Assessing the Performance of the TDNN-BLSTM Architecture for Phoneme Recognition of English Speech

Delft University of Technology

Irene Klom – i.a.klom@student.tudelft.nl Supervisor: Dr. Siyuan Feng Responsible Professor: Dr. Odette Scharenborg

1. Background

- Automatic Speech Recognition has its limitations.
- The **TDNN-BLSTM** architecture improved the Phoneme Recognition for Dutch speech.[1]
- This architecture has not been tested on English speech.

/hat/ vs. /kat/
/seɪk/ vs. /ˈsɑːki/

2.Research questions

How does the **TDNN-BLSTM** architecture perform on English **read** and **spontaneous** speech?

Comparing results **quantitatively**:

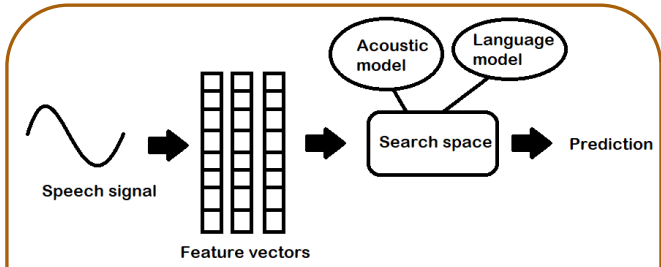
- What is the Phoneme Error Rate for TDNN-BLSTM on spontaneous and on read speech?

Comparing results **qualitatively**:

- What phonemes have a large PER difference between read and spontaneous speech?

3. Methodology

- Using the **Timit** and **Buckeye** corpora
- Preparing the data for Kaldi framework
- Training and testing the TDNN-BLSTM model
- Focus on layer dimension, epochs and learning rate
- Evaluating results with **PER** metric
- Evaluating results based on confusion matrix



Phoneme Recognition System

4.Results

31.78% PER on read speech
54.03% PER on spontaneous speech

Phonemes in read speech are recognized better than in spontaneous speech

Comparing to research on Dutch PR[1], parallel research[2] and literature values[3]:
TDNN-BLSTM does not perform as well as other acoustic models for both spontaneous and read speech.

Some terms explained:

TDNN-BLSTM:
Projected Bidirectional Long Short-Term Memory Time Delayed Neural Network

---

Read speech: (Timit corpus)
Participants read predetermined sentences

Spontaneous speech: (Buckeye corpus)
Participants have an informal conversation

PER:
Phoneme Error Rate

[1] R. Levenbach, "Phon times: Improving Dutch phoneme recognition," Master's thesis, 2021.
[2] J. van der Tang, "Evaluation of phoneme recognition through TDNN-OPGRU on Mandarin speech." 2021; G. Genkov, "Training and testing the TDNN-OPRGU acoustic model on English read and spontaneous speech." 2021; M. Chiroşca, "Evaluating the performance of the TDNN-BLSTM on Mandarin read and spontaneous speech." 2021.
[3] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition,"IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, pp. 92–102, Apr. 2018;
R. Qader, G. Lecorvé, D. Lolive, and P. Sébillot, "Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features," in International Conference on Statistical Language and Speech Processing (SLSP), (Budapest,Hungary), pp. 229–241, Nov. 2015.

[1] R. Levenbach, “Phon times: Improving Dutch phoneme recognition,” Master’s thesis, 2021.
[2] G. Genkov, “Training and testing the TDNN-OPRGU acoustic model on English readand spontaneous speech.” 2021.
[3] M. Chiroşca, “Evaluating the performance of the TDNN-BLSTM on Mandarin read and spontaneous speech.” 2021.
[4] J. van der Tang, “Evaluation of phoneme recognition through TDNN-OPGRU on Mandarin speech.” 2021.
[5] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, “Light gated recurrent units for speech recognition,”IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, pp. 92–102, Apr. 2018.
[6] R. Qader, G. Lecorvé, D. Lolive, and P. Sébillot, “Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features,” in International Conference on Statistical Language and Speech Processing (SLSP), (Budapest,Hungary), pp. 229–241, Nov. 2015.