ADVANCES IN AUTOMATIC TRANSCRIPTION OF ITALIAN BROADCAST NEWS

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

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy.

ABSTRACT

This paper presents some recent improvements in automatic transcription of Italian broadcast news obtained at ITC-irst. A first preliminary activity was carried out in order to develop a suitable speech corpus for the Italian language. The resulting corpus, formed by recordings covering 30 hours of radio news, was exploited for developing a baseline system for transcription of broadcast news. The system performs in different stages: acoustic segmentation and classification, speaker clustering, acoustic model adaptation and speech decoding. Major recent advances allowing performance improvement concern with speech segmentation and clustering, acoustic modeling, acoustic model adaptation and the language model.

The transcription system features a 14.3% word error rate on planned studio speech and 18.7% on the whole test set formed by recordings of radio broadcast news. When applied to a test set formed by recordings of television broadcast news, the system features 16.5% word error rate on planned studio speech and 23.2% by considering the whole test set.

1. INTRODUCTION

Technologies that make the management and the access of multimedia archives easier are receiving more and more attention due to the increasing availability of large multimedia digital libraries. Technologies for audio transcription and indexing of multimedia archives are among the emerging technologies.

This work presents a system for the automatic transcription of Italian broadcast news.

A preliminary activity was carried out in order to develop a suitable speech corpus for the Italian language. 30 hours of recordings covering radio news of several years were collected, labelled and transcribed. The annotated material formed a speech corpus called Italian Broadcast News Corpus (IBNC). This corpus was used for developing the system.

The current system performs four stages [1]: acoustic segmentation and classification, speaker clustering, acoustic model adaptation, and speech transcription. The transcription system was tested with two different sets of data formed, respectively, by recordings of radio and television broadcast news. On radio recordings, the system features a 14.3% word error rate on planned studio speech and 18.7% on the whole test set. When applied to television broadcast news, the system features 16.5% word error rate on planned studio speech and 23.2% on the whole test sample.

The paper is organized as follows. Section 2 introduces the activities of data collection and annotation ongoing at ITC-irst. Section 3 briefly sketches the audio segmentation, classification and clustering algorithms. Section 4 outlines the adopted acoustic modeling, language model, recognition engine and speaker adaptation. Experimental results are presented in Section 5.

2. BROADCAST NEWS CORPORA

Since Summer 1999, ITC-irst has been collecting Italian broadcast news corpora. First, the IBNC corpus was developed under a contract with ELRA/ELDA1. RAI, the major Italian broadcast company, supplied recordings of radio news programs sampled from its internal digital archive. The collection consists of 150 programs, for a total time of about 30 hours, issued between 1992 and 1999. The corpus contains about 7 hours of telephone speech. The IBNC was segmented, labelled and transcribed following conventions similar to those adopted by the Linguistic Data Consortium2 for the HUB-4 corpora. The corpus, which was released in April 2000, will be distributed by ELRA.

Recently, ITC-irst has started to develop an in-house corpus of television broadcast news, a small part of which has been used in this work. According to our plans, by the end of this year about 100 hours of transcribed material will be available for development and evaluation purposes.

3. SEGMENTATION AND CLUSTERING

The Bayesian Information Criterion (BIC) [2] is applied to segment the input audio stream into acoustically homogeneous chunks. Gaussians mixture models are then used to classify segments in terms of acoustic source and channel. Emission probability densities consist of mixtures of

1European Language Resources Association.
2www.ldc.upenn.edu.
1024 multi-variate Gaussian components having diagonal co-variance matrices. Observations are 39-dimension vectors (see Section 4.1). Six classes are considered for classification: female/male wide-band speech, female/male narrow-band speech, pure music, and silence plus other non-speech events.

Clustering of speech segments is done by a bottom-up scheme [2, 3] that groups segments which are acoustically close with respect to the BIC. As a result, this step should gather segments of the same speaker.

To evaluate the segmentation algorithm in detecting the break points, recall and precision are computed with respect to target (manually annotated) boundaries: with half a second time tolerance, the two measures result equal to 82.9% and 85.9%, respectively.

Classification accuracy, in terms of the six above classes, is 95.6%. Defining the purity of a cluster as the percentage of speech uttered by its dominant speaker, the clustering algorithm provides an average cluster purity of 94.1%. More details about the content of this section can be found in [4].

### 4. SPEECH TRANSCRIPTION

#### 4.1. Acoustic Modeling

Acoustic modeling is based on continuous-density HMMs. The acoustic parameter vector comprises 12 mel-scaled cepstral coefficients, the log-energy and their first and second time-derivatives. For either channel condition, a set of context-dependent units is defined, based on the SAMPA phonetic alphabet. Some additional units have also been introduced, to cover silence, background noise and a number of spontaneous speech phenomena. They are used both in training and in recognition by allowing their optional insertion between words.

Acoustic training was performed with Baum-Welch reestimation using the training portion of the IBNC corpus, augmented with other datasets collected at ITC-irst. Table 1 shows the amount of training data used for each model set.

<table>
<thead>
<tr>
<th></th>
<th>wide-band</th>
<th>narrow-band</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>26h:42m</td>
<td>15h:42m</td>
</tr>
<tr>
<td>#Words</td>
<td>236000</td>
<td>121000</td>
</tr>
<tr>
<td>#Triphones</td>
<td>6150</td>
<td>3814</td>
</tr>
<tr>
<td>#Backoff Models</td>
<td>2327</td>
<td>1181</td>
</tr>
<tr>
<td>#Gaussians</td>
<td>15752</td>
<td>10573</td>
</tr>
</tbody>
</table>

Table 1: Properties of the two model sets used in the experiments.

As previously stated, models for backoff units are built by agglomeration. For each backoff unit, a model is built whose topology corresponds to the common topology of the allophones of the base phone. The transition probabilities are then estimated by summing the counters computed during Baum-Welch reestimation of all the subsumed triphones. Similarly, each mixture of the agglomerate model is built by joining the components of all the mixtures in the same position within the triphone models, and estimating their weights by combining the statistics of the weights of the included mixtures. The backoff models therefore contain model-specific mixtures whose components belong to the common Gaussian pool of the allophones. The same procedure was applied for the wide-band and narrow-band models. The characteristics of the resulting model sets are reported in Table 1.

As for the training lexicon, the recognition lexicon is generated using triphones only when there is a minimum number of occurrences in the training set. This time, however, miss-
ing triphones are not mapped to remainders, but to backoff units, and the threshold is set to 50. Thresholding is applied to backoff units as well, so that context-independent units can appear.

4.2. Language Model

A trigram language model was developed by mainly exploiting newspaper text sources. In particular, a 133M-word collection of the nation wide newspaper La Stampa was employed, that includes all issues between 1992 and 1998. Moreover, the broadcast news transcriptions of the training data were added. Numeric expressions occurring in the texts were replaced by labels. As explained later, during the decoding phase, labels will be linked to specific rule-based LMs.

A lexicon of the most frequent 64K words was selected. The lexicon gives a 2.2% OOV rate on the newspaper corpus and about 1.6% on the IBNC corpus. An interpolated trigram LM was estimated by employing a non linear discounting function and a pruning strategy that deletes trigrams on the basis of their context frequency. A good trade-off was obtained by using a shift-1 discounting function and by pruning all trigrams with context frequency less than 10. This results in a pruned LM with a perplexity of 188 and a size of 14M.

4.3. Recognition Engine

The recognizer is a single-step Viterbi decoder. The basic Viterbi beam-search algorithm is extended [5] to deal with recurrent transition networks, which in this case are used to represent subgrammars in the LM for common entities such as cardinal numbers, ordinal numbers, and percentages. The use of subgrammars and a recursive decoder has the effect of reducing the network size during decoding, as there is no need to embed the subgrammars in the main trigram network for each occurrence, and also allows for a more robust estimation of the LM.

The 64K-word trigram LM is mapped into a static network with a shared-tail topology [6]. The main network has about 11M states, 10M named transitions and 17M empty transitions.

4.4. Acoustic Model Adaptation

On each cluster of speech segments, unsupervised acoustic model adaptation is carried out by exploiting the transcriptions generated by a preliminary decoding step.

Gaussian components in the system are adapted using the Maximum Likelihood Linear Regression (MLLR) technique [7, 8]. A global regression class is considered for adapting only the means or both means and variances. Mean vectors are adapted using a full transformation matrix, while a diagonal transformation matrix is used to adapt variances.

5. EXPERIMENTAL RESULTS

Transcription experiments were carried out with two different test sets: the IBNC test set, formed by 6 radio news programs for a total of 1h:15m, and a test set formed by two television news programs, for a total of 40 minutes.

In all cases, automatic transcription experiments exploited the segmentation automatically computed by the acoustic segmentation and classification modules. According to their class, speech segments were processed with the general wide-band or narrow-band acoustic models. The resulting transcripts were used to perform acoustic model adaptation on each cluster. Then, the final decoding step was performed by using the adapted acoustic models.

Table 2 reports results obtained on the IBNC test set. The Word Error Rate (WER) is reported for speech material corresponding to the different focus conditions and for the whole test set. Recognition experiments were carried out without (Baseline column) and with acoustic model adaptation (MLLR column). Acoustic model adaptation refers to Gaussian mean adaptation. Adaptation of both means and variances did not improve further the performance.

Table 3 reports results obtained on recordings of television broadcast news. Comparing these results with those obtained on the IBNC test set, it can be noted that they are tangibly worse.

It should be pointed out that the first test set consists of material recorded in homogeneous conditions with respect to the training portion of the IBNC corpus. Instead, the second test set consists of recordings from television programs which are different from radio programs for many aspects such as the studio acoustic conditions and the structure of the programs. This is also shown by the different distributions of the
uttered words with respect to the focus conditions, as reported in Tables 2 and 3. In the television news programs telephone speech (focus condition F2) is not present, while in the radio programs telephone speech represents a significant part of the whole speech material. On the contrary, speech with generic background (focus condition F4) represents a significant part of the television recordings while in the first test set material corresponding to condition F4 represents a marginal part of whole set of data.

6. CONCLUSION

Work is in progress at ITC-irst to augment the training data and to improve algorithms.

A comparison with a previous version of the system [1], which used less training material, shows that augmenting the training data and improving acoustic modeling allowed a significant improvement of performance, from 21.5% WER to 18.7% on the IBNC test set. Furthermore, the difference, in terms of transcription performance of radio and television broadcast news, suggests that the training set should be enlarged in order to contain an adequate amount of examples of recordings of television news programs. Activity for collecting a suitable amount of data, for development and evaluation purposes, has been already scheduled.

Future work will be devoted mainly to acoustic model-

Table 2: Performance of the speech transcriber on the IBNC test set.

<table>
<thead>
<tr>
<th>Focus condition</th>
<th>Word distrib.</th>
<th>Baseline</th>
<th>MLLR (means)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>57.4%</td>
<td>15.6%</td>
<td>14.3%</td>
</tr>
<tr>
<td>F1</td>
<td>0.9%</td>
<td>42.6%</td>
<td>29.6%</td>
</tr>
<tr>
<td>F2</td>
<td>23.1%</td>
<td>27.8%</td>
<td>26.0%</td>
</tr>
<tr>
<td>F3</td>
<td>6.1%</td>
<td>23.8%</td>
<td>21.6%</td>
</tr>
<tr>
<td>F4</td>
<td>3.0%</td>
<td>35.1%</td>
<td>29.8%</td>
</tr>
<tr>
<td>F5</td>
<td>0.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FX</td>
<td>9.4%</td>
<td>25.0%</td>
<td>20.9%</td>
</tr>
<tr>
<td>Global</td>
<td>100.0%</td>
<td>20.6%</td>
<td>18.7%</td>
</tr>
</tbody>
</table>

Table 3: Performance of the speech transcriber on recordings of television broadcast news.

<table>
<thead>
<tr>
<th>Focus condition</th>
<th>Word distrib.</th>
<th>Baseline</th>
<th>MLLR (means)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>36.4%</td>
<td>18.1%</td>
<td>16.4%</td>
</tr>
<tr>
<td>F1</td>
<td>0.4%</td>
<td>48.0%</td>
<td>52.0%</td>
</tr>
<tr>
<td>F2</td>
<td>0.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F3</td>
<td>3.9%</td>
<td>33.3%</td>
<td>24.4%</td>
</tr>
<tr>
<td>F4</td>
<td>27.8%</td>
<td>36.3%</td>
<td>28.5%</td>
</tr>
<tr>
<td>F5</td>
<td>0.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FX</td>
<td>31.5%</td>
<td>34.5%</td>
<td>25.9%</td>
</tr>
<tr>
<td>Global</td>
<td>100.0%</td>
<td>29.0%</td>
<td>22.3%</td>
</tr>
</tbody>
</table>

7. ACKNOWLEDGMENTS

The here presented work has been carried out within the European project CORETEX (IST-1999-11876).

8. REFERENCES


