AUTOMATIC TRANSCRIPTION OF COMPRESSED BROADCAST AUDIO

Claude Barras, Lori Lamel and Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{barras,lamel,gauvain}@limsi.fr

ABSTRACT

With increasing volumes of audio and video data broadcast over the web, it is of interest to assess the performance of state-of-the-art automatic transcription systems on compressed audio data for media indexation applications. In this paper the performance of the LIMSI 10x French broadcast news transcription system is measured on a two-hour audio set for a range of MP3 and RealAudio codecs at various bitrates and the GSM codec used for European cellular phone communications. The word error rates are compared with those obtained on high quality PCM recordings prior to compression. For a 6.5 kbps audio bit rate (the most commonly used on the web), word error rates under 40% can be achieved, which makes automatic media monitoring systems over the web a realistic task.

1. INTRODUCTION

In the past few years, major advances in speech recognition technology have allowed new ranges of applications, for example broadcast news indexation systems. In the framework of the recent Spoken Data Retrieval task conducted by NIST, 500 hours of radio and TV news were automatically transcribed and indexed [5]. The word error rates reported on this data are about 20% illustrating the ability of todays state-of-the-art systems to deal with such data. However, it should be noted that these data consist of good quality recordings of the audio sources, typically sampled at 16kHz with a resolution of 16 bits per sample, for a total bit rate of 256 kbps. With the development of internet, there is a growing amount of audio and video data broadcast over the web. Some data are re-broadcasts of traditional radio and TV programs, but other programs are available only on the web. In both cases, bandwidth limitations of individual internet connections, with standard modem connections being generally under 56 kbps, only allow for broadcasting of highly compressed programs. The bit rate devoted to audio encoding is even lower when the program contains video. Storage of archives is also a concern, and compression makes it less costly. LIMSI is involved in the European ALERT project[1] which aims to associate state-of-the-art speech recognition with audio and video segmentation and automatic topic indexation to develop an automatic media monitoring demonstrator and evaluate it in real world applications. In this context it is necessary to deal with data in the format used by the service providers for broadcast and storage. In fact, some indexation systems of web audio sources already exist, such as the the SpeechBot system developed by Compaq[2].

In light of these considerations it is necessary to assess the performance of todays systems on compressed audio data. We chose to test the effects of some of the most widespread compressed audio formats on the Web, i.e. Real Audio format[3] and MP3 format[4] [5], on automatic transcription of French broadcast news. We wanted to know what level of compression can be achieved without a significant decrease in the transcription quality, and how to improve the system for audio at higher compression rates.

In the next section we give a brief overview of the LIMSI broadcast news transcription system for French, followed by experiments carried out to assess the automatic transcription quality of uncompressed and compressed data at various bit rates. We explore two ways to reduce the difference between PCM and compressed data, either by using bandwidth limited acoustic models, or by training the models on compressed data.

2. AUTOMATIC TRANSCRIPTION SYSTEM

The LIMSI broadcast news automatic transcription system consists of an audio partitioner [7] and a speech recognizer [6]. Combined with an indexation module, it has been used for building the SDR indexation system [8]. The recognition system initially developed for American English has been ported to the French language [1, 2], which is one of the target languages of the ALERT project and was chosen

[1] http://alert.uni-duisburg.de/

[2] http://speechbot.research.compaq.com/

[3] http://www.real.com/

[4] http://www.cselt.it/mpeg

[5] http://www.tnt.uni-hannover.de/project/mpeg/audio/

for the experiments of this paper.

Prior to word recognition, the acoustic signal is partitioned into homogeneous segments, labeling and structuring its acoustic content. Partitioning consists of identifying and removing non-speech segments such as music, and then clustering the speech segments and assigning bandwidth and gender labels to each segment. For the purposes of this work, the main advantage of partitioning is that all segments from the same speaker can be clustered and acoustic model adaptation can be carried out on a per cluster basis, as opposed to on a single segment basis, thus providing more adaptation data. Removing the non-speech segments does not significantly reduce the word error rate, but does considerably reduce the processing time.

For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with each word. The speaker-independent large vocabulary, continuous speech recognizer makes use of n-gram statistics for language modeling and of continuous density HMMs with Gaussian mixtures for acoustic modeling. Word recognition is usually performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The hypotheses are used in cluster-based acoustic model adaptation using the MLLR technique [9] prior to word graph generation, and all subsequent decoding passes. The final hypothesis is generated using a 4-gram language model.

For all the experimental results given in this paper, the following training conditions were used. The acoustic models were trained on about 100 hours of French broadcast news data. The phone models are position-dependent triphones, with about 10,000 tied-states for the largest model set. The state-tying is obtained via a divisive, decision tree based clustering algorithm. A single set of speaker-independent models are used, however different bandwidth are used for some configurations. Fixed language models were obtained by interpolation of n-gram backoff language models trained on different data sets: 22 M words of press services transcripts; 332 M words of *Le Monde* and *Le Monde Diplomatique* newspaper texts; 63 M words from Agence France Press newswire texts; 0.75 M words corresponding to the transcriptions of the acoustic training data. The interpolation coefficients of these LMs were chosen so as to minimize the perplexity on the development data. The 4-gram LM contains 15M bigrams, 15M trigrams and 13M fourgrams.

The recognition lexicon contains 65333 words, and has a lexical coverage of 98.8% on the development data. Each lexical entry is described as a sequence of elementary units, taken from a 33 phone set plus 3 models used for silence, filler words, and breath noises. The lexical pronunciations are initially derived from the Graphon grapheme-to-phoneme converter developed at LIMSI, and manually verified adding alternate and contextual pronunciations.

All the baseline runs were done at 10 times real-time on a Compaq XP1000 500MHz machine with Digital Unix.

3. EFFECTS OF COMPRESSION ON WORD ERROR RATE

In this section we present the experiments carried out to assess the performance of our broadcast news transcription system on uncompressed and compressed audio data.

Experimental Conditions

The test set consists of 12 broadcast news shows from the French television channel M6, for a total of 2 hours of data. These data were recorded at 16kHz in mono, with 16 bits resolution, thus at 256 kbps. The audio data and manual transcriptions were provided by Vecsys, a partner in the ALERT project. These manual transcripts are taken as the reference transcriptions for the test data. A baseline word error rate (WERR) of the French BN transcription system was determined for the complete test set using the standard NIST alignment and scoring programs.

The 2 hours of data were then compressed using different encoders at various bit rates and uncompressed back to a standard PCM with appropriate up- or down-sampling to 16kHz when necessary. The resulting files were processed by the partitioning and transcription system and scored to provide the global word error rate.

A first set of experiment was carried out using the MP3 format, which stands for the MPEG 1/2 layer III standard and is used for example for encoding DVD audio [3] (http://www.tnt.uni-hannover.de/project/mpeg/audio/). It is now a widespread format for encoding high quality audio broadcast over the web, especially music and songs. The MP3 compression was tested at bit rates ranging from 64 kbps to 8 kbps. The bandwidth of the MPEG decoder is reduced from the initial full bandwidth of 8kHz down to 3kHz for the lowest compression rate.

A second set of experiments was done using the RealAudio format (http://www.real.com). Data was compressed using the Voice codec at bit rates ranging from 64 kbps to 5 kbps. For compression rates under 16 kbps, the encoder limits the bandwidth to 4kHz. It should be noted that the Real format is proprietary and that the encoding scheme is not documented, in contrast to the MPEG layer III format. Furthermore, to the best of our knowledge no tool is available for uncompressing a RealAudio file into a standard PCM file, which appears to be a protection mechanism. Uncompressing the data thus involves playing the file and recording it simultaneously, which can only be achieved in real time. However, given the widespread use of this format on the web, especially for video and speech documents at low bit rates, it certainly warrants being evaluated. Since our evaluation scheme considers the compression/decompression as a separate, black box process, these limitations do not affect our experiments.

Finally, we also tested the GSM full rate codec at 13 kbps used in European digital cellular phone communication [4]. Since this is a widely used format, the efficiency of speech recognition over mobile communications is clearly an important issue, albeit with a different perspective than web-based audio indexing. GSM is considered as a reference of low rate speech quality.

Results

Table 1 gives word error rates for the 2 hours of French broadcast news for the original PCM signal and GSM, Real and MP3 codecs at various compression bit rates. The data bandwidth after compression is also provided. For the reference PCM format, the word error rate is 28.3% with a confidence interval of +/-0.6%. This error rate is somewhat higher than reported error rates for the LIMSI American English transcription system. This difference can be explained by several factors: less acoustic and linguistic training data are available for training the French system than for the English one; there are no broadcasts of the M6 news shows in the training data; and each show duration is only 10 minutes so there is only limited data available for unsupervised speaker adaptation. In a production system, we would expect to incrementally adapt the acoustic models to the shows, thus significantly reducing the word error rate.

The MP3 compression at 64 kbps shows no difference in the results with the original PCM. The word error rate increase for Real and MP3 compression at 32 kbps is significant, but limited to about a 4% relative increase. The word error increases quickly for MP3 compression under 32 kbps, with almost 70% errors for 8 kbps. This very high recognition error is consistent with the poor perceived quality when listening to the compressed data. The codec used in RealAudio Voice format gives better results, especially at low bit rates with the error rate at 5 kbps being only slightly higher than MP3 at 16 kbps. Using a GSM codec the word error rate increases to 37.7%, (a relative increase of 33%), which places it in between Real and MP3 at 16 kbps. With the exception of MP3 at 8 kbps, the resulting word error rates are below 50% and can be used for an automatic indexation of the document via its automatic transcription [5].

4. IMPROVING PERFORMANCE ON COMPRESSED AUDIO

The majority of audio currently available on the web are broadcast at very low bit rates, so it is important to improve as much as possible the performance of the speech recognizer under these conditions. One way to reduce the mismatch between the full bandwidth training data and the limited bandwidth test data is by training models on limited bandwidth data. A second approach to reducing the mismatch between training and testing conditions is to train models on data compressed using the same codec as the one used to process the test data. Both of these approaches are under investigation, with some experimental results reported in this section.

Coder	Bit rate (kbps)	Data band-width (kHz)	8 kHz models WERR (%)	Models with reduced bandwidth	
				band-width	WERR (%)
PCM	256	8	28.3	7	28.5
				4	31.5
GSM	13	4	37.7	4	37.1
Real	64	8	28.7	-	-
	32	8	29.1	-	-
	16	8	31.0	-	-
	8.5	4	40.2	4	37.2
	6.5	4	41.8	4	39.1
	5	4	46.4	4	45.3
MP3	64	8	28.4	-	-
	32	7	29.4	7	29.7
	24	6	34.8	6	33.5
	16	4	45.8	4	44.2
	8	3	69.3	4	67.5

Table 1: Word error rate (WERR) on the automatic transcription of two hours of French broadcast news after various compression schemes, for full- and limited bandwidth acoustic models.

Reducing training bandwidth

The acoustic models in the LIMSI French BN system were trained on 8kHz bandwidth data. Three new sets of acoustic models were trained by lowpass filtering the training data during feature analysis at 7, 6 and 4kHz in order to match the bandwidth of the test data. The compressed data were recognized using the set of acoustic models most closely matching their bandwidth. As can be seen in the right part of Table 1, using models with the appropriate bandwidth reduces the word error rate compared to the original full band models. The largest improvement is for the Real compression at 8.5 kbps with a relative decrease in word error of 7%. Smaller improvements were obtained on the GSM and MP3 data. To provide a reference contrast, the PCM data were also processed using bandwidth limited acoustic models at 7kHz and 4kHz. In the latter case, there is a 11% relative decrease in performance which can be attributed to the loss of high frequency information, without any compression.

Training sets of acoustic models at different bandwidths can be done once for the system, and automatically selecting the set of models best matching the test data is an easy task. Figure 1 shows the word error rates obtained for the best fitting set of acoustic models, for each tested combination of coder and bit rate, along with the reference result on the PCM data.

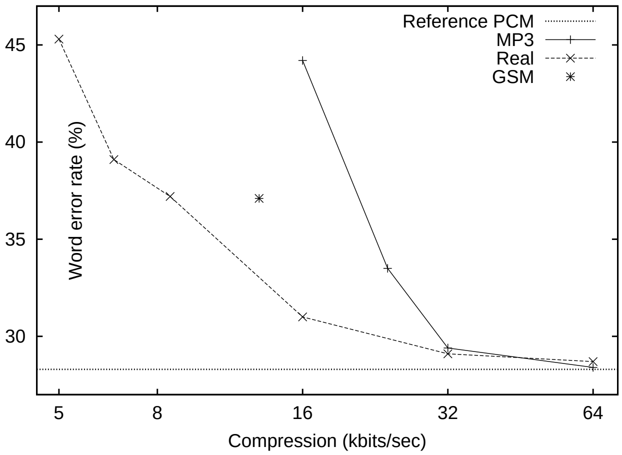

Figure 1: Word error rates for GSM, MP3 and Real compression algorithms on a two-hour set of French broadcast news programs transcribed using matching bandwidth models.

Training on compressed data

Apart from bandwidth limitation, the compression also introduces distortions of the sound. If the acoustic models are trained on data compressed by the same codec as the one expected for the test data it may be possible to obtain better recognition results. Compressing and uncompressing the 100 hours of the training set is a time consuming process, especially for the RealAudio format where it cannot be carried out in less than real time. We chose to process the training data using the MP3 compression format at 16 kbps, Real compression at 6.5 kbps and GSM compression. By doing so, the word error rate for MP3 at 16 kbps is reduced to 42.9%, corresponding to about 3% relative reduction when compared to the 4kHz limited bandwidth case, and total 6% relative reduction compared to the standard full band case. However, the word error rate is 39.7% for Real at 6.5 kbps, and 37.6% for GSM compression, which is better than the results obtained with standard models but worse than the results with bandwidth limited models.

During development of the SpeechBot audio search system at Compaq's Cambridge Research Laboratory, a word error rate of 60.5% was reported on 6.5 kbps RealAudio encoded American English data using standard acoustic models. Using acoustic models trained on RealAudio encoded/decoded data, the word error rate was reduced to 49.6%, thus achieving almost 20% relative improvement [10]. Similarly, we expected better performance under matched training/testing conditions, but using bandwidth limited models proved to be more efficient. In the latter case, the acoustic models were trained on clean speech and adapted to compressed speech during recognition with MLLR technique, which appears to be a better approach than training on distorted speech.

5. CONCLUSIONS

Our experiments on automatic transcription of broadcast news shows indicates that compression to 32 kbps results in almost no performance degradation. The MP3 format was found to be by far the most easy compression scheme to use. At lower bit rates, the RealAudio Voice codec provides the best performance, with a 38% relative increase in word error rate at 6.5 kbps compared to uncompressed audio. Word error rates under 40% were obtained at this compression rate for French broadcast news. This suggests that the development of automatic media monitoring systems over the web is a realistic enterprise.

6. ACKNOWLEDGEMENTS

This work has been partially financed by the European Commission in the framework of the ALERT project and by the French Ministry of Defense. The authors wish to thank our partners in the project, especially Martine Garnier-Rizet from VECSYS, for their support, Gilles Adda for providing the language models used in these experiments, and Sylvia Hermier for carefully checking the reference transcriptions.

REFERENCES

[1] G. Adda, M. Adda-Decker, J.L. Gauvain, L. Lamel, "Text Normalization and Speech Recognition in French", *Proc. ESCA Eurospeech'97*, pp. 2711-2714, Rhodes, Greece, Sep. 1997.

[2] M. Adda-Decker, G. Adda, J.L. Gauvain, L. Lamel, "Large Vocabulary Speech Recognition in French", *Proc. IEEE ICASSP'99*, **I**, pp. 45-48, Phoenix, AZ, March 1999.

[3] K. Brandenburg, "MP3 and AAC explained," *Proc. of the AES 17th International Conference on High Quality Audio Coding*, Florence, Italy, Sept. 1999.

[4] European Telecommunications Standards Institute, European digital cellular telecommunications system (Phase 2); Full rate speech transcoding (GSM 06.10), ETSI Technical Report, Sophia Antipolis, France, 1994.

[5] J.S. Garofolo, C. Auzanne, E. Voorhees, "1999 Trec-8 Spoken Document Retrieval Track Overview and Results," *Proc. 8th Text Retrieval Conference TREC-8*, Gaithersburg, MD, Nov. 1999.

[6] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data", *Proc. ICSLP'00*, Beijing, China, Oct. 2000.

[7] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**, pp. 1335-1338, Dec. 1998.

[8] J.L. Gauvain, L. Lamel, G. Adda, "Transcribing broadcast news for audio and video indexing," Communications of the ACM, 43(2), Feb. 2000.

[9] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

[10] J.-M. Van Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, M. Swain, "SpeechBot: a Speech Recognition based Audio Indexing System for the Web", *Proc. of the 6th RIAO Conference*, Paris, April 2000.