
Abstract

A dialect identification technique is described that takes as input extemporaneous, conversational speech spoken in Latin American Spanish and produces as output a hypothesis of the dialect. The system has been trained to recognize Cuban and Peruvian dialects of Spanish, but could be extended easily to other dialects (and languages) as well. Building on our experience in automatic language identification, the dialect-ID system uses an English phone recognizer trained on the TIMIT corpus to tokenize training speech spoken in each Spanish dialect. Phonotactic language models generated from this tokenized training speech are used during testing to compute dialect likelihoods for each unknown message. This system has an error rate of 16% on the Cuban/Peruvian two-alternative forced-choice test. We introduce the new "Miami" Latin American Spanish speech corpus that is capable of supporting our research efforts into the future.