Supervisor: Dr. Siyuan Feng
Responsible Professor: Dr. Odette Scharenborg

# Evaluation of phoneme recognition through TDNN-OPGRU on Mandarin speech

J. van der Tang BSc j.vandertang@student.tudelft.nl

TU Delft Delft University of Technology

## 1. Background

- **Automatic Phoneme Recognition** (APR) is a form of Automatic Speech Recognition (ASR).
- APR focuses on recognizing phonemes rather than words.

- APR systems are able to identify phonemes even when the words that are spoken were not part of the training data.

- APR can be useful for a variety of tasks including the identification of mispronounced phonemes and aiding people with a speech impediment[1]

## 4. Results

- A higher learning rate together with a large layer size produced the lowest Phoneme Error Rate (**PER**[6]).

- A reduced layer size results in a greater performance drop with tone information than without.

| data set | PER |
|---|---|
| Prepared speech | 39.99 |
| Prepared speech no tones | 29.34 |
| Spontaneous speech | 30.76 |
| Spontaneous speech no tones | 23.27 |

Table 2: Achieved PER on different types of speech

- TDNN-OPGRU performs better on spontaneous speech than on prepared speech.

- TDNN-OPRU performs better without tone information.

- The removal of tone information has a similar impact on the PER for both prepared and spontaneous speech.

| data set | Tone error rate |
|---|---|
| Prepared speech | 27.48 |
| Spontaneous speech | 20.42 |

Table 3: Tone error rate for prepared and spontaneous speech

| data set | substitution % |
|---|---|
| Prepared speech | 25.94 |
| Spontaneous speech | 25.99 |

Table 4: Percentage of substitutions caused by tone-only errors

- Tone error is higher in spontaneous speech than in prepared speech.
- The percentage of errors that are tone-only is similar between the types of speech.

| Type of speech | Error-prone phonemes | | | | | | |
|---|---|---|---|---|---|---|---|
| Prepared speech | N | UW | NG | ER | AH | AW | AE |
| Spontaneous speech | N | UW | NG | ER | AH | D | AA |

Table 5: Error-prone phonemes for prepared and spontaneous speech

- Five phonemes are **error-prone**[7] for both prepared and spontaneous speech.

## 2. Problem

- Recent research has presented **TDNN-BLSTM**[2] and **TDNN-OPGRU**[3] as the best performing networks for prepared speech and spontaneous speech respectively for Dutch[4].

- Mandarin could pose different issues for an APR system because it is a tonal language.

- Evaluating the network on Mandarin will provide more  insight into the general performance ofthe networks.

| Tone | pinyin | Translation |
|---|---|---|
| 1 | mā | mom |
| 2 | má | hemp |
| 3 | mǎ | horse |
| 4 | mà | scold |

Table 1: Tone example in Mandarin

- **Goal:** To investigate the performance of the TDNN-OPGRU architecture when decoding phonemes in Mandarin prepared and spontaneous speech.

## 5. Discussion

- The difference in PER between prepared and spontaneous speech is unexpected[8].
- There are several aspects of the research setup that could have contributed to this result: A difference in the amount of speakers, data preparation and gender distribution in the training set.
- However, similar results were obtained in the research of a colleague with the TDNN-BLSTM network.
- TDNN-OPGRU appears to perform worse than on Dutch. It performs better on spontaneous speech in Mandarin, and better on prepared speech in English.
- There is very little overlap between error-prone phonemes in Mandarin and Dutch[4], indicating how that TDNN-OPGRU has difficulties with different aspects.

## 6. Conclusion

- The TDNN-OPGRU architecture obtains a PER of 39.99% on prepared speech and a PER of 30.76% on spontaneous speech.

- The phonemes **N, UW, NG, ER** and **AH** are error-prone phonemes when decoding with the TDNN-OPGRU architecture.

- Tone errors make up a substantial amount of the errors during decoding, but do not impact the difference in PER between prepared and spontaneous speech.

## 3. Method

1. Prepare data for phoneme recognition

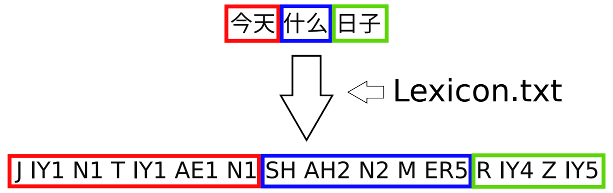

Figure 1: Using the lexicon to replace characters with phoneme sequences

2. Extract features and generate forced alignments

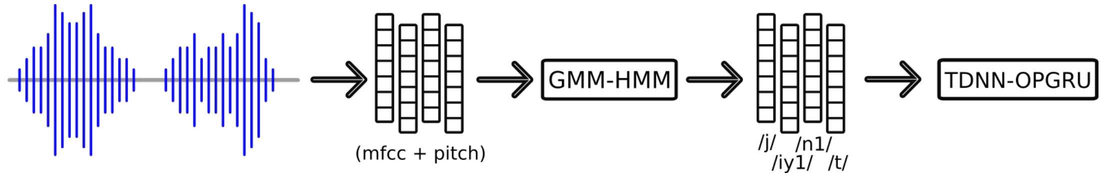

Figure 2: Using a GMM-HMM[5] to create forced alignments for TDNN-OPGRU training

3. Optimize neural network on prepared speech development set

4. Train and evaluate neural network test sets for prepared and spontaneous speech with and without tone information

## 7. Future work

- A better PER could be obtained by testing the TDNN-OPGRU network with larger training sets
- The impact of the imbalance in gender distribution on the prepared speech should be investigated.
- More value can be gained when looking at tone and base phonemes combined rather than separated[9].

## 8. Footnotes

1- . J. Witt S. M. & Young, “Phone-level pronunciation scoring and assessment forinteractive language learning,”Speech communication, vol. 30(2-3), pp. 95–108, 2000.
2- Time-Delayed Bi-directional Long Short Term Memory Recurrent Neural Network
3- Time-Delayed Open-Gate Projected Recurrent Unit Recurrent Neural Network
4- R. Levenbach, “Phon times: Improving dutch phoneme recognition,” M.S. thesis, DelftUniversity of Technology, Delft, 2021.
5- Gaussian Mixture Model- Hidden Markov Model
6- Phoneme Error Rate - Combined substitution,deletion and insertion errors divided by the amount of phonemes in the ground truth
7- Error-prone phonemes are phonemes with an above average PER, and an above average contribution to the PER, which is the errors made on this phoneme divided by the total amount of errors.
8- R. Dufour, “From prepared speech to spontaneous speech recognition system: A comparative study applied to french language,”CSTST’08, Cergy-Pontoise, France: Association for Computing Machinery, 2008, pp. 595–599
9- X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, “Improved tone modelingfor mandarin broadcast news speech recognition,”INTERSPEECH-2006, paper 1752–Tue3A2O.4. 2006.