Speech Recognizer Training – How Much Manual Segmentations Do We Need?

Péter Mihajlik, Péter Tatai, Géza Gordos
Department of Telecommunications and Media Informatics, BUTE
mihajlik@tmit.bme.hu

Keywords: Automatic speech recognition, speaker-independent telephone speech recognition, training, speech database, transcriptions, segmentations

Today's speaker-independent automatic speech recognizers require hundreds of hours of training speech data. The basic units of recognition are typically the speech sounds; therefore these units and their positions have to be identified in the database. This process is called phonetic segmentation and can be performed either manually or automatically. While the worldwide approach is to perform the (implicit) phonetic segmentation entirely automatically, sometimes a big amount of manual segmentations are made, see e.g., the recently collected HTSD - Hungarian Telephony Speech Database.[1]

We conducted some experiments on one of the mentioned database (HTSD) in order to determine the optimal approach to database development towards increasing the efficiency of speech recognizers. First we trained the ASR system based on the manual phonetic segmentations of the 500 speakers. Context dependent (CD) phone models were used. The recognizer was tested on independent test data; the task was isolated word recognition with a vocabulary size of 1000. The best recognition error rate we could achieve was 6.85%. Then we trained the recognizer based on the segmentations of only 10 speakers by a sophisticated training method developed at our laboratory. The process needed only the annotations and the waveforms of the 500 speakers additionally. Phonetic transcriptions containing pronunciation alternatives were generated automatically based on phonological rules[2]. These special transcriptions were applied at the forced alignment phase where the actual phonetic realizations and the phone boundaries were determined automatically. So, effectively, the human listener was replaced by the computer. Finally, we obtained an error rate of 7.02% on the reference recognition task.

We can conclude that using our training method that utilizes deeper phonological knowledge the recognition error rate is practically the same as in the case of fully manual segmentation. At the same time the required amount of the highly expensive manual segmentations was reduced by *50 times*.

[1] http://alpha.ttt.bme.hu/speech/MTBA.htm

[2] Mihajlik, P., Révész, T. and Tatai, P., Phonetic transcription in automatic speech recognition, Acta Linguistica Hungarica, Vol. 49, pp. 407–425, 2002