The FBK ASR system for Evalita 2011

Ronny Ronny[1], Aamir Shakoor, Fabio Brugnara, and Roberto Gretter

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

Abstract. This report describes the system developed in FBK for participating in the Evalita 2011 evaluation campaign, providing some details on the techniques included in the transcription system.

Keywords: Automatic Speech Recognition, Large Vocabulary, Constrained transcription, acoustic normalization, language model adaptation

1 Introduction

FBK participated in the Evalita evaluation campaign with the objective of setting a baseline that validates the distributed data. It participated in both tasks *transcription* and *contrained transcription* in *closed* training modality.

2 System Description

The FBK Evalita transcritpion system is based on several processing stages. The run-time architecture is the same for both tasks, the only difference being a special LM for the *constrained* case, as described in the following.

The first stage is Voice Activity Detection, which is based on an energy-based segmenter, providing the initial segmentation and discarding possible long silence segments. The active segments are then classified by a GMM classifier into several classes, including noise, music, male speech, female speech and so on. The classification may refine the initial segmentation. Inside each class, segments are then clustered by means of a BIC-based agglomerative clustering algorithm. These clusters will be the target for the following stages of both unsupervised and supervised acoustic normalization.

Then, a sequence of feature vectors including 13 MFCC along with 1^{st}, 2^{nd} and 3^{rd} derivatives is computed, applying average normalization on a segments basis. Cluster-based CMLSN normalization (Giuliani *et al* [1]) is performed on this sequence with respect to a GMM target with 1024 Gaussian components.

The output of the normalization is projected into a 39-dimensional feature space by means of an HLDA transformation (Kumar and Andreou [4], Stemmer and Brugnara [2]).

[1] Ronny developed the Evalita system during his summer student internship at FBK

On this sequence, the first decoding is performed with acoustic models trained on GMM-normalized data and a 4-gram language model. The output of this first step is used as a supervision for a stage of supervised CMLSN normalization, with respect to a set of simple target models (Stemmer *et al.* [3]). The sequence of normalized vectors is then fed to a second decoding step that produces the final output.

The overall processing time on the development data is around 3×RT on a single core Xeon CPU at 2.27GHz. In the actual run, however, the system exploits parallelization with a load-balanced dispatching of segments across several decoder instances.

2.1 Lexicon and Phonetic Alphabet

The lexicon is the one provided with the development data. It is composed partly by hand-written phonetic transcriptions, and partly by transcriptions that were generated by an automatic rule-based system. The phonetic alphabet is derived from the Sampa phonetic alphabet. There are 48 phonetic units, including 18 units that are geminate variants of basic phonemes. In addition, 16 filler units are used to model several non-speech phenomena including, beside silence and noise, several filler sounds that commonly occur in spontanoeous speech. These filler units are labeled in the provided phonetic annotation files, and were placed by an automatic alignment procedure during the preparation of the training data.

2.2 Language Model

The language model applied in both decoding step is a 4-gram language model trained on the files provided as normalized texts (`.ntxt`) in the development data. Vocabulary size is ≈67K words.

For the *constrained transcription* task, this LM was adapted with mixture adaptation to the provided normalized text of the minute, and recognition was performed as for the other task.

2.3 Acoustic Model

As explained above, the system uses two acoustic models, onr for the first stage and one for the second stage. Both have a similar complexity and are trained according to the same procedure, with the only difference that the model of the first pass process data after unsupervised GMM-based CMLSN, while the model for the second pass process data after supervised CMLSN performed with simple triphone models.

Acoustic units are cross-word triphones, represented by three-state left-to-right HMMs. HMM states are shared across models according to a Phonetic Decision Tree. On the given data, both models had ≈8.7K HMMs, built out of a set of ≈6.7K tied-states, for a total of ≈37K Gaussians. Beside state tying,

Gaussians are shared across mixtures according to a Phonetic Tying scheme. These means that all the states of the allophones of a certain phoneme share components from the same phoneme-dependent pool of Gaussians, and therefore differentiate among themselves only through the weights assigned to these components. During training, components with low weight are detached from the mixtures, so that in the end the average mixture length is about 94, even if the Gaussian pool for each phoneme includes about 1024 Gaussians.

3 Discussion

The results obtained in the official evaluation run for the two tasks, in the *Closed* modality, are as follows:

Task	WER (%)
Transcription	8.4
Constrained Transcription	7.2

However, in a post-evaluation check it was discovered that the run was performed by setting a wrong value for the language model weight. This parameter is used to balance the influence on the recognition decision between the acoustic model and the language model. It depends on many factors, such as the acoustic features, the acoustic model topology and the tying scheme. It is usually chosen by tuning on a development set. In this case, the weight was not set when running the system, so that a default value (7) was applied, that was not appropriate for the given configuration. After setting the LM weight to an higher value (10), the performance changed as reported in the following table:

Task	WER (%)
Transcription	7.5
Constrained Transcription	6.1

References

1. Giuliani, D., Gerosa, M., Brugnara, F.: Improved automatic speech recognition through speaker normalization. Computer Speech and Language, 20:107–123 (2006)
2. Stemmer, G., Brugnara, F.: Integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into Adapative Training. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse (2006)
3. Stemmer, G., Brugnara, F., Giuliani, D.: Adaptive training using simple target models. In: Proceedings of ICASSP, vol. 1, pp. 997–1000 (2005)
4. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Communication, vol. 26, pp. 283–297 (1998)