Leszek Gajecki, Marek Jaszuk, Teresa Mroczek, Leszek Puzio, Grażyna Szostek

University of Information Technology and Management (Rzeszów)

Deep Neural Networks in Acoustic Modelling for Polish Language Speech Recognition

The main objective of our research was to develop an efficient acoustic model, to be used in telephonic automatic speech recognition system (ASR). The research was conducted on LUNA - the Polish dialogue corpus, consisting of call center recordings. This kind of data is characterized by low quality of sound, spontaneous talk without grammatical correctness, emotions during conversations.

The focus of our research was put on exploring different models of deep neural networks and their combinations. We used eleven network configurations, on which 139 tests were performed, to verify the impact of particular topologies and training/decoding parameters on recognition efficiency. To evaluate the performance of the ASR system, we used the word error rate (WER) metric. The research was conducted on the unigram and bigram language model (LM). The results for the both models are presented in the table.

Acoustic model architecture	Acoustic model	unigram LM	bigram LM
CNN&TDNN&BLSTM	tri3a	38.35	22.98
CNN&TDNN&LSTM&BLSTM	tri4a	39.08	24.44
CNN&TDNN&GRU	tri3a	37.52	22.95

The novelty of our approach is applying CNN networks along the time axis to reduce temporal variability, which has proven to be successful in normalizing speech spectral features. The final model is more resistant to small disturbances occurring simultaneously in the time and frequency input spaces, because the extracted local signal patterns cross the dimensions of time and spectrum.

The obtained WER results are relatively high, in comparison to other reports. This is, however, justified, considering the difficulties met in the considered problem. The system had to handle many speakers. An efficient recognition of Polish language requires a bigger dictionary than the equivalent task in English language. Polish is a highly inflectional language. Free word order also makes it difficult for the n-gram language models. The quality of speech recognition could be improved if bigger speech data sets were used. However, a significant improvement could be expected, if the size of the speech transcripts database would be several times as large as we used.

The publication is a result of the project: "IVA service platform of virtual voice agents to automate the operation of emergency call centers", co-financed under Measure 1.2 Research and development activities of enterprises, Regional Operational Program of the Masovian District (Poland) for the years 2014-2020, grant RPMA.01.02.00-14-5746/16.