ISSUES IN AUTOMATIC TRANSCRIPTION OF HISTORICAL AUDIO DATA

F. Brugnara, M. Cettolo, M. Federico, D. Giuliani

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo di Trento - Italy
{brugnara,cettolo,federico,giuliani}@itc.it

ABSTRACT

This work deals with some interesting issues that arose when the ITC-irst broadcast news transcription system was applied to transcribe the audio track of historical documentary films. Due to an evident acoustic and linguistic mismatch between the broadcast news and the new application domain, the initial word error rate was of 46.4%. By exploiting a limited amount of manually annotated training data, adaptation of all components of the transcription system was performed, namely the audio partitioner, the acoustic model, and the language model. This permitted to achieve a word error rate of 30%, which makes automatic transcription of documentary films effective for information retrieval applications.

1. INTRODUCTION

This work builds on previous experience [1, 2] about porting the ITC-irst Italian broadcast news (BN) speech recognition system to a spoken dialogue domain. In this work, a more difficult task is approached, which consists in transcribing the audio track of historical documentary films, made available within the European project ECHO (European CHronicles On-line).

In general, when an automatic transcription system trained and tailored for the BN domain is applied to transcribe audio streams of a new domain, interesting issues arise, mainly due to mismatches at the acoustic and linguistic levels. In particular, the peculiar characteristics of the historical film data make them a challenging test bed for the porting techniques explored within the European project CORETEX, which aims at improving genericity and adaptability of automatic transcription technology. The historical data represent a formidable benchmark to evaluate, and possibly improve, the adaptability of all components of our transcription system, namely the audio partitioner, the acoustic model, and the language model. As these components are typically trained from manually transcribed and annotated data, the effectiveness of porting was related to the used amount of domain specific training data versus the achieved performance.

As will be explained in the paper, different trade-offs result for each component, which depend on the applied adaptation technique and on the complexity, i.e. number of free parameters, of the underlying statistical model.

This paper is organized as follows. First, the historical audio data are introduced in Section 2; then, the ITC-irst BN transcription system is summarized in Section 3. Section 4 describes the porting experiments carried out. Finally, conclusions about the results are reported in Section 5.

2. THE HISTORICAL CHRONICLES TASK

The ECHO project aims at developing a digital library service for historical film collections belonging to large national audiovisual archives. These collections of documentary films, going from the twenties until the sixties of the last century, are of extraordinary value since they document the different aspects (social, cultural, political, economic) of life in different European countries (France, Italy, the Netherlands and Switzerland) during this period of time. Automatic transcriptions will be used to create meta-data, to produce indexes for cross-language information retrieval, and to generate automatic summaries.

For the development of the Italian component, 240 Italian video documents were made available by Istituto Luce, one of the ECHO partners. The video documents, mainly newsreels and documentary films, were selected from an archive of 4,000 hours of historical films. The selected films cover a variety of topics and were produced between the ’30s and ’60s of last century.

Video documents were available in digital format as MPEG-1 files. For each file, associated textual data were also provided. These data contained information about the document, i.e. title, date of issue, genre, ownership, etc., and its content, i.e. a terse description, keywords, theme, etc.

2.1. Analysis of Audio Data

The audio MPEG-1 files, which contained stereo signals sampled at 44100 Hz and encoded at 128 kbit/s, were converted and downsamped into 16000 Hz mono signals. In order to train and evaluate system performance, the audio files were manually annotated and transcribed following the conventions used for transcribing the BN shows. Finally, these historical chronical data (HC) were split into a training set of 157 files, and a test set of 83 files. In the following, these sets will be referred to as HC-train and HC-test, respectively.

Analysis of the audio data and of the manual transcriptions showed a great variability of acoustic conditions and covered topics. Acoustic variability is mostly due to the use, throughout the years, of different recording equipments and film media. Analysis of signal spectrograms showed, for example, that the effective bandwidth is not the same for all the signals, but ranges between 5kHz and 8kHz. The general quality of the audio also appears significantly lower than that found in recent BN recordings. Moreover, most of the speech in the videos has background music or is noisy (see Table 1). To summarize, the 240 audio tracks contain, on the average, about 46% of non speech signal, 48% of speech with background noise or music, and less than 6% of clean speech.
3. THE BN TRANSCRIPTION SYSTEM

The ITC-irst BN transcription system [1] features an audio partitioner, which is described in Section 4.1, and a speech recognition module. Speech recognition exploits context dependent HMMs AM, a 64K-word trigram LM, beam-search Viterbi decoding, and maximum likelihood linear regression (MLLR) AM adaptation.

The acoustic front-end uses a sliding window of 20ms, with a step of 10ms, to compute 12 mel-scaled cepstral coefficients, the log-energy and their first and second time-derivatives. Mean subtraction is applied to cepstral coefficients, while log-energy is normalized by subtracting the maximum value in the utterance.

The BN transcription system employs different AMs for wide and narrow band data. However, only the wide band AM was considered in the experiments on porting. This AM was estimated on 57h of BN recordings by using an agglomerative training procedure [3]. The LM was estimated on a 226M-word corpus including newspaper articles, for the largest part, and BN transcripts. Newspaper articles were taken from issues covering the period 1992-1999. The BN transcripts consisted of about 0.5M words, corresponding to the manual transcription of the AM training set. The LM was compiled into a static network with a shared-tail topology [3]. Table 2 reports statistics concerning the AM and LM training and representation.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Test</th>
<th>Training</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Films</td>
<td>83</td>
<td>137</td>
<td>240</td>
</tr>
<tr>
<td>Duration</td>
<td>3h:15m</td>
<td>6h:01m</td>
<td>9h:16m</td>
</tr>
<tr>
<td>Clean Speech</td>
<td>0h:10m</td>
<td>0h:21m</td>
<td>0h:31m</td>
</tr>
<tr>
<td>Speech with Music</td>
<td>1h:31m</td>
<td>1h:55m</td>
<td>3h:26m</td>
</tr>
<tr>
<td>Noisy Speech</td>
<td>0h:14m</td>
<td>0h:50m</td>
<td>1h:03m</td>
</tr>
<tr>
<td>Non Speech</td>
<td>1h:20m</td>
<td>2h:55m</td>
<td>4h:16m</td>
</tr>
</tbody>
</table>

Table 1. Composition of the HC audio data.

Table 2. Statistics concerning the broadcast news AM and LM.

Audio transcription is decomposed into several stages: first the input signal is segmented, and identified segments are clustered into homogeneous groups; then, speech segments are transcribed by means of two decoding steps, which are interleaved by a cluster-based unsupervised AM adaptation exploiting transcriptions generated by the first decoding step. This is performed with MLLR [4], in which Gaussian means are adapted by means of a global affine transformation.

4. PORTING EXPERIMENTS

Preliminary experiments were carried out by applying the BN transcription system on HC-test. Recognition results confirmed the mismatch existing between the BN and the HC tasks. In fact, the baseline transcription system achieved WERs of 54.2% and 46.4% with, respectively, one and two decoding steps. This badly compares with the performance achieved on the BN task, i.e. 19.7% and 17.7%.

Hence, adaptation and tuning methods were investigated for all the components of the transcription system. In order to evaluate the cost of porting, adaptation of the modules was carried out with increasing amounts of training data.

4.1. Audio Partitioning

The partitioner consists of two main modules (see Figure 1): the segmenter, based on the Bayesian Information Criterion (BIC), and the classifier, using Gaussian Mixture Models (GMMs).

BIC-based Segmentation. Segmenting an audio stream means to detect the time indexes corresponding to changes in the nature of audio, in order to isolate segments that are homogeneous in terms of bandwidth and speaker.

Segmentation is based on a statistical model selection criterion, namely the BIC [5, 6]. According to the BIC, the decision of hypothesizing or not a change in a particular time index can be based on a threshold \( \lambda \) that determines the sensitivity of the method.

Segment Classification. The final goal of the partitioning stage is to classify each acoustically homogeneous segment in terms of broad acoustic classes. Here, the acoustic classes are modeled by GMMs and the classification is done through the Viterbi algorithm on a search space in which the activation of a new class is possible at each time, by using a network with loop topology. This process induces a refinement of the segmentation made by the BIC, since the time indexes of class changes correspond to new segment boundaries.

Finally, speakers are clustered through a BIC-based hierarchical method [6]. In this paper, the porting of this algorithm is not investigated.

Automatic Tuning of the BIC [6]. The decision threshold of the BIC can be estimated by exploiting the segmentation induced by the classifier in the following way. The whole audio stream is given as input to the GMM-based classifier. For each hypothesized boundary, the BIC threshold value is computed, that would allow to detect that boundary. The set of these \( \lambda \) values is then used to select the operating point of the BIC algorithm.
combination schemes are possible. Actually, the best experimental results were obtained by setting the BIC decision threshold to the mean of all computed values.

Recognition experiments were carried out to evaluate different configurations of the partitioner, by exploiting domain adapted AMs (see Section 4.2), and a single pass Viterbi decoding. A first set of experiments was conducted by training the classification module on increasing amounts of manually annotated data. The BIC parameter $\lambda$ was instead estimated empirically on a development set and kept fixed. In Figure 2 both classification results in terms of frame classification accuracy (FCA) and recognition results in terms of WER on HC-test set are given as functions of the amount of training data. The point $\text{BN}$ corresponds to the BN configuration of the classifier. FCA is given in terms of 5 generic acoustic classes: wide- and narrow band speech from female and male speakers, and non-speech. It emerges that GMMs trained on BN data poorly classify HC documents, although the impact on speech recognition is not so severe. Moreover, even a small amount of HC-train data (i.e. 1 hour) permits to obtain a level of performance close to that corresponding to using all the available training data (about 6 hours).

![Fig. 2](image)

Fig. 2. FCA and WER as functions of the training data: using BN GMMs and using an increasing amount of HC data for training.

Other experiments were performed to evaluate the empirical and automatic tuning of the BIC parameter. In Table 3, recognition results are reported. The first row shows WERs obtained by using manual annotations instead of any automatic partitioner; these figures are taken as references. The second row shows WERs obtained with domain specific partitioners, in which both the BIC threshold and the GMMs are tuned/trained on domain specific data. The last row refers to experiments in which the automatically tuned segmenter is employed. It has to be noted that in these cases, while GMMs are domain specific, the segmenter is exactly the same for the two domains. In the experiments of the following sections, the partitioner with the automatic estimation of the BIC threshold will be adopted.

<table>
<thead>
<tr>
<th>partitioner</th>
<th>HC</th>
<th>BN</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual</td>
<td>32.8</td>
<td>18.5</td>
</tr>
<tr>
<td>$\lambda_{\text{Echo}}, \lambda_{\text{BN}}$</td>
<td>35.9</td>
<td>19.7</td>
</tr>
<tr>
<td>$\lambda_{\text{AutExt}}$</td>
<td>35.0</td>
<td>19.7</td>
</tr>
</tbody>
</table>

Table 3. Recognition results (WERs) with different partitioners.

In conclusion, the following comments can be made:
(a) The degradation in recognition performance due to the automatic partitioner is comparable across the two domains.
(b) The automatic estimation of the BIC parameter works as well as the empirical method using an annotated development set. Moreover, in the case of the HC domain, the former gives even a better accuracy. This is possible because, unlike the empirical tuning method, the automatic one is applied on a file by file basis, thus permitting a more refined behavior, at least in theory.
(c) Both the classification and the automatic $\lambda$ estimation need a set of domain specific GMMs. Figure 2 shows that in order to train such models little data are required. The reason is that acoustic classes to be discriminated are few, well separated, and their models have a small number of parameters to be estimated (in our set-up, GMMs have 32 components, and the feature space has dimension 39).

4.2. Acoustic Model Adaptation

Domain adapted AM was obtained through MLLR using manually transcribed data (see Table 1). To better exploit adaptation data, a regression class tree was employed for dynamic definition of regression classes. Gaussian means were adapted by using full transformation matrices while diagonal transformation matrices were employed for adapting variances [7].

In Table 4 recognition results are reported, which were obtained by adapting the AM with 1h up to 6h (about 3h of speech) of data. Experiments were carried out with a single decoding step. Row \textit{automatic} gives results obtained with the partitioner trained on the same data. In row \textit{manual}, results obtained with the manual segmentation are reported for comparison purposes. Experiments show that 4 hours of training data permit to significantly reduce the acoustic mismatch, i.e. the WER reduces by 34.7%, from 54.2% to 35.4%. By using two more hours just another 1% relative improvement is achieved. The application of the second decoding step, after cluster-based AM adaptation, reduces the WER by another 8.6% relative, i.e. to 32.3%.

![Table 4](image)

Table 4. WERs achieved adapting, with different amounts of adaptation data, both the partitioner and the AM of the BN transcription system.

4.3. Language Model Adaptation

An analysis of the performance of the BN LM on HC-test showed value of perplexity and out-of-vocabulary (OOV) word rate about three times higher than those usually measured on the BN domain. The reason for this mismatch is mostly in the long period and wide range of topics covered by the archive. This involves for instance many proper names that are nowadays quite unusual.

LM estimation for the HC domain had, unfortunately, to cope with the unavailability of language resources directly related to the contents of the historical films. In particular, only 30k words from the manual transcripts of the training data were available. Hence,
starting from available corpora of newspapers and broadcast news transcripts, different mixtures of LMs were tested [8]. The reference BN LM, which results from the combination of newspaper and BN transcripts, achieves a perplexity of 1,013 and an OOV word rate of 4.4%. In Table 5, dictionary size (column V), perplexity (column PP) and OOV word rate (column OOV) are reported for different estimated LMs.

A slight improvement in perplexity was achieved by removing the BN transcripts from the LM, thus by just using the newspaper material (row NP). The resulting LM was then combined with all the available task specific transcripts, by keeping the dictionary fixed (row NP + Transcr), i.e. by not including words found in the HC domain. Practically, this allowed to better estimate the OOV word probability. Finally, the LM vocabulary was extended, just at the unigram level, with further 120,000 words, selected according to frequency in the newspaper corpus (row NP + Transcr + Lex). This solution allowed to further reduce the perplexity by 8.2% and the OOV word rate by over 50%.

4.4. Transcription Results

This section summarizes recognition results obtained on the HC task with different settings of the ITC-irst transcription system. In particular, results are shown after adapting single components of the system to all the available training data of HC-train (6 hours). The second and third columns of Table 6 show WERs obtained by applying one or two decoding steps. It is worth noticing that, by following the order in Table 6, the largest improvement is achieved with AM adaptation, which alone accounts for a 28.4% WER relative reduction on the single decoding step, i.e. from 54.2% to 38.8%. The tuning of the partitioner provides a further 9.8% WER reduction. Finally, LM adaptation and lexicon extension provides a 7.4% WER reduction. Globally, the adaptation process permitted to reduce the WER by 40.2% and 35.3%, respectively, with one and two decoding steps.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>1-pass decoder</th>
<th>2-pass decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN Baseline</td>
<td>54.2%</td>
<td>46.4%</td>
</tr>
<tr>
<td>+ AM adaptation</td>
<td>38.8%</td>
<td>-</td>
</tr>
<tr>
<td>+ Partitioner porting</td>
<td>35.0%</td>
<td>-</td>
</tr>
<tr>
<td>+ LM adaptation</td>
<td>32.4%</td>
<td>30.0%</td>
</tr>
</tbody>
</table>

Table 6. Recognition results (WERs) obtained on HC-test under different settings of the BN baseline system.

5. CONCLUSION

In this work we dealt with porting issues arisen when the ITC-irst BN transcription system was applied to transcribe historical documentary films. Analysis of historical audio data revealed a severe mismatch, at acoustic and linguistic level, with respect to the BN data. Hence, porting of each components of the system was separately considered, by exploiting about 6 hours of task dependent training data. Different trade-offs resulted between cost of porting, i.e. amount and quality of training data, and achieved performance. In the following the outcomes for each component are briefly summarized.

The partitioner requires training data annotated at the level of acoustic content, i.e. acoustically homogeneous segments must be detected and classified. This annotation is much cheaper than transcribing speech. It results that porting would require about one hour of annotated audio signal, providing it is representative of the application.

Adaptation of the AM model exploited limited amounts of speech data with manual transcriptions, which were used for supervision. Previous work [2] has shown that AM adaptation does not require very detailed transcriptions, i.e. the annotation of spontaneous speech phenomena can be omitted. In our experiments most of the performance gain was achieved with four hours of training data, roughly corresponding to two hours of speech.

Language model adaptation requires texts related to the domain, which could be acquired from scripts or other sources, e.g. close captions. In this work, a 30K-word corpus of manual transcriptions was used, which roughly corresponds to 3 hours of read speech.

In conclusion, this work showed that even a difficult speech transcribing task can be approached with statistical adaptation methods once few hours of task-specific transcribed data are available. In particular, exploiting 6 hours of films, i.e. 3 hours of transcribed speech, permitted to achieve a level of performance acceptable for information retrieval applications.

6. REFERENCES


