

Complementary System Combination and Generation for ASR

Mark Gales

Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK.
mjfg@eng.cam.ac.uk

Abstract

Large Vocabulary Continuous Speech Recognition (LVCSR) systems often use a multi-pass recognition framework where the final output is obtained from a combination of multiple acoustic models. By using a combination of multiple the final error rate is usually lower than any of the individual systems. There are two important issues that must be addressed when combining multiple systems together. The first is how the various "scores" from the individual models should be combined together. The standard approaches are to align and combine either the 1-best hypothesis (ROVER) or Confusion Networks (CNC). Alternatively it is possible to combine at the acoustic model likelihood level, for example, in product framework (PoE) or a simple linear combination.

The second issue that needs to be addressed is the selection of the systems that are to be combined together. The systems should be complementary to one another. That is, the errors made by one system should not be made by the second, or subsequent, systems. Approaches to obtaining complementary systems can be split into two broad classes. The first class may be considered as "random" selection. LVCSR acoustic models are based on a large number of design choices. These range from the front-end used, to topology and context-span of the HMMs, to the design of the decision tree, to the choice of the segmentation and clustering algorithms for tasks like Broadcast News Transcription. The simplest, and currently most successful, approach to selecting systems for combination is to choose a set of configurations from the space of all possible models and then see which systems are complementary to one another on some development data. Though this approach has been successfully applied, it requires the training of a large number of systems with no guarantee which will work well.

An alternative approach is to generate systems that are explicitly trained to be complementary to one another. These class of techniques is popular in the Machine Learning community, for example the ADABOOST algorithm and extensions. However, to date, successful application of these schemes to LVCSR have been limited. An exception to this is the "code-breaking" framework. Here a standard LVCSR system is used in an initial decoding pass and then confusions are resolved using classifiers specifically tuned to that set of confusions.

This talk will give an overview of existing combination schemes and approaches for building complementary systems. The assumptions and motivations for the schemes will be described as well as possible future directions. Examples of how these approaches have been applied to a range of LVCSR tasks, including Broadcast News Transcription and Conversational Telephone Speech Transcription, will be given.