switzerland, genoud@idiap.ch,

*CNRS URA-820, ENST, 46 rue Barrault, 75634 PARIS cedex 13, chollet@sig.enst.fr

ABSTRACT

Recently, some large-scale text dependent speaker verification systems have been tested. They show that less than 1% Equal Error Rate can be obtained on a test set score distribution. So far, the majority of impostor tests are performed using speakers who don't really try to fool the system. This can be explained by the lack of databases recorded for this purpose, and the difficulty for a normal speaker to transform his voice characteristics. Nevertheless, actual automatic analysis/synthesis techniques, such as Harmonic plus Noise Model (H+N) [1], allows very good speech/speaker transformations. Thus, it becomes possible to transform the voice of a speaker in the voice of another speaker in order to make voluntary impostures. This paper evaluates these kind of intrusive impostures and proposes a new speech pre-processing method, based on harmonic subtraction [7], making speaker verification less insensitive to these spectral transformations. A state-of-the-art Hidden Markov Model is used as reference system to assess the transformation results. The speech is parameterised by LPCC coefficients. The results are obtained on a database of telephone speech quality. The speaker verification system works in text dependent mode.

1. INTRODUCTION

Understanding the characteristics of imposture is a problem for large-scale industrial speaker verification systems. In speaker verification, a trade-off is necessary between the false rejection of a client and the false acceptance of an impostor. Indeed, impostor modelling is used to build the speaker models, the normalisation models (*world model*) and to set the decision thresholds. The experiments reported here were performed on a Swiss French speaker verification database (Polycode) [4]. The client passwords are sequences of connected digits. On a test trial, such a sequence is first recognised automatically and therefore segmented. The speaker verification is done at word level allowing a better control of the pronounced utterance. In the **section 2**, transformation of impostors will illustrate the problem of imposture using automatic analysis/synthesis techniques. The **section 3** will describe new pre-processing techniques inspired by the harmonic plus noise model, which can make speaker verification systems less sensitive to spectral (cepstral) transformations.

2. A TECHNIQUE FOR DELIBERATE IMPOSTURE

If we suppose that an impostor could record some samples of a registered customer, it could be possible for him to transform his voice in order to mimic the client voice. In the first part of this paper, we investigate possibilities of transforming word by word a digit sequence uttered by a source speaker (the impostor) into a sequence looking like an utterance of the target speaker (the client). This exploratory system shows that it is possible to fool a state-of-the-art text dependent speaker verification system. Future work will use phoneme to phoneme transformations.

2.1. Harmonic plus Noise modelling

The speech is modelled by a Harmonic plus Noise model (H+N) which allows good quality spectral modifications. This model is normally used in text-to-speech applications [1]. The H+N model decomposes the speech into a harmonic part and a noise part. From the harmonic analysis of speech, cepstral coefficients (c_i) are extracted and from the noise part reflection coefficients (k_i) are estimated. These coefficients (c_i and k_i) are then used to re-synthesise the speech.

2.1.1. Analysis of the harmonic part:

The voiced part of the speech signal can be modelled as a fundamental frequency and its harmonics. The fundamental frequency of the human speech (often called *pitch*) is varying with the prosody. In order to estimate the amplitudes of the fundamental frequency harmonics, a signal analysis is performed pitch synchronously on short temporal windows (typically, 25ms overlapped each 10ms). On every window, it is supposed that the pitch f_0 is constant, and that the harmonics are the sum of complex exponential functions (equation 1).

$$\hat{h}(t) = \sum_{k=-L}^{L} A_k(t_a^i) e^{j2kf_0(t_a^i)(t-t_a^i)} \quad (1) \qquad \varepsilon = \sum_{t=t_a^i-N}^{t_a^i+N} \omega^2(t)(s(t)-\hat{h}(t))^2 \quad (2)$$

In equation (1) A_k denotes the complex Amplitude of each harmonic *k*, *L* denotes the number of chosen harmonics and the pitch synchronous analysis instant t_a^i. The harmonics are determined by minimisation of the quadratic error ε between the original signal *s(t)* and the estimated harmonics $\hat{h}(t)$ (equation 2). The cepstral coefficients are extracted from the harmonic analysis part. Phase and amplitude envelope estimations are performed to allow for modifications of the pitch when re-synthesising the signal. Then real cepstral coefficients can then be estimated [2].

2.1.2. Synthesis of the harmonic part

The signal can be re-synthesised (equation 3) by re-composition of the harmonic $\hat{h}(t)$ at each synthesis time instant t_s^i using a sum of cosine functions. The amplitudes a_k are directly extracted from the cepstral coefficients, and the phases are extracted by re-sampling the spectral phase envelope at the synthesis instant t_s^i.

$$\hat{h}(t) = \sum_{k=0}^{L} a_k(t_s^i)\cos(\Phi_k(t_s^i) + 2k\pi f_0(t_s^i) * t) \qquad (3)$$

This synthesis method permits pitch modifications between the analysis instants and synthesis instants.

2.1.3. Analysis of the noise part

All unvoiced parts of the speech can be viewed as a noise source passed through filters [6]. In this approach, the spectral density function of the noise is estimated by a 16-order all-pole filter using the autocorrelation method [2]. The reflection coefficients k_i are then estimated on a 40[ms] window around the analysis instant t_a^i. As an estimation of the maximum voicing frequency is performed [1] the noise part can also be extracted from the voiced segments of speech.

2.1.4. Synthesis of the noise part

The noise part is re-synthesised using a Gaussian noise source and a normalised lattice filter using the k_i coefficients extracted at analysis time.

2.2. Speaker transformations

Given the coefficients for the harmonic and the noise parts of a source [0734_1.wav] [0734_2.jpg] and a target [0734_3.wav] [0734_4.jpg] speaker, the idea is to map these coefficients from the source to the target and use them to re-synthesise the utterance of the transformed source. The cepstral coefficients are independent [2], and we assume that, on short speech events, the distribution of each coefficient follows a Gaussian law. At first we consider that the speech event has the duration of a word and that each Gaussian distribution $N(\mu_{i,source},\sigma_{i,source})$ of the source coefficient $c_{i,source}$ will be mapped to the distribution $N(\mu_{i,target},\sigma_{i,target})$ of the target using equation (4).

$$T(c_{i,source}) = \left(\frac{\sigma_{i,target}}{\sigma_{i,source}}\right)(c_{i,source} - \mu_{i,source}) + \mu_{i,target} \qquad (4)$$

Subsequently the duration of a speech event will be a word part determined by the state occupation of a HMM. One speaker independent HMM by word is used to align the vectors of the source and the target, state by state. The distribution of each vector component in each state is then assumed Gaussian and the transformation of the source is performed state by state using the equation (4).

2.3. Speaker re-synthesis

The transformed coefficients of the harmonic part are then injected into the synthesis part of the H+N model (see paragraph **2.1.2**). As tentative trials for noise transformations gave imposture results worst than unmodified imposture tests, two ways were followed: the first one keeps the original *noise source* [0734_5.wav] [0734_6.jpg] re-synthesised with the transformed harmonic part. The noise source is extracted by subtracting the harmonic part from the source signal. The second one adds only a *random background noise* [0734_7.wav] [0734_8.jpg] to the transformed harmonic part. The random background noise is built from randomly selected samples of a non-speech part utterance.

3 RESULTS OF TRANSFORMATIONS

3.1. Database

The results are obtained on a database [4] composed of a subset of 28 speakers recorded over a telephone line in several sessions. During the same session, each speaker had to say, among other sentences in French, 4 times his own 7-digit PIN code and one time a 10-digit sequence (all the digits from 0 to 9 in different order for each sequence). All these sequences are time-labelled digit by digit using a speech recogniser. Some subsets are extracted from this Polycode database (see **figure 1**).

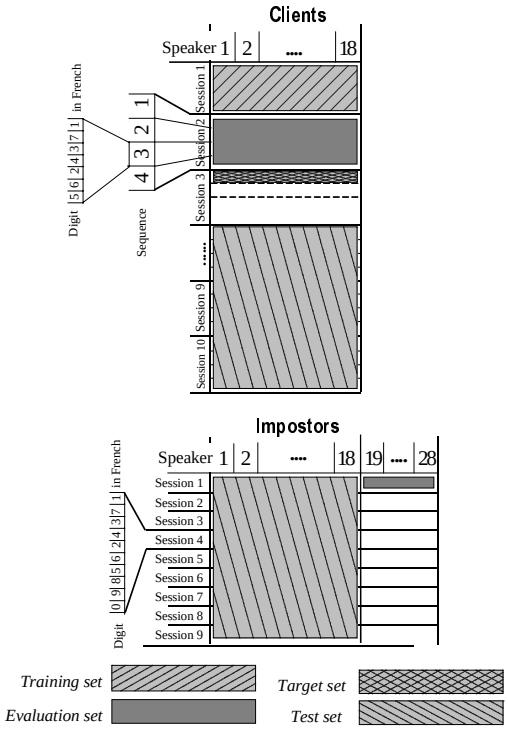

Figure 1: The sets composing the database.

3.2. The reference system

3.2.1 Modelling

The automatic speaker verification system (ASV) used here as reference is a state-of-the art HMM (Hidden Markov Model) based system and operates in text dependent mode. Two HMM models are created. One is speaker independent (the ***world model***), trained with 300 speakers on a database different from Polycode. This world model is used as a normalisation model [4]. A ***speaker model*** derived from the world model is then re-estimated on the *Training set*. The scoring is done by computing the log likelihood ratio (*LLR*) of the log likelihood of the speaker

model $Lk_{speaker}$ and the log likelihood of the world model Lk_{world} along an utterance (equation 5).

$$LLR = \log(Lk_{speaker}) - \log(Lk_{world}) \quad (5)$$

3.2.2 Parameterisation

The input signal of the reference system is windowed every 10ms, each window has a duration of 25ms. A pre-emphasis of 0.97 and a Hamming window are applied on each window. Then 12 LPCC coefficients, the energy, their derivatives and accelerations are extracted, constituting a vector of 39 parameters.

3.3. Speaker transformations

The reference system is trained with the *Training set* and **a speaker dependent a priori threshold (set at Equal Error Rate)** is computed using the *Evaluation set*. The reference system is then used with the *Test set* in two different way:

1. The impostor data of the *Test set* are given to the reference system, a decision is taken by comparing the scores of the utterances of the clients and impostors to the speaker dependent a priori threshold. The false acceptance (FA) rate is computed as the percent of impostor utterances that are accepted wrongly as client ones.

2. For each speaker the impostor data of the *Test set* (source) are transformed using the mimic systems (Gaussian and HMM) into the utterances of the *Target set*. These transformed utterances are then given to the reference system and a new false acceptance rate is computed.

The **table 1** shows the false acceptance rate (FAR) when the Gaussian and HMM transformation systems are used with the **a priori** threshold. The first column indicates the FAR when the noise part of the source speaker is used. In this case the transformation is not efficient, indicating that the noise part of speech contains speaker dependent and probably channel information**.** When using a random noise (col.3 of **table 1**) the FAR increases considerably. The ROC curves in **figure 2** confirm the results given in **table 1**.

Method	FAR% source	FAR% random
Reference system	4.19%±0.7	4.19%±0.7
Gaussian transformation	3.59%±0.6	14.45%±1.3
HMM transformation	4.65%±0.8	**23.09%±1.5**

Table 1: False Acceptance Rate with an *a priori* threshold set at EER using Gaussian or HMM transformations with the original source noise (source), and random noise.

4. REMOVING PARTS OF SIGNAL

The previous paragraphs show an important increase of the error rate when an impostor mimics the voice of a client. As these transformations are based on spectral (cepstral) modifications, a way to hold out against them would be to suppress the spectral (cepstral) parts that can be modified. This section will evaluate the performances of this approach. The first pre-processing technique will suppress the harmonic part of the signal using the H+N model (see section 2). This technique will be compared to a technique, which keep only the residual part of a LPC signal parameterisation.

To test the two techniques are evaluated on the *Test set* (see section 3).

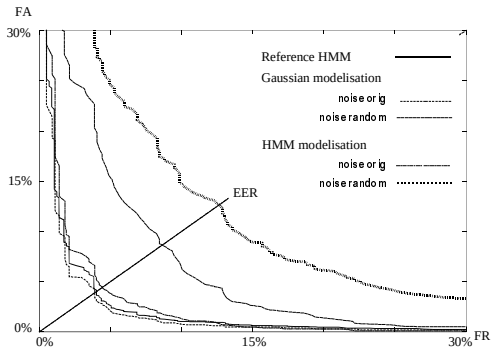

Figure 2: ROC curves for the different impostor transformations

4.1 Suppression of harmonics

This technique subtracts the harmonic part of the signal from the speech: Initially the input speech is analysed and re-synthesised by the H+N model using the equations (1) and (3). Then the synthetic signal is subtracted from the original one, sample by sample. This process suppresses all the harmonic parts, which can be modelled by the H+N system. Finally, the residual signal is given to the input of the reference system. The **table 2** shows that no loss of performance can be observed after suppression of the modelled harmonics when the system is used with an a priori threshold and no transformation of the impostors. Moreover, the roc curves of **figure 3** show no significant difference in the performance when the harmonics are removed. Obviously not all harmonics are suppressed as the analysis/synthesis is not perfect (error in pitch estimation, in the amplitude and phase determination). Thus, the remaining part contains the noise and some residual parts of the harmonics. An example of the original signal [0734_9.wav] [0734_10.jpg] and the remaining signal can be found here [0734_11.wav], and its narrow band spectrogram here [0734_12.jpg]. From this experiments we can conclude that the residual part of the H+N modelling hold relevant information able to characterise speakers.

System	FRR%	FAR%	HTER%=(FAR+FRR)/2
Reference	$2.33^{\pm 1.24}$	$4.19^{\pm 0.72}$	3.26
Sub Harmonics	$5.06^{\pm 1.81}$	$2.03^{\pm 0.50}$	3.54
LPC residual	$3.90^{\pm 1.59}$	$1.43^{\pm 0.42}$	2.66

Table 2: Computation with an *a priori* threshold of False Rejection, False Acceptance, Half Total Error (=FA+FR)/2) rates of the original reference system, with the suppression of the harmonics, and with LPC residual as pre-processing.

4.2 LPC residual

The suppression of harmonics, which consists in removing the reconstructed signal and in keeping only the residual part of the

speech, can be compared with the extraction of the residual of a LPC (Linear Prediction Coding). If the LPC can be seen as a production model of speech [3,6], the residual of the LPC is known to contain the excitation signal and the error of LPC reconstruction. The residual *u(n)* can be defined by the difference between the original signal *s(n)* and the reconstructed signal identified by an all pole filter of order *p* and coefficients a_k weighted by the gain *G* (Equation 6).

$$u(n) = \frac{1}{G}\left(s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k) \right) \qquad (6)$$

Thévenaz [5] has shown that the cepstrum coefficients of the residual are a relevant parameterisation in text independent speaker verification applications. The **table 2** shows the results obtained by using the LPC residual as a pre-processing of the reference system. These results are not significantly different from the subtraction of harmonics or from the original reference system. However, the ROC curve of **figure 3** indicates weaker performances of the LPC residual pre-processing.

4.3 Robustness to intentional imposture

The **table 3** shows the false acceptance error rate when the transformed impostors of **section 3.3** are given to the input of the reference system. The suppression of the harmonics decreases the FAR nearly of a factor 3, however it cannot suppress all the influence of the speaker transformations. These results confirm the ones of section 4.1. On the other hand, The LPC-residual doesn't hold out against the spectral transformations. The ROC curves of **figure 4** confirm the results of table 3.

System	FAR%
Reference	**23.09%±1.5**
Sub Harmonics	$8.17^{\pm 1.02}$
LPC Residuals	$15.65^{\pm 1.36}$

Table 3: Computation with an *a priori* threshold and *transformed impostors* of the False Rejection rate of the original reference system, with the suppression of the harmonics by H+N, and with LPC residual as pre-processing.

5. CONCLUSION

The results show that the harmonic part of the speech signal contains speaker dependent information that can be transformed to mimic another speaker. They also demonstrate the possibility to become partly robust to these kinds of transformations by suppressing the harmonic part of the signal. The modelling of the noise is critical: it seems that ASV systems based on HMM are very sensitive to the noise modifications. Finally they confirm that the speech signal is redundant and contain speaker dependent information in the harmonic part but also in the noise part. Further investigations will be done by using a different approach rather than the H+N model for the analysis/synthesis part especially to find a better way to model the noise and its transformation. It should be noted that these transformations are not robust against concatenation of parts of client voice that an impostor could have done.

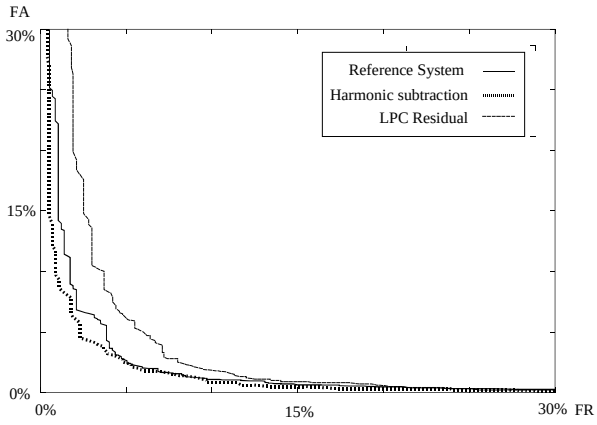

Figure 3: Computation of ROC curves of the reference system, and with harmonic subtractions or LPC residual pre-processing.

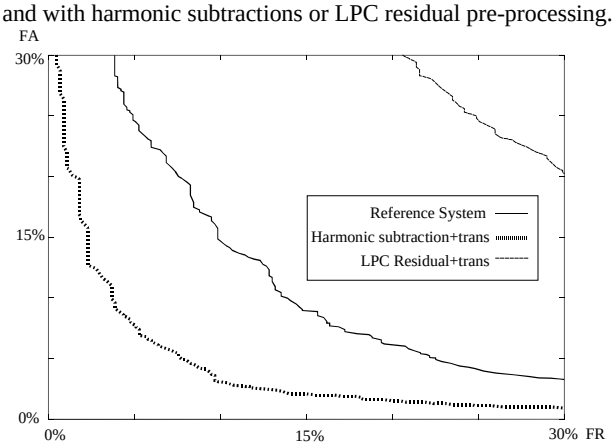

Figure 4:Computation of ROC curves when using transformed voice for impostors, for the same systems as in figure 3

7. REFERENCES

1. STYLIANOU Ioannis, Modèles Harmoniques plus Bruit combinés avec des méthodes statistiques, pour la modification de la parole et du locuteur. PhD thesis , ENST Paris 1996.
2. RABINER Lawrence, BIING-HWANG Juang, Fundamentals of speech recognition, Prentice Hall, 1993, Englewood Cliffs, NJ.
3. ROSENBERG A.E., LEE C.H., GOKOEN S, Connected Word Talker Verification Using Whole Word Hidden Markov Model", pp 381-384, in proceedings of ICASSP-91", 1991
4. GENOUD D., CHOLLET G., Polycode a Verification Database, internal doc., IDIAP 1995
5. THEVENAZ Philippe, Résidu de prédiction linéaire et reconnaissance de locuteurs independante du texte., PhD thesis, Univeristé de Neuchâtel, 199
6. MARKEL J.D., GRAY A.M, Linear prediction of speech, Springer Verlag. 1976 Berlin
7. GENOUD D., CHOLLET G., Voice Transformations, some tools for the imposture of speaker verification systems, IPS-98, Washington, 1998