Abstract

The availability of a standard and phonetically rich speech corpus provides a common platform for comparing the performance of different speech recognition approaches and therefore is the first step for the research in a language. This work presents the development of a large multilingual speech corpus of Pashto, Urdu and English. Recordings have been made from a total of 194 speakers in the three languages, covering diverse dialects, age groups, genders and professions. Pashto and Urdu both native and non-native speakers have been considered while for English, all the speakers were non-native. The corpus comprises of three categories of phonetically rich spoken data in each language, that is, short questions regarding speaker’s personal information, read speech and spontaneous speech from the domain of tourism. Although the corpus is developed primarily for research on Automatic Spoken Language Identification purpose, nevertheless, it can also be used for research on other topics such as Automatic Speech Recognition, Accent Recognition, Automatic Speaker Identification and the study of effects of non-nativeness on Language and Speaker Identification.
References


5. “LDC - Linguistic Data Consortium” [Online]. Available at: https://catalog.ldc.upenn.edu/LDC93S1


Index Terms

Computer Science

Information Sciences
Keywords