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 design, setup and results of the speech recognition task in the framework of the EVALITA campaign for the Italian language, giving details on the released corpora and tools used for the challenge. A general discussion about approaches to large vocabulary speech recognition introduces the recognition tasks. Systems are compared for recognition accuracy on audio sequences of Italian parliament. Although only a few systems have participated to the tasks, the contest provides an overview of the state-of-the-art of speech-to-text transcription technologies; the document reports systems performance, computed as Word Error Rate (WER), showing that the current approaches provide effective results. The best system achieves a WER as low as 5.4% on the released testset.

Keywords: automatic speech recognition, large vocabulary, constrained transcription, EVALITA.

1 Introduction and Motivation

This contribution reports on the motivations and the setup of the speech recognition task in the framework of the EVALITA campaign for the Italian language.

Research in Automatic Speech Recognition (ASR) has a long history [1] and, given the maturity of the field, high performance is achieved through the implementation of sophisticated systems, for example using huge language models that use prior information to constrain the hypothesized utterances. As a consequence, improving speech recognition often means to deal with large-scale tasks, although small-dictionary tasks can also be difficult; tasks characterized by spontaneous speech acquired in noisy and reverberant condition, even if based on small vocabulary, may be realistic and significant of the robustness of the investigated approach. Nonetheless, even after decades of research and many successfully deployed commercial products, the performance of ASR systems in some real-usage scenarios is behind human level performance [2].

Hence, the trend in ASR is toward increasingly complex models, with the purpose of improving accuracy in different acoustic conditions and with larger

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 274–285, 2013.

vocabularies. There have been notable recent advances in discriminative training [3], in large-margin techniques [4], in novel acoustic and language models [5]. Also, a major improvement has been made in training densely connected, directed belief nets with many hidden layers [6].

1.1 State-of-the-Art in ASR Technology

State-of-the-art ASR systems incorporate various processing layers in order to output hypotheses. The usual signal processing chain is composed by high-pass filtering, windowing, short-term spectral analysis, critical band integration and cepstral transformation [7]. Recent work has shown improvements using learned parameters for non-linear functions of the spectral values, inspired by the amplitude compression observed in human hearing [8]. The spectrum can possibly be warped through *Vocal Tract Length Normalization* (VTLN) [9]. VTLN uses statistical learning techniques to determine the maximum-likelihood warping of the spectrum for each speaker and this factor is derived from unsupervised learning.

Another common component is *Heteroscedastic Linear Discriminant Analysis* (HLDA) [10]: this transformation maps the cepstral features, typically over several neighboring frames, into observations of reduced size for the purpose of maximizing phonetic discrimination.

The resulting features are then used to train a set of Hidden Markov Models (HMM) that are used to generate likelihoods for particular speech sounds in the different phonetic contexts. The popular model is based on mixtures of Gaussians that are trained with the popular Expectation-Maximization algorithm using a Maximum Likelihood (ML) criterion [11].

Other objective functions are typically used to train the Gaussian parameters discriminatively [3]. Discriminative training attempts to optimize the correctness of a model by formulating an objective function that in some way penalizes parameter sets that are liable to confuse correct and incorrect answers. Many discriminative training schemes have been proposed based on different objective functions such as Maximum Mutual Information (MMI), Minimum Word Error (MWE) or Minimum Phone Error (MPE). Recently, many attempts have been made to incorporate the principle of large margin (LM) into the training of HMMs in ASR to improve the generalization abilities [12]: significant error rate reduction over the traditional discriminative training on the test sets has been observed on both small vocabulary and large vocabulary continuous ASR tasks. The parameters of the resulting acoustic model are then altered further by incorporating methods for adaptation, for instance Maximum a Posteriori (MAP) [13] or Maximum-Likelihood Linear Regression (MLLR) [14]. The resulting acoustic likelihood is then used in combination with a language model probability, which has been trained on a large quantity of written text. The interpolation coefficients between language and acoustic level likelihoods are also optimized and finally, the recognizers usually incorporate multiple contrastive systems that combine their information at various levels [15,16].

MLP techniques developed for computing discriminant emission probabilities of HMMs have been recently proposed to derive features useful for phonetic