EVALITA 2011
Automatic Speech Recognition
Large Vocabulary Transcription

Marco Matassoni, Fabio Brugnara, and Roberto Gretter

FBK-irst, via Sommarive 18, Povo (TN), 38123, Italy
{matasso,brugnara,gretter}@fbk.eu
http://www.fbk.eu

Abstract. In this paper we describe motivations and setup of the Speech Recognition task in the framework of the EVALITA campaign for the Italian language. Systems are compared with respect to recognition accuracy on audio sequences of Italian parliament. Although only a few systems participated to this task, the recognition results give an overview of the performance. A more general discussion about approaches to large vocabulary speech recognition concludes the presentation.

Keywords: Automatic Speech Recognition, Large Vocabulary, Constrained transcription

1 Motivation

The trend in Automatic Speech Recognition (ASR) is toward increasingly complex models, with the purpose of improving accuracy in different acoustic conditions and with larger vocabularies. In order to select a sufficiently complex recognition task and at the same time allow a wide participation, some constraints have been introduced in the definition of the contest. Thus, the recognition task proposed at Evalita 2011 has been designed according to these preferred features:

- large vocabulary;
- large number of speakers;
- controlled recording conditions;
- spontaneous speech but with limited colloquial or dialectal expressions;
- availability of data for acoustic and language model training.

Moreover, the required data for a suitable training can be distributed in order to favor a wide spectrum of interested researchers/practitioners not owing proprietary technologies or specific audio and linguistic resources. Italian Parliament speeches satisfy these requirements: audio and minutes of all the sessions are publicly available and the additional effort to manually annotate a small portion of the corpus has already been made by FBK. The chosen context also gives the opportunity of defining two realistic subtasks, with different goals and different prospective application scenarios.

A similar task, related to the European Parliament, was taken as reference in the TC-Star European Project (see e.g. [1]) in which the Spanish and English languages were considered.

2 Definition of the Task

In the Large Vocabulary Transcription task, systems are required to transcribe recordings of sessions of the Italian Parliament. Two subtasks are defined, and applicants may choose to participate in any of them.

In the **transcription** subtask, participants are required to produce an automatic transcription of the session by exploiting only the corresponding audio file. This task corresponds to the scenario in which unknown content has to be extracted from the audio material.

In the **constrained transcription** subtask, the accompanying minutes are provided, and participant can exploit them to produce a more accurate transcription. This task corresponds to the scenario in which the goal is to align as close as possible an existing report of the session with the actual spoken content.

For each task, two training modality have been defined. In **closed** modality only distributed data are allowed for training and tuning the system while in **open** modality the participant can use any type of data for system training, declaring and describing the proposed setup in the final report.

3 Dataset

The data set distributed for model training consists in:

- about 30h of parliament audio sessions along with corresponding automatic transcriptions
- 5-years (1 legislature) of minutes of parliament sessions, for a total of about 32 millions running words;
- a 74K-word lexicon covering acoustic training data and most of language model data

The development set contains:

- 1 hour parliament audio session
- the minutes of the session
- the reference transcription

The evaluation test set includes a 1 hour recording of a parliament session, and the corresponding minutes to be used only in the constrained recognition task.

4 Evaluation Measure

The evaluation was based on Word Accuracy, computed as Minimum Edit Distance (Levenshtein distance) between the recognizer output and the reference annotation. The evaluation tool, called *sclite*, was developed by NIST and was provided in the distribution.

The reference transcriptions for the development and evaluation data were produced by manual annotation and did not include punctuation. Numbers were written in words and split in their basic tokens, e.g. 1998 → *mille nove cento novantotto*. Evaluation is case-insensitive.

5 Results

Two sites took part in the evaluation:

- Vocapia Research, Orsay, France
- Fondazione Bruno Kessler, Trento, Italy.

Vocapia submitted two outputs for the transcription task in open training modality, while FBK submitted one output for the transcription task and one for the contrained transcription task, both in closed training modality. Results are reported in Tables 1,2.

Table 1. *Transcription task*: WER (%) of the participant systems for the two different modalities (Closed and Open).

Closed	System	WER (%)
	FBK	8.4
Open		
	Vocapia (run 1)	6.4
	Vocapia (run 2)	5.4

The two Vocapia run differ in system complexity, the first one is a single-pass real-time system, while the second one is a two-pass system that includes AM adaptation and word-lattice rescoring, running in about 5×RT. The FBK system is a two-pass system that includes acoustic normalization, and runs in about 3×RT.

Table 2. *Constrained Transcription task*: WER (%) of the participant system.

Closed	System	WER (%)
	FBK	7.2

6 Discussion

The system outputs are not directly comparable as they are using significantly different training data sets. One thing that can be noted from FBK results[1] is that the provided data, albeit reduced in size, are sufficient to build a reasonable recognition system. Moreover, in the accompanying paper Vocapia reports a performance gain of 1.9% absolute when adapting the baseline system, that was tuned on Bradcast News, to the distributed data. The system share some common choices, such as the use of tied-state left-to-right 3 state HMMs with Gaussian mixtures for acoustic modeling. Both system use some sort of Speaker Adaptive Training. Apart from that, however, they differ considerably in many aspects.

Concerning language modeling, while FBK use a 4-gram LM in both decoding passes, Vocapia exploits continuous space Neural Network LM in the main decoding passes, and applies a 4-gram LM only in the rescoring stages.

Another substantial difference is the front-end for acoustic modeling. FBK adopts a conventional MFCC+derivatives 52-dimensional feature vector, that undergoes GMM-based acoustic normalization followed by a HLDA projection into a 39-dimensional feature vector. Vocapia, on the other hand, combines conventional PLP-like features with probabilistic features produced by a Multi Layer Perceptron with a bottleneck architecture, resulting in a 81-dimensional feature vector. It appears that this enriched representation is very effective in capturing significant characteristics of the speech signal, as the large difference in final performance can hardly be attributed only to the difference in the training data. As for the complexity of the acoustic models, the number of triphone models is similar, about 8K for Vocapia, and 8.6K for FBK. The FBK system uses Phonetically Tied mixture components, with a total number of 37K Gaussians shared by 6.7K tied-states. Vocapia reports a typical values of 32 Gaussians per state, but does not mention the total number of tied-states. Assuming it is also similar to that of FBK, there may be about 200K Gaussians in the Vocapia system. Beside, Vocapia uses gender-dependent models, while FBK uses a gender-independent acoustic model.

A more thorough discussion will be given in the post-workshop proceedings.

References

1. Lamel, L., Gauvain, J.L., Adda, G., Barras, C., Bilinksi, E., Galibert, O., Pujol, A., Schwenk, H., Xuan, Z.: The LIMSI 2006 TC-STAR EPPS Transcription Systems. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 997–1000 (2007)

[1] Note that, after the evaluation took place, FBK realized that a mistake was made when performing the evaluation run, because a wrong value for an option was given to the system. The real performance is therefore better than what appears in the official table, and is reported in the participant report. However, the general observations presented here still apply.