TWO DECADES OF RESEARCH ON ASR

Isabel Trancoso

INESC ID Lisboa / IST-UTL
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt/

ABSTRACT

Last May, the IEEE Signal Processing Magazine published a special issue on spoken language technology that was followed by a companion special issue of the IEEE Transactions on Audio, Speech and Language Processing, in September, on new approaches to statistical speech and text processing. Several times in these companion issues, the recent advances in automatic speech recognition (ASR) are described as "dramatic". As a consequence of these advances during the past decade, the performance is now good enough to stimulate new research areas involving the cross-fertilization of the written and spoken language processing communities. The new areas span spoken language understanding, spoken dialogue systems, spoken document retrieval and summarization, information extration from speech, and speech-to-speech machine translation, among several other exciting topics.

Much of this progress would not have been possible without the availability of large linguistic corpora which, coupled with significant advances in machine learning, makes it possible to apply statistical techniques to both spoken and written language processing. The existence of benchmark evaluation campaigns conducted by the US National Institute for Science and Technology (NIST), and in a smaller scale by some European projects, has also been recognized as a strong drive of research progress.

This maturity may lead us to wonder if we have indeed solved the major problems in ASR. The answer to this question can be derived from counting the number of papers on ASR submitted to major speech conferences such as Interspeech in the last years. Since the percentage is still almost 30% (statistics over Interspeech 2007), it is clear that the community is far from considering this a solved problem.

A very significant percentage of the accepted papers (21%) deals with what we could almost call standardized corpora in English such as TI, TIMIT, AURORA, and describes experiments that could be easily be generalized to other languages. English is in fact the most widely used language for the corpora used in the remaining papers (30%), with Japanese (13%), Chinese (7%), French (5%), and Arabic (4%), following next, and the remaining languages totalling only 20%. The importance of widely available corpora at a national level can be seen from the significant percentage of Japanese-language papers using CSJ (38%), and French papers using ESTER (33%). Hence multilinguality remains an issue, but definitely not the only challenge in ASR.

Of all the major blocks in speech recognition systems, acoustic modeling gets the largest share of current research efforts (21%), motivating specific sessions on discriminative training, attribute transcription, template-based and structure-based recognition, etc. This percentage does not include the large number of papers dealing with robustness to noise and reverberation (20%, also including voice activity detection and feature extraction), nor the papers focusing on adaptation to specific speaker characteristics (8%). The other half of papers is mostly devoted to lexical and prosodic modeling, language modeling, confidence measures, fusion of speech and other modalities, and large vocabulary approaches.

A growing area is the extraction of meta-data from the audio signal, in terms of domain and topics of utterances, context, semantics, speaker diarization, etc. Although this area is now included in separate sessions outside the ASR topic, the extracted meta-data constitute very valuable knowledge sources, and incorporating knowledge sources into the statistical ASR framework is one important step towards approaching human speech recognition.

This paper attempts to summarize the so-called "dramatic" advances in ASR and list the current research topics and challenges in this area. We shall try to do this by looking at the evolution of the topics of papers on ASR through the last two decades.