The effect of speech and audio compression on speech recognition performance

— **Source link** ↗

Laurent Besacier, Carole Bergamini, Dominique Vaufreydaz, Eric Castelli

Related papers:

- An Enhanced Full-Rate Speech Coder for GSM and PCS1900

- An efficient transcoding algorithm for G.723.1 and EVRC speech coders

- Method of Converting Speech Codec Formats between GSM 06.20 and G. 729

- Low bit rate speech coding for single speakers

- A Qualified ITU-T G.729EV Codec Candidate for Hierarchical Speech and Audio Coding

THE EFFECT OF SPEECH AND AUDIO COMPRESSION ON SPEECH RECOGNITION PERFORMANCE

Laurent Besacier, Carole Bergamini, Dominique Vaufreydaz, Eric Castelli

THE EFFECT OF SPEECH AND AUDIO COMPRESSION ON SPEECH RECOGNITION PERFORMANCE

L. Besacier, C. Bergamini, D. Vaufreydaz, E. Castelli

Laboratoire CLIPS-IMAG, équipe GEOD, Université Joseph Fourier,
B.P. 53, 38041 Grenoble cedex 9, France
Phone : (+33) 04 76 63 56 95 ; fax : (+33) 04.76.63.55.52
Laurent.Besacier@imag.fr

Abstract - This paper proposes an in-depth look at the influence of different speech and audio codecs on the performance of our continuous speech recognition engine. GSM full rate, G711, G723.1 and MPEG coders are investigated. It is shown that MPEG transcoding degrades the speech recognition performance for low bitrates whereas performance remains acceptable for specialized speech coders like GSM or G711. A new strategy is proposed to cope with degradation due to low bitrate coding. The acoustic models of the speech recognition system are trained with transcoded speech (one acoustic model for each speech / audio codec). First results show that this strategy allows to recover acceptable performance.

1. INTRODUCTION

Speech recognition technology tends to be more and more embedded in mobile phones or other communication terminals. If keyword recognition for basic commands (isolated speech recognition for small vocabulary) can be performed locally on the terminal with DSP processors, it is generally not the case for large vocabulary continuous speech recognition which implies more complex treatments and the use of huge language models (for instance, a 60000 words trigram language model requires about 300 Mbytes of RAM). In this case, a distant speech recognition server is accessed by the terminal which needs to use continuous speech recognition technology. Different strategies can be proposed to access a distant speech recognition server:

-Recently, some solutions were proposed where speech analysis (the extraction of acoustic features for speech recognition, which is a low-cost task) is performed on the client terminal and the acoustic parameters are directly transmitted to the distant speech recognition server. The expression *distributed speech recognition[1]* is generally used to define this type of architecture [4,7]. In this kind of architecture, the terminal needs to be able to extract the adequate acoustic parameters needed a the input of the speech recognition server.

-Sometimes, one may want to access to a speech recognition service from a very basic terminal which is just able to capture and transmit speech. In this case, the speech signal is transmitted from the client terminal to a distant speech recognition server. Coding of the speech signal is generally necessary to reduce transmission delays (GSM, MPEG or G7XX coding, depending on the client terminal used). *Figure 1* illustrates this client/server architecture. The problem with this architecture is that transcoding (the process of coding and decoding) modifies the spectral characteristics of the speech signal, so it is likely to have an influence on speech recognition performance.

---

[1] http://www.icp.inpg.fr/ELRA/aurora.html

Thus, this paper first proposes an in-depth look at the influence of different speech and audio codecs on the performance of our French continuous speech recognition engine. An new strategy is then proposed to cope with speech degradation due to speech coding. The solution we propose is to train the acoustic models of the speech recognition system with transcoded speech (one acoustic model for each speech/audio codec).
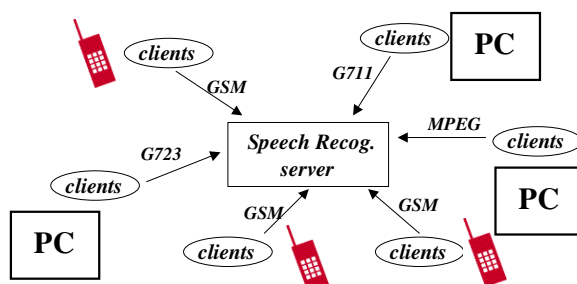

Figure 1 : speech recognition server accessed by different client terminals

2. AUDIO CODECS TESTED

Different human-machine interfaces use speech recognition technology. For instance, vocal servers (used to obtain information via the telephone) are more and more developed. Nowadays, access to a vocal server is not only made through the conventional telephone network, but voice can also be transmitted through wireless networks (with mobile phones or mobile devices) or through IP networks (with H323 videoconferencing software for example). As a consequence, we decided to study the following coders:

- *GSM (Global System for Mobile Communications) coder :*

GSM is the pan-European cellular mobile standard. Three speech coding algorithms are part of this standard. The purpose of these coders is to compress the speech signal before its transmission, reducing the number of bits needed in its digital representation, while keeping an acceptable quality of the decoded output. There exist three different GSM speech coders, which are referred to as the *full rate*, *half rate* and *enhanced full rate* coders. Their corresponding European telecommunications standards[2] are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively. These coders work on a 13 bit uniform PCM speech input signal, sampled at 8 kHz. The input is processed on a frame-by-frame basis, with a frame size of 20 ms (160 samples). We used in our experiments the FR coder which was standardized in 1987. This coder belongs to the class of Regular Pulse Excitation - Long Term Prediction - linear predictive (RPE-LTP) coders. In the encoder part, a frame of 160 speech samples is encoded as a block of 260 bits, leading to a bit rate of 13 kbits/s. The decoder maps the encoded blocks of 260 bits to output blocks of 160 reconstructed speech samples. The GSM full rate channel supports 22.8 kbits/s. Thus, the remaining 9.8 kbits/s are used for error protection. The FR coder is described in GSM 06.10 down to the bit level, enabling its verification by means of a set of digital test sequences which are also given in GSM 06.10. A public domain bit exact C-code implementation of this coder is available[3].

---

[2] http://www.etsi.fr
[3] http://kbs.cs.tu-berlin.de/~jutta/toast.html

- G711 and G723.1 coders

H323 is a standard for transmitting voice and video. A famous H323 videoconferencing software is for example NetMeeting™. H323 is commonly used to transmit video and voice over IP networks. The audio codecs used in this standard are G711, G722, G723, G728 and G729. We used in our experiments the codec which has the lowest bit rate : G723.1 (6.4 and 5.3 kbits/s), and the one with the highest bitrate : G711 (64 kbits/s : 8 kHz, 8 bits). Experiments on VoIP speech were made in the context of the NESPOLE![4] European Project which is a common EU NSF funded project exploring future applications of automatic speech to speech translation in e-commerce and e-service sectors [3]. The scenario for the first showcase of NESPOLE! involves an Italian speaking agent, located in a tourism agency in Italy and a client located anywhere (English, German or French speaking) using a simple terminal (PC, sound and video cards, H323 videoconferencing software like NetMeeting). Thus, the French speech recognition module, which is an essential part of the complete speech to speech translation chain, needs to cope with VoIP speech.

- MPEG audio coders

We also made some experiments with MPEG coders. Unlike GSM and G7XX which are specific speech coders, MPEG coders allow to compress any audio signal. In fact, MPEG audio coding is generally not use for transmission of speech data but for compression of audiovisual data (TV programs for instance). Another application of speech recognition is the transcription of broadcast news and TV programs or films for archiving and retrieval. It is thus interesting to test the influence of MPEG audio coding algorithms on speech recognition performance. Moreover, MPEG audio coding supports a variable bitrate, which allows us to test speech recognition on more and more compressed speech. For the experiments on MPEG transcoded speech, we used a PCX11+ specialized board (from *Digigram* society) for layers 1,2 and 3 of MPEG I and for different bit rates.

3. SPEECH RECOGNITION SYSTEM

Our continuous French speech recognition system RAPHAEL uses Janus-III toolkit [6] from CMU. The context dependent acoustic model was learned on a corpus which contains 12 hours of continuous speech of 72 speakers extracted from Bref80 [2] corpus. The first vocabulary (5k words) contains nearly 5500 phonetic variants of 2900 distinct words ; it is specific to reservation and tourist information domain. The second vocabulary (28k words) is more general. The trigram language model that we used for our experimentation was computed using Internet documents because it was shown that they give very large amount of training data for spoken language modelling [5].

4. RECOGNITION EXPERIMENTS ON CODED SPEECH

We conducted a series of recognition experiments with either 120 recorded sentences focused on reservation and tourist information task (CSTAR120 database) or with 300 sentences read from a French newspaper (AUPELF database). We transcoded (coded and then decoded) both test databases with the coders described

---

[4] http://nespole.itc.it/

in *section 2* and we calculated the Word Accuracy of our recognition system with an acoustic model learned either on BREF 80 database (16kHz sampling rate) or on its downsampled version (8kHz). In fact, we had to learn a "8kHz acoustic model" because GSM and G7XX coders need 8kHz input signals. It was thus unrealistic to test the effect of GSM and G7XX coding algorithms on 16kHz sampled signals.

- Effect of GSM coder

Training database	Coder for test	Word Accuracy
Bref80 – 16kHz	None (16kHz sig.)	92.3 %
Bref80 – 8kHz	None (8kHz sig.)	91.6%
Bref80 – 8kHz	GSM FR	91,6 %

Table 1 : effect of GSM coding on speech recognition performance
on CSTAR120 DB with 5k words vocabulary

Training database	Coder for test	Word Accuracy
Bref80 – 16kHz	None (16kHz sig.)	61.2 %
Bref80 – 8kHz	None (8kHz sig.)	49.7%
Bref80 – 8kHz	GSM FR	46.3 %

Table 2 : effect of GSM coding on speech recognition performance
on AUPELF DB with 28k words vocabulary

From the results given in *Table 1* we could conclude that GSM transcoding has no influence on the performance of our speech recogniser. However, it was observed in a former paper [1] that GSM speech coding has influence on speaker recognition performance (the task of recognising speakers by their voice). Both results can be explained: speaker recognition technology uses "purely" acoustic models which can be degraded by GSM transcoding, whereas speech recognition technology uses acoustic models but also language models. Thus, degradation of the speech recogniser acoustic model can be recovered by the language model, especially if the language model is well adapted to the test data (which was actually the case for results presented in *Table 1*). From this experiment, we could also notice that downsampling from 16 kHz to 8 kHz does not introduce important performance degradation. However, the other experiment on AUPELF transcoded database (*Table 2*), with a more "general" vocabulary (28k words) and consequently with a less "constrained" language model, shows that in this case, the bandlimiting resulting from the downsampling to 8 kHz degrades the performance. We also observe in *Table 2* that GSM transcoding has a slight negative influence on the performance. So, we can say that with a less constrained language model, the degradation of the acoustic model is not completely recovered by the language model ; but this degradation is mainly due to the bandlimiting and not to the GSM transcoding.

- Effect of G711 and G723.1 coders

In this experiment, the test speech material was either transcoded with G711 coder or with G723 coder.

Training database	Coder for test	Word Accuracy
Bref80 – 8kHz	G711	91.9 %
Bref80 – 8kHz	G723	91.2 %

Table 3 : effect of G711 and G723.1 coding on speech recognition performance
on CSTAR120 DB with 5k words vocabulary

The results of *Table 3* show that G711 and G723 transcoding do not degrade the speech recognition performance. Moreover, we do not see much difference between G711 and G723 performance whereas G723 is a very low bitrate coder (5.3 kbits/s) compared to G711 coder (64 kbits/s). We must however admit that only the influence of speech codecs was evaluated here, whereas packet loss during transmission, which is certainly the biggest source of degradation, is not investigated. Experiment on AUPELF database (with a less "constrained" language model) transcoded with G723.1 confirms the law degradation due to G723.1 transcoding since performance goes from 41.2% accuracy for 8khz none coded signals to 41.1% accuracy for G723.1 transcoded signals.

- Effect of MPEG audio coders

Training database	Coder for test	Word Accuracy
Bref80 – 16kHz	None (16kHz sig.)	92.3 %
Bref80 – 16kHz	MPEG Lay3 64kbits/s	92.2 %
Bref80 – 16kHz	MPEG Lay3 32kbits/s	92,1 %
Bref80 – 16kHz	MPEG Lay3 24kbits/s	91,6 %
Bref80 – 16kHz	MPEG Lay3 16kbits/s	85,5 %
Bref80 – 16kHz	MPEG Lay3 8kbits/s	33,8 %
Bref80 – 16kHz	MPEG Lay2 64kbits/s	92,5 %
Bref80 – 16kHz	MPEG Lay2 32kbits/s	92,3 %
Bref80 – 16kHz	MPEG Lay2 24kbits/s	70,6 %
Bref80 – 16kHz	MPEG Lay2 16kbits/s	58,3 %
Bref80 – 16kHz	MPEG Lay2 8kbits/s	6.2 %
Bref80 – 16kHz	MPEG Lay1 32kbits/s	73.0 %

Table 4 : effect of MPEG coding on speech recognition performance
on CSTAR120 DB with 5k words vocabulary

Results of *Table 4* show that above 32 kbits/s bitrate, no significant degradation of speech recognition performance is observed, whereas below this threshold, performance starts to decrease dramatically. Moreover, performance is better for MPEG layer 3 than for MPEG layer 2 which is also better than MPEG layer 1. These results are in correspondence with the perceptual speech quality of the different MPEG layers.

5. ROBUSTNESS ISSUES

In the experiments of *section 4*, acoustic models training was always performed on a "clean" and not transcoded database. We observed that a mismatch between test data (compressed with MPEG) and training data can introduce a significant degradation of performance (*Table 4*). In this section, we intend to cope with this degradation. For this, we propose to reduce the mismatch between test and training data by training speech recognition system acoustic models with transcoded speech. Thus, in this experiment, two acoustic models were trained on BREF80 database transcoded with MPEG layer2 24 kbits/s and 8 kbits/s.

Training database	Coder for test	Word Accuracy
Bref80 – 16kHz	MPEG Layer2 24kbits/s	70,6 %
Bref80 – mpeg layer 2 24kb/s	MPEG Layer2 24kbits/s	**93.4 %**
Bref80 – 16kHz	MPEG Layer2 8kbits/s	6.2 %
Bref80 – mpeg layer 2 8kb/s	MPEG Layer2 8 kbits/s	**64,6 %**

Table 5 : training on the original database versus training on transcoded database

Results of *Table 5* show that this strategy allows to recover acceptable performance since, for instance, word recognition can increase from 70.56% to 93.45% for speech signal transcoded with MPEG, Layer 2, 24 kbits/s and from 6.19% to 64.63% for MPEG, Layer 2, 8 kbits/s. We can also notice that results obtained with 24kbits/s bitrate are even better than those obtained on the original speech. An explanation could be that passing the training database inside the coder   allowed to filter some disturbing information. These encouraging results should allow us to propose a speech recognition system architecture for which one acoustic model for each speech/audio codec is trained. To know which speech coder is used on the user terminal (and thus to know which acoustic model must be used on the server) the information concerning the coder used by the client could be encoded in the data stream transmitted from the client to the speech recognition server.

6. CONCLUSIONS

This paper investigated the influence of different speech and audio codecs on the performance of our continuous speech recognition engine. GSM full rate, G711, G723.1 and MPEG coders were investigated. It was shown that MPEG transcoding degrades the speech recognition performance for low bitrates whereas performance remains acceptable for specialized speech coders like GSM or G711. A new strategy was proposed to cope with degradation due to low bitrate coding. The acoustic models of the speech recognition system were trained with transcoded speech. First results have shown that this strategy allows to recover acceptable performance.

As a perspective, a filtering strategy could be proposed : instead of degrading the acoustic models to reduce mismatch between training and test data, we could keep the acoustic model trained on clean speech and learn a linear transformation f between "clean" and "degraded" signals. During the recognition stage, the inverse transformation f^{-1} would be applied to the degraded test signal to reduce mismatch between training and test. We also intend to study new coders like G729 and more recent MPEG standards.

References

[1] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge and F. Pellandini, "GSM Speech Coding and Speaker Recognition" *ICASSP 2000*, Istambul, Turkey, 5-9 june, 2000 .

[2] L. F. Lamel, J. L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Coprus for French", *Eurospeech*, Gênes, Italy, Vol 2, pp. 505-508, 24-26 september 1991.

[3] Lazzari G., "Spoken translation : challenges and opportunities", *ICSLP'2000*, Beijing, China.

[4] D. Pearce, Motorola Labs, "An Overview of ETSI Standards Activities for Distributed Speech Recognition Front-Ends", *AVIOS 2000 : The Speech Applications Conference*, May 22-24, 2000 San Jose, CA, USA.

[5] Vaufreydaz, D., Akbar, M., Rouillard, J., "Internet documents : a rich source for spoken language modeling", *ASRU'99 Workshop*, Keystone Colorado (USA), pp. 277-280.

[6] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "Recent Advances in JANUS : A Speech Translation System". *Eurospeech*, 1993, volume 2, pages 1295-1298.

[7] W. Zhang, L. He, Y. Chow, R. Yang, and Y Su, "The Study on Distributed Speech Recognition System", *ICASSP 2000 International Conference en Acoustic Speech & Signal Processing*. June 5-9 2000, Istambul, Turkey.