AUTOMATIC SPEAKER RECOGNITION ON A VOCODER LINK

Frédéric Jauquet, Patrick Verlinde and Claude Vloeberghs
Signal and Image Processing Center (SIC)
Electrical & Telecommunication Dept.
Royal Military Academy, Av. de la Renaissance, 30 - 1000 Brussels - Belgium
Tel : +32 2 737 62 53, FAX : +32 2 737 62 53, E-mail : frederic.jauquet@tele.rma.ac.be

ABSTRACT

Automatic speaker recognition on a vocoder link has rarely been explicitly tested. In this paper, we show how the automatic speaker recognition could be used on a vocoder link. In a first experiment where we consider the "coder-link-decoder" speech system as a black box, a classic speaker recognition method (applied on the reconstructed speech) is shown to be able to provide an objective measurement of the voice quality of the vocoder. In a second experiment, the same speaker recognition method is directly applied on the information contained in the coded frames. In latter case, the recognition scores provide an interesting analysis.

1. INTRODUCTION

At the present, the use and the improvement of coding technology are continuing to boom. As the quality of low bit rate speech coders continues to improve, speech intelligibility is not a major issue and voice quality becomes the major factor. More and more, the users are now asking for speaker recognition capabilities on coded speech. Unfortunately, if the subjective tests (DRT, DAM, MOS) of voice intelligibility and quality are widely used to compare and evaluate vocoders, no speaker recognition has been tested on vocoders yet.
In order to have a good representation of the evolution of vocoders [1], four classic vocoders at different bit rates coming from standards and/or recommendations [2] are used. The four vocoders are listed in Table 1.

VOCODER	STANDARD
LPC10 (2.4 kbits/s)	NATO-STANAG 4198
CELP (4.8 kbits/s)	US FS-1016
CELP (8 kbits/s)	ITU-T Recom G.729
LD-CELP (16 kbits/s)	ITU-T Recom. G728

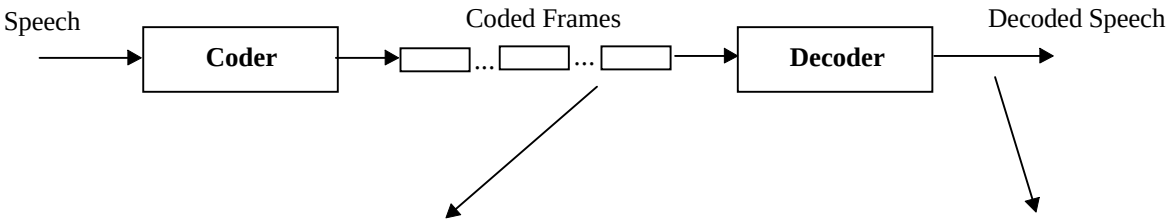
Table 1 : The 4 different vocoders used in the experiments

2. EXPERIMENTS

Figure 1 represents the schemes and the conditions of the different realized experiments.
In a first step, we do not use the technical performances of the different vocoders and each vocoder is considered as a black box. For each vocoder, the transmission of a vocoder link is simulated without any transmission error. Then a classic speaker recognition method is applied on the reconstructed speech : "Second-Order Statistical Measures" [3]. In section 4, we show that the ROC (Receiver Operating Characteristic) curves could be used to provide objective measurements to compare and evaluate the vocoders.
In a second step, we implement a speaker recognition method on a vocoder link by directly using the characteristic parameters transmitted in the coded frames. Indeed, in the coded frames, depending on the sophistication of the vocoder, speaker dependent information is transmitted such as reflection coefficients (R) or linear spectrum pair (LSP) coefficients, information about voicing and pitch and/or information about the gain-shape of the excitation. In this approach, the R and the LSP coefficients are picked up from the frames coded respectively by the LPC10 and the CELP (8 kbits/s) vocoders. From these coefficients, the 12 cepstral coefficients are computed and then passed along to our speaker recognition method. The recognition scores are less good.

3. SPEAKER RECOGNITION

The speaker recognition method considered in this work is a free-text method based on Second-Order Statistical Measures [3]. In the experiments, the speaker is characterised by two prediction matrixes estimated by the 12 cepstral coefficients from the training speech samples. To perform the recognition, only the basic form of the measure, referred to as "Arithmetic-geometric sphericity measure", is used since the aim of this study is not to achieve the very best speaker



Second Experiment :

In the transmitted frames, we pick up the characteristic parameters (12 cepstral coefficients) and we apply them into the Speaker Recognition method (AR).

First Experiment :

From the decoded speech, we compute the characteristic parameters (12 cepstral coefficients) and we apply them into the Speaker Recognition method (AR).

Figure 1 : Block diagram of experiments.

recognition performance, but only to assess the effect of the vocoders in the recognition performance.

4. DATABASE AND RESULTS

Our experiments have been performed on a reference database of 25 Dutch speakers (16 males and 9 females). Over a period of one month, each speaker recorded 2 sessions of 30 sentences. The speech acquired through a microphone is sampled at 8 kHz, and linearly coded at 16 bits/sample.

To train the model of a speaker, the first twenty concatenated sentences of each session are used. Therefore, the training duration for each speaker is about 40 seconds. For the test, 3 successive sentences are concatenated out of the last ten sentences of each session (8 verification tests per session and per speaker). The test duration is about 8 seconds.

In the first experiment we applied the standard speaker recognition method on the coded/decoded signals (on which both enrolment and verification tests were performed). The identification error rates (IER) are presented in Table 2. To achieve the verification tests, the improved Decision Logic (DL), suggested in [4], is performed. The ROC curves for the different coders are shown in Figure 2.

	IER (%)
Unprocessed Speech	0.5
LPC10 (2.4 kbits/s)	8.0
CELP (4.8 kbits/s)	7.5
CELP (8 kbits/s)	0.5
LD-CELP (16 kbits/s)	0.5

Table 2 : Identification Error Rates with the different coders

In comparison with the quality scores obtained by the different subjective tests, we can suggest that the ROC curves or the IER with less accuracy be used as objective tests. In the future, the objective will be to find the relationship between the subjective tests and the results obtained by speaker recognition methods. This relationship will be accompanied by a training/test protocol and a specific speakers database to define.

For the second experiment, the reference database is coded by the LPC10 and CELP (8 kbits/s). From respectively the R and the LSP coefficients transmitted in the coded frames, we compute the 12 cepstral coefficients on which both enrolment and verification tests were performed. The training/test protocol is the same as the one used in the first experiment. The identification error rates, obtained in the first and second experiment, are compared in Table 3.

	Experiment N°1	Experiment N°2
LPC10 (2.4 kbits/s)	8.0 %	8.75 %
CELP(8 kbits/s)	0.5 %	1 %

Table 3 : Identification Error Rates in the first and second experiment

The ROC Curves, obtained in the first and second experiment, are also compared in Figure 3. With respect to table 3 and Figure 3, we can observe that the implementation of the speaker recognition on a vocoder link, at the level of the transmitted coded frames, is not straightforward. Some observations and explanations can be given :

- With the LPC-10, the recognition in the second experiment is very degraded. Indeed, the transmitted parameters by the LPC-10 are the PARCOR coefficients. In comparison with the LSP coefficients

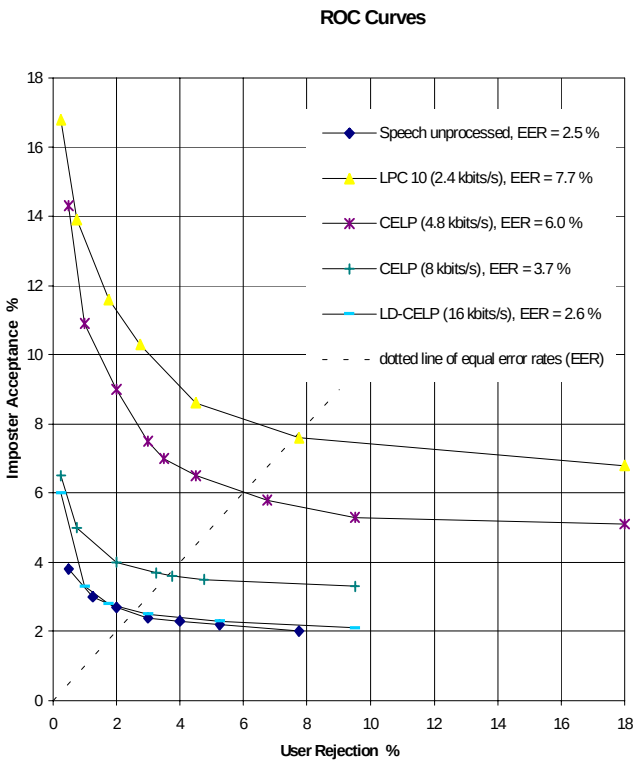


Figure 2 : ROC Curves for different vocoders. The dotted line shows the points of equal error (false acceptance/false rejection).

transmitted by the CELP vocoder, we know that the PARCOR coefficients do not have good quantization properties [5]. We can see that the speaker recognition is better when using the CELP vocoder than when using the LPC-10 vocoder.

- The speaker recognition scores are better when we use the 12 cepstral coefficients computed from the decoded speech than when we use the 12 cepstral coefficients picked up from the coded frames. A possible explanation comes from the "analysis-by-synthesis" concept of vocoders. In the coded frames, some important information about the speech and the speaker is transmitted in other parameters than the 12 cepstral coefficients, such as the information about the excitation (voiced/unvoiced, pitch, gain-shape). Contrary to the second experiment where we pick up only 12 cepstral coefficients, the synthesis of the signal (decoder) uses all these parameters with all their dependencies. To implement the speaker recognition at the level of the coded frames, the development of a new speaker recognition method using all the transmitted parameters must therefore be investigated.
- Comparing the speaker recognition scores obtained on the unprocessed speech and those with the decoded speech by the vocoders, we can see that the recent vocoders (CELP at 8 kbits/s, LD-CELP at 16 kbits/s) provide a high enough quality to achieve the implementation of the speaker recognition system on a vocoder link in real-world conditions.

5. CONCLUSION

In this paper, we have applied a speaker recognition method, using the Second-Order Statistical Measures, on several vocoders. From the experiments, we have presented two different uses of a speaker recognition method in application to a vocoder link.

If we consider the vocoder system as a black box, the use of a speaker recognition method could give an objective measurement of the quality of this system. In this case, the speaker recognition could be used to validate and to choose one vocoder amongst the others.

When we use the speaker recognition method to provide an additional functionality to the system, by applying it directly to the coded frames, the recognition scores are function of the considered vocoder. The recent vocoders have a high enough quality to carry out the implementation of the speaker recognition in real-world conditions. The speaker recognition performances will probably improve if we investigate and use a new speaker recognition taking into account all the transmitted parameters by the vocoder.

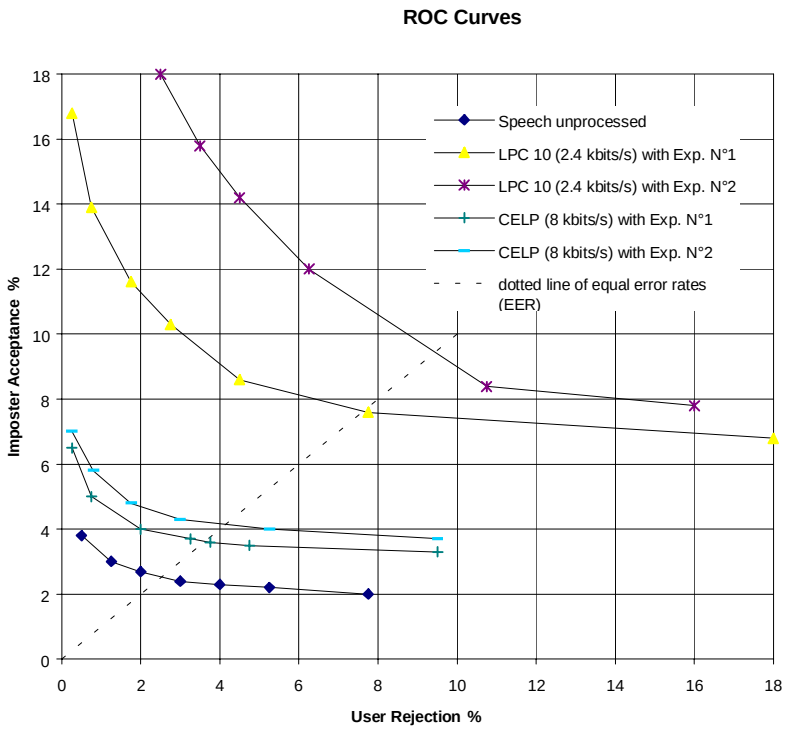


Figure 2 : ROC Curves between the first and the second experiment on 2 different vocoders (LPC-10, CELP 8kbits/s). The dotted line shows the points of equal error (false acceptance/false rejection).

ACKNOWLEDGEMENTS

The authors acknowledge Hervé Bourlard, Olivier van der Vrecken, Laurent Hubaut and all the members of SIC Lab for their helpful comments and feedback.

REFERENCES

[1] S. Spanias, “Speech Coding : A Tutorial Review”, Proc. IEEE, vol. 82, pp 1541-1582, Oct. 1994.

[2] M. R. Schroeder, and B. Atal, “Code-Excited Linear Prediction (CELP) : High Quality Speech at Very Low Bit Rates”, in Proc. IEEE ICASSP, pp. 937, Apr. 1985.

[3] F. Bimbot, I. Magrin-Chagnolleau, L. Mathan, “Second-order statistical measures for text-independent speaker identification”, in Speech Com., vol 17, pp. 177-195, Aug. 1995.

[4] A.E. Rosenberg, “Evaluation of an automatic speaker verification system over telephone lines”, Bell Syst. Tech. J., vol 55, No 6, pp 723-744, 1976.

[5] N. Sugamura and F. Itakura, “Speech analysis and synthesis methods developed at ECL in NTT -From LPC to LSP-, Speech Communication 5, North-Holland, pp 199-215, 1986.