TRANSCRIBING AUDIO-VIDEO ARCHIVES

Claude Barras[1], *Alexandre Allauzen*[1,2], *Lori Lamel*[1], *Jean-Luc Gauvain*[1]

[1] Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France

[2] Institut National de l'Audiovisuel (http://www.ina.fr/)
4 Avenue de l'Europe, 94366 Bry-sur-Marne cedex, France

{barras,allauzen,lamel,gauvain}@limsi.fr

ABSTRACT

This paper addresses the automatic transcription of audiovideo archives using a state-of-the-art broadcast news speech transcription system. A 9-hour corpus spanning the latter half of the 20th century (1945-1995) has been transcribed and an analysis of the transcription quality carried out. In addition to the challenges of transcribing heterogenous broadcast news data, we are faced with changing properties of the archive over time, such as the audio quality, the speaking style, vocabulary items and manner of expression. After assessing the performance of the transcription system, several paths are explored in an attempt to reduce the mismatch between the acoustic and language models and the archived data.

1. INTRODUCTION

Over the last 5 years tremendous progress has been made in speech recognition of broadcast data as a support for random access to relevant portions of audio documents, reducing the time needed to identify recordings in large multimedia databases. This progress has led to automatic systems that can transcribe contemporary broadcast news audio data with word error rates of about 20% in various languages. The transcription quality is good enough to enable a variety of applications such as content-based document retrieval with almost no degradation compared to manual closed-captions, using state-of-the-art indexation and retrieval techniques [1]. There has also been growing interest, both at a national and international level, in preserving and facilitating access to historical archives, and speech recognition is a key technology to deal with processing the audiovisual data from different media sources. Recent progress in information processing has been accompanied with tremendous increases in the computational and storage capacities at reduced cost, leading to a increased interest in preserving and annotating huge audiovisual archives. Many archives are currently on film or video, and there are efforts underway to convert these to digital format for more efficient storage and easier access. These conversion efforts will also enable much wider access to the content, which is often limited in order to minimize degradation to the media.

The objective of the European Community funded ECHO project (http://pc-erato2.iei.pi.cnr.it/echo/) is to provide technological support for digital film archives, and to improve accessibility and searchability of large historical audio visual archives. Archive institutions in France, Italy, the Netherlands and Switzerland are involved in the project. In France, the Institut National de l'Audiovisuel (INA) archives contain about 1.5 million hours of radio and television programs (dating back to 1933 for radio and 1949 for television), and a program for digitizing 500,000 hours of them is underway. The associated manual annotations of these documents provide only general information about the entire document, such as the broadcast date and main theme. Detailed manual annotation of such a large amount of data is beyond the means of the conversion effort. An automatic analysis of the content, based on an automatic transcription, can be used to produce much more detailed information for document retrieval purposes. One of the aims of the ECHO project is to study the use of current speech recognition technology to produce automatic transcriptions of historical documents for the French, Italian and Dutch languages, and LIMSI is in charge of the French transcriptions.

This paper describes our recent research in transcribing French archive data from the latter half of the last century, provided by INA for the ECHO project. The archive documents are different from contemporary broadcast news, which raises several issues for processing them: they are often noisier (specific acoustic conditions), the speaking style has (and is still) evolving, there are epoch-specific proper names that need to be added to the lexicon, and the range of topics is very broad compared to contemporary news. Training the statistical models of the transcription system requires a large amount of acoustic and text data, representative of the documents to be processed and in electronic format. Finding a sufficient amount of data for this adaptation, especially electronic texts about historical periods, is one of the biggest challenges. The acoustic and linguistic models of the contemporary broadcast news transcription are adapted for use in ECHO.

The next two sections describe the archive data and provide a brief overview of the LIMSI contemporary broadcast news transcription system used as the basis for this work. The experimental results are given in Section 4. Transcription results obtained using the baseline system are given and several ways to improve the transcription quality for the archive data are explored: training reduced bandwidth acoustic models; training Gaussian mixture models for speech detection on matching data; and estimating language models on texts covering news of the same period as the archives.

2. ARCHIVE DATA DESCRIPTION

The experiments reported here were conducted on native French documents from the Eurodelphes database provided by INA. These

This work was partially financed by the European Commission under the IST-1999-11994 Project ECHO.

Decade	*Number of documents*	*Length (minutes)*
1940	13	24
1950	22	53
1960	50	204
1970	31	103
1980	16	64
1990	31	92
All	163	540

Table 1. Number of documents and total length per decade for the Eurodelphes data.

documents were previously gathered by INA in the framework of the European EURODELPHES project which aimed at developing a hypermedia pedagogical environment for teaching history (http://eurodelphes.ina.fr/).

The Eurodelphes database consists of 163 documents in MPEG-1 video or audio format, dating from 1945 to 1995. The general topics cover the building of Europe, social changes in Europe following World War II through recent times, and the East/West relationship since World War II. In total there is about 9 hours of audio data, with individual document lengths ranging from 20 seconds to 20 minutes, and a mean document duration of 3m 20s. The number of documents and accumulated length per decade is given in Table 1, where it can be seen that the most represented periods are the sixties and seventies. The audio track was extracted from the MPEG-1 files and converted to 16 kHz mono wave files for processing by the automatic transcription system. The audio data delivered by INA was observed to be bandwidth limited to 6 kHz. There is a large variability in speaking style and acoustic conditions. Although the older documents sound somewhat noisier than recent ones, estimations of signal/noise ratio did not show any clear indication of increase for older documents.

Along with the audiovisual documents, INA provided document level annotations (date, length, topic, speaker identities) and time-synchronized manual word transcriptions. These transcriptions were used as references for an a posteriori evaluation of the automatic transcriptions. These transcripts were carefully checked and manually corrected using the Transcriber environment [2], to refine the temporal synchronization and to apply the same specific normalization rules as used to train language models for the speech recognizer. Some segments or documents were excluded because no reference transcription was provided or because they contained overlapping speech, foreign speech, songs, or acoustic conditions for which accurate manual transcription was not possible. About 40 minutes of data were discarded.

3. TRANSCRIPTION SYSTEM DESCRIPTION

The LIMSI broadcast news transcription system has two main components, the audio partitioner and the word recognizer [3]. Data partitioning [4] serves to divide the continuous audio stream into homogeneous segments, associating cluster, gender and bandwidth labels with each segment.

The speech recognizer uses continuous density (CD) HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The initial hypotheses are used for cluster-based acoustic model adaptation using the MLLR technique [5] prior to word graph generation. A 3-gram language model (LM) is used in the first two decoding steps. The final hypotheses are generated with a 4-gram LM and acoustic models adapted with the hypotheses of step 2.

The acoustic models were trained on about 100 hours of recent 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 bandwidths 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 Presse* 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 a set of 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 a set of 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 grapheme-to-phoneme rules 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.

4. EXPERIMENTAL RESULTS

4.1. Standard BN system and error analysis

The baseline recognition results obtained with the LIMSI standard French BN transcription system are summarized in Table 2. The total number of words, the out-of-vocabulary (OOV) rate and word error rate are given per decade. The average OOV rate is 0.9%, ranging from a high of 1.5% down to a low of 0.6%. It is a bit lower than the usual OOV rate on recent data (about 1.2%), which is probably due to the restricted range of topics covered in the corpus. 75% of the OOV words occur only once and are mainly proper names (37% of occurrences), conjugated forms of in-vocabulary verbs (28%), gender and number agreement of in-vocabulary words (11%). The percentage of OOVs due to proper names increases with the age of the data: in the forties about half of the OOVs are proper names, whereas in the eighties and nineties, only about 20% of the OOVs are proper names. The other types of OOVs are quite constant over time. Thus it appears that the normalizations used to process the recent training data is efficient on the "older" texts. The overall word error rate (WER) obtained by the NIST speech recognition scoring toolkit (http://www.nist.gov/speech/tools/) is 36.8% with the standard 8 kHz acoustic models.

It was mentioned earlier that Eurodelphes data is band-limited to 6 kHz. To remove this mismatch, new acoustic models were trained by reducing the bandwidth of the 100 hours of contemporary French broadcast news data to 6 kHz. This resulted in a small relative reduction (3%) of the word error rate to 35.7%. Observing the results by decade, we can note a WER increase for the 1940s, but this difference was found to be insignificant by NIST tools.

Decade	*# Words*	*%OOV*	*WER 8k*	*WER 6k*
1940	3148	1.5	32.5	34.1
1950	8029	1.3	34.6	33.1
1960	28365	1.1	38.2	36.7
1970	15332	0.5	41.3	41.0
1980	9667	0.8	40.9	39.7
1990	16807	0.6	29.8	28.6
All	81348	0.9	36.8	**35.7**

Table 2. Number of words, OOV rate and baseline system word error rates with 8kHz and 6kHz bandwidth acoustic models.

Decade	$M_{0.0}$	$M_{0.1}$	$M_{0.2}$	$M_{0.3}$	$M_{1.0}$	λ_u	*WER*
1940	133	125	122	**121**	274	0.30	33.1
1950	120	114	**113**	114	293	0.16	32.9
1960	109	**106**	**106**	107	318	0.24	36.9
1970	92	**91**	92	94	335	0.08	41.1
1980	**94**	**94**	95	98	366	0.06	39.4
1990	**81**	82	83	86	374	0.04	28.6
All	99	**97**	98	100	333	0.16	35.7

Table 3. Perplexity computed on the test data, with the standard and interpolated LMs and interpolation coefficient λ_u estimated on baseline transcripts. Word error rate (WER) using the interpolated language model with the lowest perplexity for the decade.

From Tables 2 and 3 we can see that there is no trend of increasing WER towards the past, despite an increase in perplexity. Indeed, the highest error rates are observed for the 1970s and 1980s. This result can be attributed to the unbalanced distribution of the corpus. To understand this result, we studied the main error sources for these two decades. Some long documents present particularly poor speaking styles and acoustic conditions, which imply low performance. For the 1970s, 5 of the 28 documents account for 32% of the words, for which there is an average WER of 56%. A significant proportion of segments in these files contain spontaneous, overlapping or noisy speech and some are even rejected by the partitioner. For the remaining documents of the decade, the WER is about 33%. For the 1980s, 1 of the 16 documents contains 43% of the words. Much of the speech in this file is poorly articulated, and the WER on this document is particularly high (51%) compared to the rest of the decade (31%).

The relatively good performance for the older documents can probably be attibuted to a clearer speaking style. News clips from the 40s and 50s were generally recorded in more controlled conditions, with more preparation. The speaking rate estimated using the reference transcriptions was found to be about 160 words per minute for the forties, 170-175 for the fifties, sixties and seventies, and over 185 words per minute for the eighties and nineties. The much slower delivery style found in older documents seems to facilitate the transcription task, partially counterbalancing the effects of increased perplexity and OOV rates.

4.2. Language model adaptation

The baseline n-gram LMs were trained on texts dating from 1987 through 1999. Table 3 shows perplexity of the baseline fourgram LM ($M_{0.0}$) on the reference transcripts. It can be seen that the perplexity of the older data is higher than the more recent data, increasing from a perplexity of 81 for data from 1990 to 133 for the 1940's data. Given the importance of the training data epoch [6], a source of texts covering the data period (i.e., dating back to 1945) is required. However, it is not easy to locate sufficient electronic texts from this period. We found a French video archive web site with documentary summaries covering the period from 1945 to 1979 (http://www.newsreels.gaumont.com/). These texts were cleaned and normalized in order to create an additional training corpus. Enumerations, technical and structural phrases were removed by filtering out short sentences (under 10 words). The resulting corpus contains 2.7M of words (1.4M words for the fifties, 1.0M words for the sixties and 0.3k words for the forties). Since there is not enough data to train a LM per decade, standard backoff LMs M_{epoch} were estimated on the complete text set. Since the OOV rate is low, the vocabulary was not modified. The M_{epoch} perplexity has the inverse characteristic of the baseline LM, increasing from the forties to the nineties. We therefore attempt to use this model to add linguistic information from the older periods by interpolating the two LMs: $M_\lambda = \lambda.M_{epoch} + (1 - \lambda).M_{baseline}$. To measure the impact of the interpolation coefficient, we computed three LMs with $\lambda = 0.1,\ 0.2,\ 0.3$. An alternative is to directly estimate the interpolation coefficients using the EM algorithm to minimize the perplexity on a development set. However, the ECHO corpus is too small to extract a development set, so we decided to investigate an approach using unsupervised adaptation [7]. For each decade, we compute the interpolation coefficient λ_u which minimizes the perplexity on the transcripts generated by the standard BN system. The perpexities per decade with the different LMs are shown in Table 3. Rounding the λ_u to the nearest coefficient (0.0, 0.1, 0.2 or 0.3), we obtain LM with the lowest perplexity for the decade (shown in bold). The transcripts obtained with the standard BN system can therefore be used to select the LM interpolation coefficient. Figure 1 shows the relative gain in perplexity for the earlier decades : 9%, 6%, 3%, 1% from forties to seventies. For the two most recent decades, the lowest perplexity is obtained with the baseline LM. The M_{epoch} weight and the associated gain in perplexity decrease with time, as the baseline LM training corpus becomes more relevant. However, as can be seen in the last column Table 3 the reduction in perplexity does not result in an equivalent reduction in WER.[1]

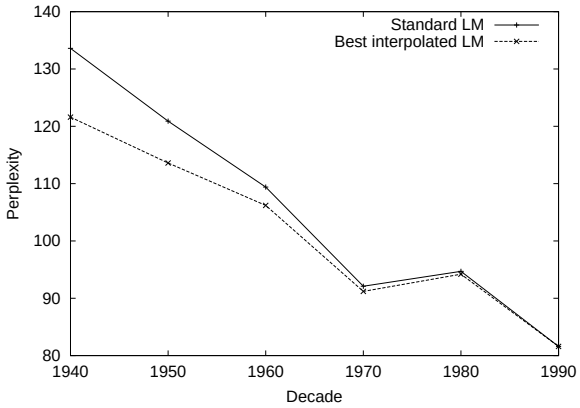

Fig. 1. Perplexity by decade for the standard LM and the best interpolated LM.

[1] In fact, the only significant gain was observed for the forties using the $M_{0.2}$ LM, which has a slightly higher perplexity than $M_{0.2}$, and resulted in a word error of 32.9% (not shown in the table). The differences for the other decades are not significant.

S/NS models	*Baseline*	*Baseline*	*Adapted*
Test set	*Full*	*1h subset*	
MD (%)	5.6	8.9	1.0
FA (%)	4.8	5.2	1.6
FSE (%)	9.5	12.5	2.4
WER (%)	35.7	41.4	37.7

Table 4. Speech/non-speech missed detections (MD), false alarms (FA), global frame segmentation errors (FSE) and resulting word error rate (WER) on the complete Eurodelphes corpus and a 1 hour test subset using baseline and adapted partitioning models.

4.3. Speech/non-speech detection

As was noted in the error analysis (see Section 4.1), some speech segments were discarded by the partitioner [4], resulting in unrecoverable errors for the transcription system. Comparing the automatic labeling with the manual transcriptions, there is a 9.5% speech/non-speech frame segmentation error using the baseline models, with 5.6% missed detections (MD) of speech frames, and 4.8% of false alarms (FA) (see Table 4). False alarms (non-speech frames taken as speech by the partitioner) are less harmful than missed dectections since they can be rejected later by the recognizer, thus only errors on non-speech segments lasting over 2 seconds were considered erroroneous. Also, since the manual transcription is segmented at the sentence or breath group level rather than at the word level, short inter-word silences are labeled as speech and thus the MD rate is slightly over-estimated. The frame segmentation error on this data is more than twice that on contemporary English broadcast news data [4]. A detailed analysis per decade shows that the frame segmentation error rate is under 2% for the nineties, and is over 9% for all other decades.

Supervised adaptation of the speech/non-speech Gaussian mixture models was performed as follows: a 1 hour subset of the Eurodelphes database selected by INA for a demonstration of the ECHO project was reserved for testing and the remaining 7 hours were used for training the acoustic speech and non-speech models. The selected test subset is clearly harder than the whole set, with a 41.4% word error rate, to be compared to 35.7%. On this 1 hour test set, both the speech frame missed detections and false alarms decrease dramatically with the adapted models, resulting in a 80% reduction in the frame segmentation error and about 9% relative reduction of the resulting word error rate.

5. SUMMARY

In this paper we have assessed the transcription performance of a state-of-the-art contemporary broadcast news transcription system on a set of audio-video archives spanning the latter half of the 20th century. The word error of this transcription system is 35.7% on a 9-hour set of archive data. Several sources of mismatch between the training data and the archive data were identified, including the acoustic quality (bandwidth and background conditions) which affects the acoustic models used both during partitioning and recognition, language model (changes in vocabulary and linguistic style), pronunciation and articulation. Although there is an increase in perplexity and out-of-vocabulary rate for older documents, these values remain within a reasonable range. In contrast, no systematic trend was observed for the word error rate, probably due to the high acoustic variability of the documents. However, the speaking rate was found to be slower on older documents, which may have made the decoding easier.

Some attempts were made to improve transcription performance on the archive data. The data partitioner rejected a much larger proportion of speech frames (9.5%) than are rejected for recent data (3.7%). Supervised adaptation of the baseline audio partitioner on a few hours of archive data dramatically improved speech/non-speech detection on a set aside portion of the data. The acoustic models were retrained to match the bandwidth of the archive data, resulting in a small, but significant error reduction.

Unsupervised language model adaptation was carried out using documentary summaries found on the Internet site. However, despite a reduction of perplexity on older documents, only a minimal decrease in word error was observed. It may be that the current mismatch of the acoustic models is partially masking the improvements of the language model, or it may be that using a single historical language model is too general and that more specific language models are needed for the different epochs. This requires obtaining additional training texts, which may also allow adaptation of the vocabulary.

These experiments illustrate that a transcription system developed on contemporary French broadcast news is able to transcribe historical news from national audio-visual archives with sufficient accuracy for certain automatic indexation tasks. Attempts to improve performance resulted in only a small gain.

6. ACKNOWLEDGEMENTS

The authors wish to thank our partners in the project, especially Laurent Vinet from INA, for their support; Gilles Adda for providing the baseline language models used in these experiments and normalizing the transcriptions, and Sylvia Hermier for carefully checking and refining the reference transcriptions.

7. REFERENCES

[1] J. Garofolo, C. Auzanne, E. Voorhees, W. Fisher, "1999 TREC-8 Spoken Document Retrieval track overview and results," in *Proc. TREC-8*, 1999.

[2] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1–2), 2001.

[3] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, 2002, to appear.

[4] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. ICSLP*, 1998.

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

[6] M. Adda-Decker, G. Adda, J.L. Gauvain, L. Lamel, "Large vocabulary speech recognition in French," in *Proc. ICASSP*, 1999.

[7] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Language model adaptation for broadcast news transcription," in *Proc. ISCA ITR Workshop*, 2001.