Deep Neural Networks in Acoustic Modelling for Polish Language Speech Recognition

Leszek Gajecki, Marek Jaszuk,
Teresa Mroczek, Leszek Puzio, Grażyna Szostek
University of Information Technology and Management
Rzeszów
{lgajecki,tmroczek,mjaszuk,lpuzio,gszostek}@wsiz.edu.pl

Acknowledgement: ”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.

Motivation

- 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 language of written text is formal, with correct grammar rules. The language of telephone speech is informal, with simplifications like informal words, grammar with more freedom in comparison to written text.

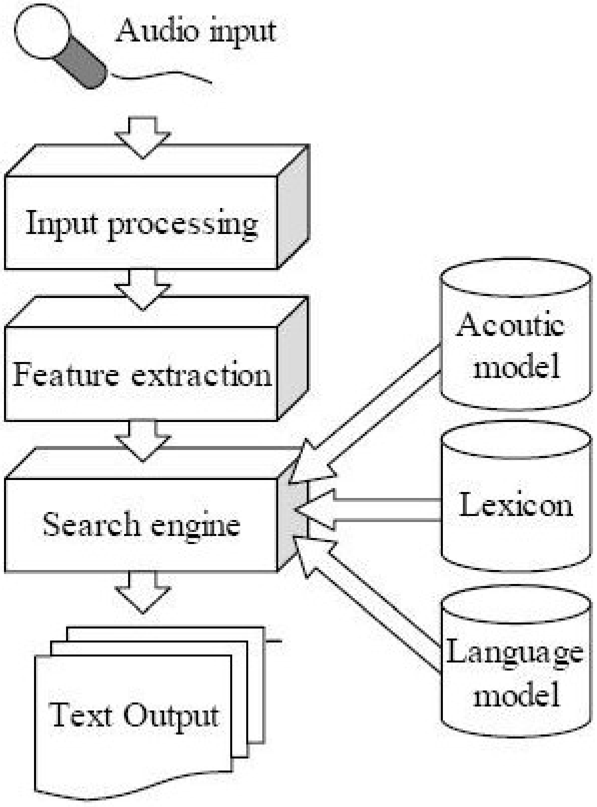
Cooperation with

- Poznań Supercomputing and Networking Center (PCSS), Bioorganic Chemistry Institute- Polish Academy of Science
- Emerline sp. z o.o (formely Haxon Telecom)

LUNA corpus

- We used set called LUNA, created as a result of the project “spoken Language UNdestarding in multilinguAl communication systems”.
- We took 562 speech recordings, which last about 13 h 20 min. Each conversation contains:
  - WAV file with recording,
  - TRS file with transcription, which contains data about speakers, utterances’ transcriptions, and durations.
- LUNA statistics: utterances 12 788; words 81 049; dictionary size 7 768; male speakers 255, female speakers 276

Speech recognition system



Speech recognition task

- The aim – find the optimal word sequence:

$$W_{opt} = argmaxP\left(W_1^k \middle| X\right)$$

- The cost of hypothesis:

$$f\left(X, W_1^k\right) = AClogP\left(X \middle| W_1^k\right) + logP\left(W_1^k\right) + CC$$

- where:
  - Acoustic model: $P\left(X \middle| W_1^k\right)$
  - Language Model: $P\left(W_1^k\right)$

Language model

- The acoustic model gives the probability of respective acoustic units (e.g. triphones).
- The longer units, i.e., words, are constructed from acoustic units using lexicon.
- Such module is created once, and it is not trained. More precisely, the lexicon is not yet a final structure, which can be applied in recognition process.
- Such structure is a decoding graph, which is constructed using lexicon. It contains all possible paths, in which nodes are acoustic units, that constitute words.
- Additionally, for nodes representing words, there is superimposed language model.

Language model probability

- The language model gives probability of a current word given previously occurred (recognized) words:

$$P(W_i) = P\left(W_i \middle| W_1^{i-1}\right)$$

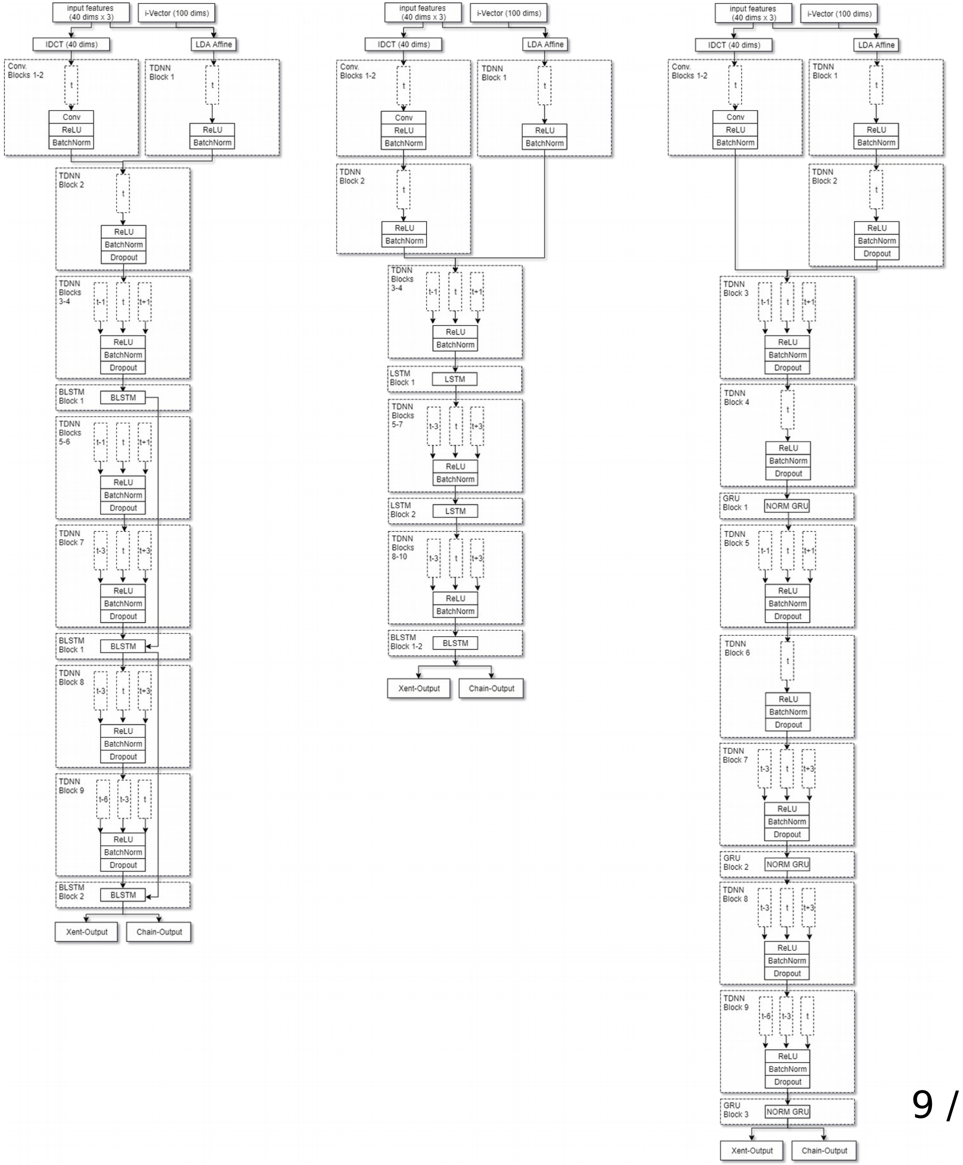
- where needed probability of recognized word sequence is:

$$P\left(W_1^k\right) = \prod_{i=1}^{k} P(W_i) = \prod_{i=1}^{k} P\left(W_i \middle| W_1^{i-1}\right)$$

- One of the simplest models are n-gram models, where the length of context is limited to n:

$$P(W_i) = P\left(W_i \middle| W_{i-n}^{i-1}\right)$$

Most effective Neural networks



Results

- 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).

Results for the unigram and bigram models

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

Results discussion

- The obtained WER results are relatively high because:
  - the system had to handle many speakers,
  - Had bigger dictionary than the equivalent task in English,
  - Polish is a highly inflectional language. Free word order also makes it difficult for the n-gram language models.
- 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 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.

Summary

- 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.