Labeling in the Wild: Crowdsourcing versus Categorical Perception

Mark Hasegawa-Johnson, Jennifer Cole, Preethi Jyothi and Lav Varshney

Modern speech technology and speech science depend on large labeled corpora. These large corpora are collected under three assumptions: speech is perceived in terms of discrete categories; there is consistency within and across labelers in the categories that are perceived in a given speech sample; a language can be reliably labeled only by a first-language speaker (and not by non-natives). Machine learning requires labeled corpora, and labeled corpora require technologically literate labelers, therefore any language without a large paid population of such labelers finds itself ignored by speech technology. Recent machine learning theory provides three mechanisms by which we might alleviate the growing speech technology gap between well-resourced and under-resourced languages. First, active learning is a set of methods in which the machine learning algorithm plays an active role in its own education, e.g., by selecting the speech waveforms whose labels would be most informative. In the best case, theoretical guarantees show that active learning can reduce the error rate of a trained classifier from 1/n to exp(-n), where n is the number of labeled data samples. Algorithms for active learning can be tested on well-resourced languages, and having been proven on known tasks in known languages, can then be deployed in order to rapidly develop technology in under-resourced languages. Second, methods of crowdsourcing allow speech labeling tasks to be distributed to a large number of anonymous labelers, some of whom may be less reliable than others. Finally, we can consider the possibility of mismatched crowdsourcing: the acquisition of speech labels from labelers who are not native speakers of the language under study, or by native speakers who lack adequate linguistic training to perform a desired labeling task. Mismatched labeling is a kind of lossy communication channel, according to which some of the information in the original signal has been systematically deleted by the untrained ears of the labeler. As in the case of any other lossy channel, the problem can be solved by a low bit-rate auxiliary channel: in this case, a small amount of data transcribed by a trained expert native speaker may be enough to recover all of the information lost by the mismatched crowdsourcing experiment. In the paradigms of active learning and matched crowdsourcing, theoretical performance guarantees specify upper bounds on the error rates that result from labeling with any given number of labelers. The paradigm of mismatched crowdsourcing has been much less heavily studied, either in theory or in practice, but work in multilingual speech recognition, rapid prosody transcription, and second-language pronunciation error detection suggests the possible shape of future results.