ARCHIVES OF ACOUSTICS
32, 4 (Supplement), 187–192 (2007)

TESTS OF ROBUSTNESS OF GMM SPEAKER VERIFICATION IN VoIP TELEPHONY

Piotr STARONIEWICZ

Wrocław University of Technology
Institute of Telecommunications, Teleinformatics and Acoustics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: piotr.staroniewicz@pwr.wroc.pl

(*received July 15, 2007; accepted November 7, 2007*)

The paper presents the scores of the GMM (Gaussian Mixture Models) based speaker verification system for speech signal transmitted in VoIP (Voice over IP) telephony conditions. The speaker verification problem was partly solved over traditional PSTN networks (Public Switched Telephone Network), however nowadays it is also important to assess how specific distortions of VoIP transmission influence the speaker verification scores. As a reference database XM2VTS (Extended multi Modal Verification for Teleservices and Security appliations Data Base) containing English speech (strings of digits) was applied. Three coder degradations (PCM, G.711A and G.723.1) and three network conditions were examined in various configurations to estimate the influence of each, coding and transmission degradation for the final verification scores.

Keywords: speaker verification, VoIP.

1. Introduction

The influence of the VoIP transmission on speaker identification was partly examined in the author's previous works [6, 8]. The specific recognition task addressed in commercial systems is rather a verification than identification. Most simple speaker verification applications are text-dependent or text-constrained systems. These are quite convenient solutions providing there are cooperative users pronouncing a fixed password or prompted phrases from a small vocabulary. A more flexible, text-independent systems do not demand speaker cooperation but require more sophisticated algorithms applied in the recognition process. Text-independent speaker verification system was examined in the VoIP conditions and the influence of the individual Internet transmission factor on the system scores is presented in the paper.

2. Speaker verification system

A classical speakers verification system is composed of two phases, a training and a testing one (Fig. 1) [2]. The first step of both the training and the testing is the speech

feature extraction process [2, 8]. In the front-end procedures of the applied verification system standard speech parametrization methods were applied, namely, pre-emphasis, windowing, extraction of one of the cepstral coefficients vectors: MFCC (Mel Frequency Cepstral Coefficients), LPCC (Linear Prediction Cepstral Coefficients) or UFCC (Uniform Frequency Cepstral Coefficients). Calculated cepstral coefficients can be then centered, which is realized by subtraction of the cepstral mean vector (CMS), lowering the contribution of slowly varying convolutive noises. Afterwards the dynamic information was incorporated in the feature vectors by using Δ and $\Delta\Delta$ parameters, which are polynomial approximations of the first and the second derivates.

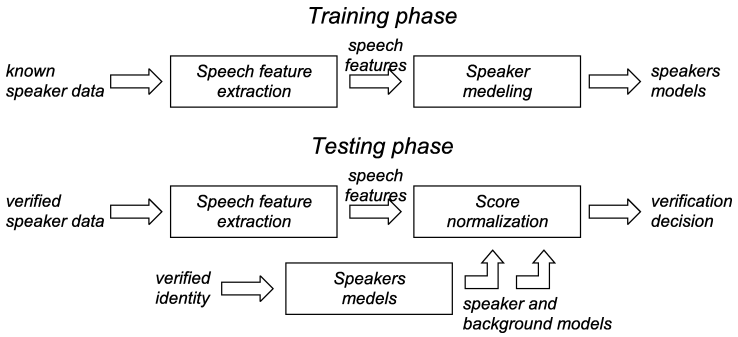

Fig. 1. The scheme of speaker verification system.

The second step, the statistical modeling (Fig. 1), was done with GMM (Gaussian Mixture Models), nowadays the most successful likelihood function [2, 5, 8]. During the presented experiments the Alize v.1.1 software platform was applied. The final step of the speaker verification process is the decision which consists of comparing the likelihood resulting from the comparison between the claimed speaker model and the incoming speech signal with a decision threshold. The claimed speaker is accepted if the likelihood is higher than the threshold, otherwise it is rejected. The tuning of the decision threshold is a troublesome problem in speaker verification because of score variability between trials (differences in contents of speech material, duration between speakers, variation in a speaker's voice caused by emotional state etc., acoustical conditions). To avoid the above problems, score normalization techniques have been introduced. Three normalization techniques have been tested: Tnorm (Test-normalization), Znorm (Zero normalization) and ZTnorm (the combination of Znorm and Tnorm).

3. Database

The XM2VTSDB (Extended Multi Modal Verification for Teleservices and Security applications Data Base) database was used for tests of the speaker verification system. The database was created within the framework of the EU ACTS (Advanced Communications Technology and Services) program. It contains the recordings of 295 voices (men and women) where each speaker uttered eight digit strings repetitions in English:

“0,1,2,3,4,5,6,7,8,9” and “5,0,6,9,2,8,1,3,7,4” and the sentence: “Joe took father’s green shoe bench out". The mean time of a single utterance is about six seconds which gives altogether about two and a half minutes of speech for each speaker. The recordings were done in PCM format with 32 kHz sampling frequency and 16 bit resolution in acoustically good conditions. The database was exposed to nine types of degradation to simulate various conditions of VoIP transmission. In the degradation process signal was down-sampled to 8 kHz, next encoded with G.711 or G.723.1 codecs and finally treated with the packet loss process according to chosen IP conditions. The packet loss simulation was done with the two-state Gilbert model [1, 7, 8]. Three transmission conditions: ideal (non-loss), average ($p = 0.1$ and $q = 0.7$) and poor ($p = 0.25$ and $q = 0.4$). The speakers were divided into clients (200 speakers), impostors whose voices were used to create a background-noise model as well as for results normalization (40 speakers) and impostors used in the testing phase (55 speakers).

4. Evaluation functions

In the speaker verification system two basic types of errors occur, namely, FAR (False Acceptance Rate) and FRR (False Rejection Rate). A false acceptance error occurs when an identity claim from an impostor is accepted, whereas a false rejection error occurs when a valid identity claim is rejected. Both FAR and FRR depend on the threshold value which is set in the verification decision process. Such a system has many operating points so a single performance number is usually inadequate to represent the capabilities of the system. The ROC curve (Receiver Operating Characteristic) (where the false rejection rate is plotted on the horizontal axis, whereas the correct detection rate is plotted on the vertical) has been used traditionally to present the performance of the speaker verification system, nowadays a variant of this which is called DET curve (Detection Error Tradeoff) proposed by NIST (National Institute of Standards and Technology) [4] is more useful. The DET curve represents the system performance as a FAR in the function of FRR which is monotonous and decreasing and is usually plotted on a normal deviate scale. In a speaker recognition system with true speakers and impostors, whose scores are Gaussians with the same variance, the result will be depicted as a linear curve with a slope equal to -1. The better the system is, the closer to the origin the curve will be. Practically the score distributions are close to Gaussians and therefore easily readable and comparable for various conditions in which the system works. The EER (Equal Error Rate) measure is sometimes used to summarize the performance of the system in a single figure. It corresponds to the operating point where the FAR is equal to the FRR.

5. Results and discussion

In the preliminary experiments the optimal parameters of the speaker verification system were settled to provide the highest effectiveness on the one hand and not to lengthen the computation time excessively on the other. In most cases the best results were obtained for the following parameters (which were applied in the main experi-

ments): pre-emphasis with the factor of 0.95; windowing with 20 ms Hamming windows and 10 ms overlap; FFT of 512 samples and 35 channels filter-bank; 108-dimensional feature vector (35 UFCC coefficients, log of frame spectrum energy and corresponding to them Δ and $\Delta\Delta$ coefficients) which was centered (CMS); UBM (Universal Background Model), individual speakers models consisting of 185 components; score normalization with Tnorm with the cohort of 39 best impostors. The experiments results are presented in DET curves (Figs. 2, 3 and 4).

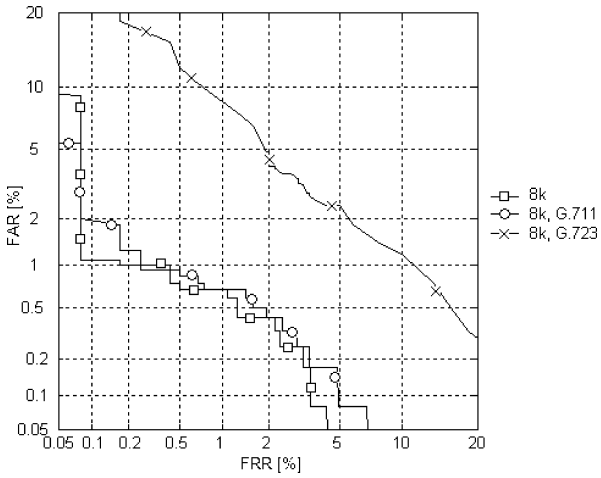

Fig. 2. DET Curves for speaker verification with 8 kHz sampling, G7.11 and G.723 coding, no packet loss and UFCC coefficients.

The EER results are presented in Table 1. Table EER obtained for the UFCC and LPCC parametrizations and various kinds of VoIP speech signal degradation (coding: none, 8 kHz and network conditions: with no packet loss, average and poor). During the speaker verification of no degraded data for both applied parametrization methods (UFCC and LPCC) the 100% accuracy was obtained. The limitation of the speech signal band to 4 kHz badly affected the verification scores (to 0.67% of EER for UFCC parameters). In practice, the G.711 encoding did not cause further slope of the verification accuracy. The applying of the G.723.1 codec crucially lowers the system efficiency (EER of up to 9% for poor network conditions). Poorer results for the G.723.1 than for the G.711 codec, besides a bigger signal compression (and what follows, bigger distortions), are caused by the fact that the CELP (Code Excited Linear Predictive) codecs, which are based on the acoustic model of the vocal tract during the speech production, are focused rather on copying information of the content of the speech than on the individual biometric features of the speaker. Besides the speech coding, the second important aspect of VoIP transmission is the packet loss. Its influence on verification scores is at the level similar to coding. The average network conditions insignificantly lower the verification result, whereas in the poor network conditions the error rates rose considerably.

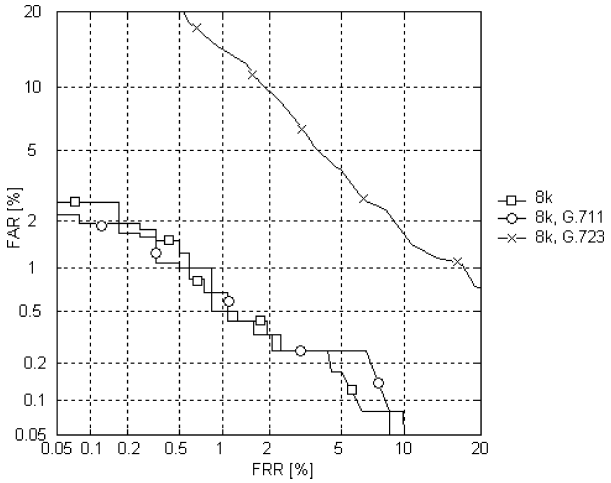

Fig. 3. DET Curves for speaker verification with 8 kHz sampling, G7.11 and G.723 coding, average network conditions and UFCC coefficients.

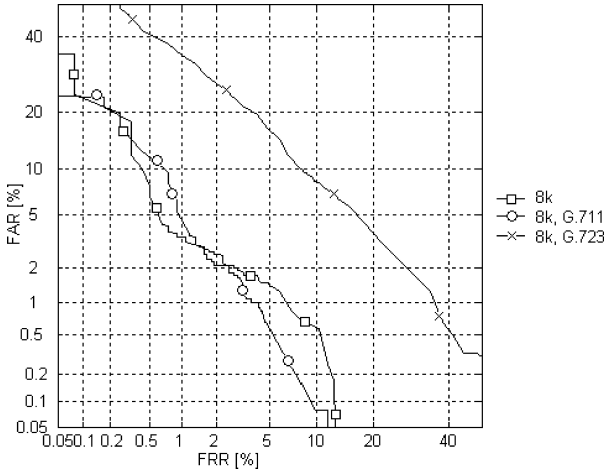

Fig. 4. DET Curves for speaker verification with 8 kHz sampling, G7.11 and G.723 coding, poor network conditions and UFCC coefficients.

Table 1. EER speaker verification scores for no degradation case, three coding types (8 kHz sampling, G.711 and G.723), three network conditions (no packet loss, average and poor) and two coefficients (UFCC and LPCC).

	None none	8k none	8k aver.	8k poor	8k G.711 none	8k G.711 aver.	8k G.711 poor	8k G.723 none	8k G.723 aver.	8k G.723 poor
UFCC	0.00	0.67	0.75	2.25	0.75	0.83	2.08	3.17	4.33	9.00
LPCC	0.00	0.75	1.50	2.96	0.83	0.75	2.08	2.92	4.33	8.50

The packet loss effect brings in gaps in the speech signal, which is the cause of reducing the number of frames that can be used in the verification process and disturbs the time structure of the signal which can be important when using Δ and $\Delta\Delta$ coefficients. Simultaneous degradation by the speech coding and network distortions caused the biggest verification errors.

6. Conclusions

The main aim of the carried out experiments was the examination of the influence of VoIP speech transmission distortions on speaker verification scores. The tested GMM-based speaker verification system use up-to-date speech processing procedures and has an efficiency comparable with nowadays professional applications used commercially. The applied speaker verification assessment methods with DET curves and EER revealed a big universality and make it possible to present performance results, in which tradeoffs of two error types are involved. The obtained results indicate that in a case of remote speaker verification through the IP network it can be an error-caused case for poor network conditions and a lower band transmission coding (important if the cost of error is very high, i.e. bank transactions, trade secrets data, etc. but acceptable for other applications i.e. games, educational etc.). The packet loss problem can be avoided by an improvement of the IP transmission (to introduce suitable QoS parameters). The other way of speaker verification improvement would be a proposal of other than G.723.1 codec for low band speech transmission, which is more focused on transmitting voice biometrics attributes.

References

[1] BESACIER L., MAYORGA P., BONASTRE J. F., FREDOUILLE C., *Methodology for evaluating speaker robustness over IP networks*, Proc. of a COST 275 Workshop The Advent of Biometrics on the Internet, pp. 43–46, Rome, Italy 2002.

[2] BIMBOT F., *et al.*, *A tutorial on text-independent speaker verification*, EURASIP Journal on Applied Signal Processing, **4**, 430–451 (2004).

[3] EVANS N., MASON J., AUCKENTHALER R., STAMPER R., *Assesment of speaker verification degradation due to packet loss in context of wireless devices*, Proc. of a COST 275 Workshop The Advent of Biometrics on the Internet, pp. 43–46, Rome, Italy 2002.

[4] MARTIN A., DODDINGTON G., KAMM T., ORDOWSKI M., PRZYBOCKI M., *The DET curve in assessment of detection task performance*, EuroSpeech 1997, **4**, 1895–1898 (1997).

[5] REYNOLDS D. A., QUATIERI T. F., DUNN R. B., *Speaker verification using adapted gaussian mixture models*, Digital Signal Processing, **10**, 19–41 (2000).

[6] STARONIEWICZ P., *Influence of specific VoIP transmission conditions on speaker recognition problem*, Archives of Acoustics, **31**, 4S, 197–203 (2006).

[7] STARONIEWICZ P., MAJEWSKI W., *Methodology of speaker recognition tests in semi-real VoIP conditions*, Proc. of 3rd Cost275 Workshop in Hatfield, pp. 33–36, UK 2006.

[8] STARONIEWICZ P., *Speaker recognition for VoIP transmission using Gaussian Mixture Models*, Computer Recogition Systems, pp. 739-745, Springer-Verlag, Berlin Heidelberg 2005.