Training and Testing the TDNN-OPGRU acoustic model on English Speech

1.Background

- Automatic phoneme recognizers (APR) can recognize separate **phonemes** from speech. This removes the limitation of conventional speech recognizers which have a finite dictionary of words.

- The research question is “What is the Best Automatic Phoneme Recognition System?”. In particular I should evaluate the performance of the **TDNN-OPGRU** model on two English **corpora** - one for prepared speech (TIMIT) and one for spontaneous (Buckeye).

2. Methodology

To evaluate **TDNN-OPGRU** on the two corpora 4 main steps need to be completed:

- Process the datasets into a format understood by the speech recognition software **Kaldi.**
- Configure the acoustic model for **TDNN-OPGRU**, train and test it on either prepared speech. Configuration from Robert Levenbach [1].
- Adjust the parameters of the acoustic model to perform better, train and test it one both corpora.
- Evaluate the results and compare with peers’ findings.

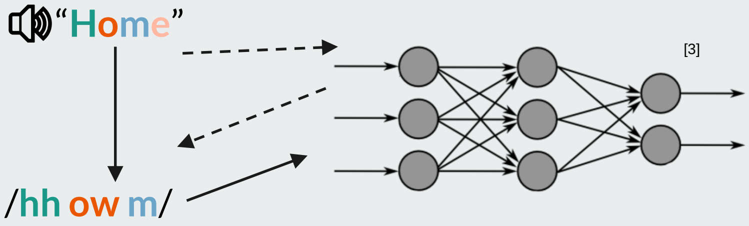

Figure 1: Transcribing spoken words into phonemes and training a NN to recognize them (solid arrows). After training a recording goes through the NN to recognize the phonemes (dotted arrows).

3. Results

- Initial configuration:

7 TDNN layers (dimension: 1024); 3 OPGRU layers (dim: 512); Initial and final learning rate: 0.001 and 0.0001; 6 epochs

Results: 32.57% **PER** for TIMIT. Configuration was modified to achieve “Baseline” results:

- Further modifications to the epochs and learning rate improved the **PER** slightly[1].

7 TDNN layers (dimension: 256); 3 OPGRU layers (dim: 128); Initial and final learning rate: 0.005 and 0.0005; 10 epochs

	TIMIT	Buckeye
Baseline conf.	31.55%	52.21%
Final conf.	25.98%	49.31%

Figure 2: PER for the baseline and final configuration

	TIMIT	Buckeye
Substitutions	67.95%	61.9%
Insertions	9.35%	4.1%
Deletions	22.68%	33.95%

Figure 3: Contributions to the PER for TIMIT and Buckeye.

Georgi Genkov
G.Genkov@student.tudelft.nl

Responsible Professor: Odette Scharenborg
Supervisor: Siyuan Feng

4. Conclusion

- Main limitations are the timespan, processing power and the use of subsets instead of the whole corpora.
- The results are worse than many previously researched acoustic models [2] but very comparable with the parallel research.
- The results are consistent with previous comparisons between TDNN-OPGRU and TDNN-BLSTM [1].

[1] Levenbach, R. (2021). "Phon Times: Improving Dutch phoneme recognition".
[2] van Geffen, et al. (2019). "A review of deep neural network-based phoneme recognition systems”.
[3] Image adapted from Wikimedia Commons: https://commons.wikimedia.org/wiki/File:MultiLayerNeuralNetworkBigger_english.png

[1] Final PER is corrected for the insertions and deletions of silences.