Connectionist mixture of experts and auditory-based parameters for a better identification of complex phonetic features

Sid-Ahmed Selouani and Douglas O’Shaughnessy
INRS-Télécommunications, Université du Québec, Canada
selouani@inrs-telecom.uquebec.ca, dougo@inrs-telecom.uquebec.ca

One of the interesting challenges that practical Automatic Speech Recognition (ASR) is faced with is to unify acoustic processing and to adapt the architecture of ASR systems to cover the broadest range of language particularities. Unfortunately, the training procedures of most ASR systems yield a rigid modeling and are individually devoted to a specific language. Hence, interesting questions impose themselves: is it possible to adapt a system dedicated to English, for instance, to another language such as Serbo-Croatian, Czech or Mandarin where tone, pitch or phoneme lengthening is semantically relevant? How would a system react when it is confronted with the discrimination of emphatic/non-emphatic opposition as is frequently encountered in Arabic, Hebrew and other Semitic languages? The monolithic approach adopted by the classical ASR systems seems to be not adapted to ‘perceive’ these features. Indeed their architecture is generally compact and frontally tackles the global recognition task, which considerably limits their performances. The approach we propose intends to ‘boost’ the performances of a modular ASR structure in the cases of the complex phonetic features such as gemination, emphasis and relevance of phoneme duration. Our solution consists of placing a hierarchical structure of neural experts downstream in a baseline ASR system. These experts are typically time delay neural networks using a version of an autoregressive back propagation algorithm. Both implementations of experts build into their respective recognition and model-training algorithms a mapping process between the acoustic and the phonetic/phonologic feature space through the use of auditory-based cues. Pertinent indicative features are extracted and constitute a computational interface between the phonological process and the hearing process. This interface is expected to optimize a parsimonious set of model parameters in order to accurately characterize the symbolic, dynamic and static components in speech. The combination of structural adaptation of ASR architecture and the inclusion of hearing/perception knowledge permit a joint characterization of the contextual and speaking-style variations manifested in speech acoustics. Our experiments show that feature representations which include fundamental frequency, zero-crossing and auditory-based cues minimize the ASR error rate and are most suited to be used in combination with the proposed mixture of sub-neural-network experts. The performance improvement reaching 5% over 32,000 phonemes, considering the results obtained by a baseline system using classical Mel-Frequency Cepstral Coefficients and their first and second derivatives, promises to overcome some key limitations of current speech recognizers.