Evaluation of phoneme recognition through TDNN-OPGRU on Mandarin speech

Jordy van der Tang BSc,
(Supervisor) Dr. Siyuan Feng,
(Responsible Professor) Dr. Odette Scharenborg

27th of June, 2021

Abstract

This research expands past research on implementing the TDNN-OPGRU network for Automatic Phoneme Recognition on Dutch speech by implementing and testing the TDNN-OPGRU network on Mandarin speech. The goal of this research is to investigate the performance of the TDNN-OPGRU architecture when decoding phonemes in Mandarin prepared and spontaneous speech. The difference in Phoneme Error Rate between prepared and spontaneous speech is being determined, and the effect that tones have on the PER is being investigated since Mandarin is a tonal language.

The results are that a substantial amount of the PER comes from substitutions that are made where only the tone is incorrectly determined. However, tone does not appear to have an impact on the difference in PER between spontaneous and prepared speech since it is responsible for an similar amount of the substitutions in both types of speech. The inclusion of tone also causes the error rate of the TDNN-OPGRU architecture on base phonemes to increase.

1 Introduction

Automatic Speech Recognition (ASR) systems convert the sound generated by speech into a sequence of discrete sentences, words or phonemes. ASR systems are widely used, well known examples are Word Recognition (WR) systems such as Siri or Google Assistant. However, a downside of WR systems is that they can not recognize words that they have not been trained on.

This is where Automatic Phoneme Recognition (APR) systems come in. A phoneme is the "smallest unit of speech distinguishing one word (or word element) from another, as the element p in 'tap', which separates that word from 'tab'" [1]. APR systems output a sequence of phonemes, even if these phonemes would not construct a word that it has been trained on. This can be useful for a variety of tasks including the identification of mispronounced phonemes [2] [3].

There are two types of speech that are defined in ASR research. Prepared speech occurs for example when reading a passage from a book, or during a rehearsed speech. Spontaneous speech occurs in unprepared situations, for example having a conversation or a spontaneous discussion with a friend. In general spontaneous speech is harder to process for ASR systems than prepared speech. Some causes for this include unexpected pauses, repetition and some phonemes not being pronounced properly [4]. To get a good impression of the performance

of an ASR or APR system it is therefore important to evaluate its performance on both types of speech.

In march 2021 a master thesis was published that evaluated four Neural Network architectures for use as the acoustic model of an APR system for the Dutch language [5]. From this research came TDNN-OPGRU (Time-Delayed Open-Gate Projected Recurrent Unit Recurrent Neural Network) and TDNN-BLSTM (Time-Delayed Bi-directional Long Short Term Memory Recurrent Neural Network) as two of the best performing architectures for spontaneous speech and prepared speech respectively.

Based on the results of the master thesis the TDNN-OPGRU and TDNN-BLSTM architectures are being further evaluated on Mandarin and English in several studies alongside this research. Some steps such as the data preparation will be shared between the studies to ensure that results between the studies are comparable; The main variation will be the network architecture. This paper will focus on the TDNN-OPGRU architecture and the Mandarin language.

TDNN-OPGRU is a DNN (Deep Neural Network) that combines TDNN (Time-Delayed Neural Network) layers with OPGRU (Output-Gate Projected Gated Recurrent Unit) layers. TDNN layers help capture contextual dependencies because temporal context can be added through a delay element. (OP)GRU is a simplified version of LSTM, which helps solve the vanishing gradient problem [5].

Mandarin could pose different issues for an APR system because it is a tonal language. This means that rather than changing the sound, for example from 'p' to 'b', the tone of a syllable alone can affect the meaning of a word; 马(Mǎ) means horse, but 麻(Má) means hemp [6], adding an additional distinction between meaning over the dutch language. There are four tones and a neutral tone in Mandarin [6], this increases the amount of phonemes that have to be recognized from 34 for Dutch [5] to 113 used in this research for Mandarin.

1.1 Related Research

Not much research has been done on phoneme recognition in Mandarin speech. However, research has been done on tone recognition for mandarin speech. Some research focuses on system that are designed to recognize only tone, and no other aspects of the phonemes or language [7]. Other research focuses on improving the tone recognition aspect within an ASR system [8]. It is stated that improving the tone recognition of an ASR system can improve the overall performance since the tone decision might impact the phonetic decision because of context dependence [8]. The results of these researches can be used as points of comparison with the results of TDNN-OPGRU regarding tone recognition, since tone recognition will be an aspect of phoneme recognition.

Parameters for the TDNN-OPGRU network are defined in [5] based on results from other papers. However, it does not explain what these parameters are, how they influence the network, or why these parameters are suited for this specific task. These parameters include dropout schedule, L2 regularization and layer size. Dropout is a mechanic that improves generalization of a system by randomly ignoring a certain percentage of the cells in a layer. A dropout schedule is a version of this that can linearly interpolate between different dropout values through several epochs [9]. The same dropout schedule that is described in this paper for ASR is also used in [5]. L2-regularization is a method that removes a small portion of the weights of cells at each iteration of the training. It is a technique that helps with reducing over fitting on the training data [10]. The value that is selected for this in [5] is chosen based on [10].

The required layer size of a network is partially dependent on the amount of training data that is used. Specifically, a layer size that is too large can result in under fitting and a layer size that is too small can result in over fitting [11]. This means that if the amount of training data differs, a different layer size from the one that is used in [5] might be more optimal for the network. Similarly, [11] shows that too many layers can lower the performance of a network for ASR. This indicates that these parameters have to be tuned to the specific data set and language.

1.2 Research Goals

The goal of this research is to investigate the performance of the TDNN-OPGRU architecture when decoding phonemes in Mandarin prepared and spontaneous speech. This will be done by training and evaluating the architecture on a data set containing spontaneous speech, and a data set containing prepared speech. Afterwards the Phoneme Error Rate (PER) will be analyzed to determine the differences between prepared and spontaneous speech. Phoneme error rate is the sum of all substitutions, deletions and insertions, divided by the total amount of occurrences of the phoneme in the ground truth. Substitutions are cases where one phoneme is wrongly decoded and substituted by another phoneme. Insertions are cases where no phoneme is present in the ground truth, but an extra phoneme is present in the decoding. Deletions are cases where a phoneme is present in the ground truth, but not present in the decoding.

Error-prone phonemes will also be determined and compared with error-prone phonemes identified in other languages [5]. Error-prone phonemes are phonemes that have an above average PER, as well as an above average contribution to the PER. The contribution to the overall PER is determined by dividing the amount of substitutions, deletions and insertions from one specific phoneme by the sum of all substitutions, deletions and insertions.

Lastly, [7] and [8] demonstrate that tone recognition is important for ASR on Mandarin speech. Therefore the performance of the architecture on tone recognition is important to determine. Since tone can be decoded separately it will also be useful to investigate the performance of the architecture without tone information, this can help determine if perhaps tone should be decoded separately for increased performance. To aid in completing the research goal the following research questions have been constructed:

- What is the PER when recognizing phonemes in Mandarin speech with a TDNN-OPGRU architecture?
- What is the difference in PER between prepared and spontaneous speech when recognizing phonemes in Mandarin speech with a TDNN-OPGRU architecture?
- What phonemes are error-prone when decoding phonemes in Mandarin speech with a TDNN-OPGRU architecture?
- What influence does the presence of tones have on the PER when decoding phonemes in Mandarin speech with a TDNN-OPGRU architecture?

1.3 Report Outline

In section two the methodology of the research will be discussed. Section three will then discuss the contributions of this research. In section four the setup of the experiment is described. Section five presents the different results and section six is dedicated to the ethical

aspects of the research and the reproduciblility of the methods. In section seven the results of the previous sections will be discussed and the results will be put into a broader context. Finally in section eight the paper will be concluded and future work will be recommended.

2 Methodology

This section focuses on the methodology that is applied in this paper. First the software that is used will be presented. Next, the data sets that will be used are discussed, together with the specific preparation of the data. Following this the training and implementation of the TDNN-OPGRU network will be discussed. Afterwards how the research questions will be answered is explained. Lastly, some notes on collaboration between different papers will be given.

2.1 Kaldi

The TDNN-OPGRU architecture will be implemented using the Kaldi framework [12], this is a framework that has been developed for ASR. Although APR is slightly different the framework can be used by replacing characters in the transcripts with their phoneme sequences and replacing a lexicon that normally maps words to phonemes with a dummy lexicon that maps phonemes to phonemes.

2.2 Data sets

In order to train and evaluate the network a data set is needed. In this case two separate data sets are used for prepared and spontaneous speech. The data sets that are going to be used are:

- Aidatatang_200zh : a Chinese (Mandarin) read speech corpus [13]
- Magic Data Chinese Mandarin Conversational Speech : a Chinese (Mandarin) spontaneous speech corpus [14]

The Aidatatang set, a read speech corpus, is freely available and used on a desktop computer. The speech was recorded in a quiet room, the data was collected from 34 provincial administrative regions across China. A gender ratio of about 3 males to 3.3 females is indicated [13]. The Magic Data set, a spontaneous speech corpus, is a paid version and only accessible on the TU Delft HPC [15]. The speech was recorded on mobile devices with speakers "from accent regions across the country" [14].

The network will not be trained on the complete data sets. For Aidatatang the following speakers have been selected from the training set: G0013, G0017, G0019, G0020, G0029, G0030, G0032, G0034, G0037, G0038. This results in about 217 minutes of training data. This has a gender distribution of 9 males and 1 female, from 9 different provinces in China. The full test and dev sets will be used for evaluation. Magic data does not have a training/test division available in the set, so the following sub sets have been selected: SPK001 - SPK026 are the training set. SPK027 - SPK030 are used as the test set. This leads to about 233 minutes minutes of training data. No specific gender or location information is included with the data.

There are slight differences in how the transcripts for Aidatatang and Magic Data are provided. Aidatatang provides word-level transcriptions, but Magic Data provides character

level transcriptions. An excerpt from Aidatatang shows spaces between the words: "播放 有没有 人 曾告诉 你", but an excerpt from Magic Data shows a lack of spaces: "你竟你竟然能做出这种事儿". This matters because the pronunciation of a character can vary slightly based on the context. Having less context about the usage of the character makes it more difficult to determine the correct phoneme sequence that was used in the speech. This is an issue for both data sets; The Aidatatang phoneme lexicon provides several possible transcripts for most words, where it is unclear which transcript is the correct one. The Magic Data phoneme lexicon provides only character level transcripts, so some information related to the context or word is lost.

The data sets will have to be pre-processed for use; Specific files are required for Kaldi, and the feature vectors need to be extracted from the audio files. This pre-processing will also be performed with Kaldi. The processing of the data sets will be partially divided between other members of the research group.

2.3 Training and Optimization

After the data has been processed, the TDNN-OPGRU architecture will be implemented and trained. In order to provide forced alignments that are needed for training the network, a Gaussian Mixture Model - Hidden-Markov-Model (GMM-HMM) is trained on the data to provide these forced alignments. There are two types of GMM-HMM models directly available in Kaldi: A Tri3 model that has one round of Monophone training and 3 rounds of triphone training with delta+deltadeltas, LDA+MLLT and SAT training to generate forced alignments. A Tri5a model that has an extra round of delta+deltadeltas training on triphones and a second SAT training round with a larger number of leaves and gaussians. First will be determined which of these results in a lower PER. The model with a lower PER will be used in consecutive steps.

The TDNN-OPGRU network will then be implemented and trained based on the parameters used in [5],. Afterwards some of the parameters will be iterated, such as the layer size, layer count and training rate since these vary based on the size of the training data and complexity of the language [11]. The aim of this is to improve the PER. These iterations will be performed on the Aidatatang dev set to avoid over-fitting to the test sets. Because of time constraints no optimization of these parameters is performed on a dev set for Magic Data.

A simple overview of the base TDNN-OPGRU architecture can be found in Figure 1. This figure shows how the layers are interlaced with each other, any experiments that change the amount of layers will not change how the layers are interlaced.

2.4 Answering research questions

After the TDNN-OPGRU network is optimized on the Aidatatang dev set, the network will be trained separately on the Magic Data set, and on versions of both sets where the tone information has been removed from the transcripts. Then the trained networks will be used to decode the respective test sets. The resulting PER information from this will be used to determine what the PER is when recognizing phonemes in Mandarin speech as well as determining what the difference in PER is between spontaneous and prepared speech when decoding phonemes in Mandarin speech with a TDNN-OPGRU architecture.

More detailed information on the error rate of phonemes will also be gathered from the results of the decoding. Kaldi will store the substitutions, insertions, deletions and correct

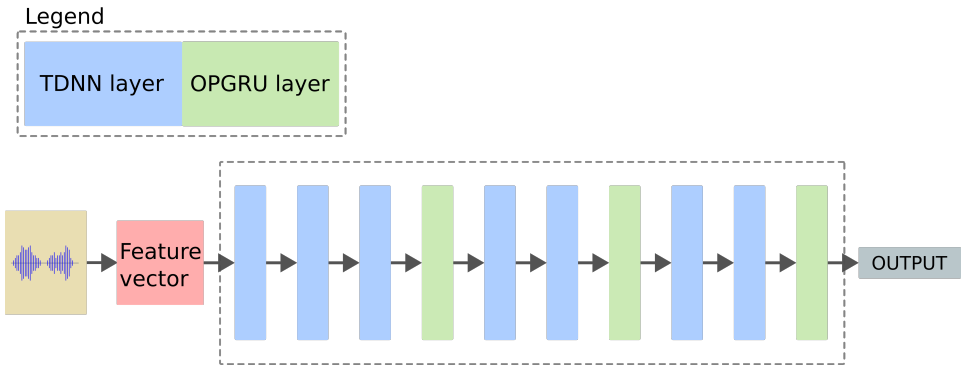

Figure 1: TDNN-OPGRU layout

results for each unique phoneme after comparing the result with the ground truth. With this information error-prone phonemes can then be identified.

2.5 Collaboration

There will be some collaboration between other members of the research group. This will be in regards to the pre-processing of data since multiple members have similar goals with different neural network/language combinations as mentioned in the introduction. For the evaluation of the TDNN-OPGRU networks, there will be dependence on the results of other members of the research group to be able to compare the results in a broader context.

3 Expanding Phoneme Recognition

As mentioned in the introduction, the goal of this research is to investigate the performance of the TDNN-OPGRU network when decoding phonemes in Mandarin prepared and spontaneous speech. Currently there does no exist research on using the TDNN-OPGRU network for phoneme recognition on Mandarin speech. This research can be used as point to continue expanding APR research for mandarin, and as a comparison for the evaluation of other acoustic models. The results of this paper will be compared with research on the TDNN-OPGRU architecture for phoneme recognition on other languages such as English [16] and Dutch [5]. And the performance of a similar neural network for Mandarin phoneme recognition, TDNN-BLSTM [17], that is being conducted in parallel with this research.

Next to making comparison with other languages and models, the goal is to perform comparisons within different permutations of the TDNN-OPGRU network, and gain insight into what phonemes the network struggles with identifying, as well as the differences in PER between prepared and spontaneous speech. Most interesting will be how the network handles the smaller differences between phonemes, as well as the differences in the network complexity and how the increase in phonemes influences this.

4 Experimental Setup

This section will discuss the setup used to train the TDNN-OPGRU. Steps such as the feature extraction, generation of the forced alignments and optimization steps will be discussed.

4.1 Feature extraction

When training an ASR or APR system, the input of that system is not the raw audio data. Rather, several features will be extracted from the audio, and this is what will be used as an input. Depending on the language and model, different features will give a different performance for the system.

The features that will be used are a combination of Mel-frequency cepstral coefficients (MFCCs) and pitch information. The pitch information is intended to aid the system specifically with the tone recognition. This results in an input vector with a size of 43.

4.2 Forced Alignments

The TDNN-OPGRU framework needs forced alignments to be trained on. This is not something that the data sets provide; They only provide character level transcripts for an sequence. In order to generate these, as mentioned before, a GMM-HMM will be used. Two different GMM-HMM setups will be tested, to determine which one is better suited for the data set. The two GMM-HMM setups are Tri3 and Tri5a. In both cases each consecutive model will be trained using alignments generated by the previous model.

For Tri3, first a monophone model is trained, and after that triphone models with delta + delta-deltas, LDA + MLLT and SAT are trained. Each of these models is trained with 2500 leaves and 15000 gaussians. For Ttri5a, two triphone delta + delta-deltas models are trained consecutively instead of one. A second, larger, SAT model is also trained at the end with 4000 leaves and 25000 Gaussians.

4.3 Optimization

During the optimization step of the network, some parameters will remain fixed with values based on [5]. The fixed parameters can be found in Table 1. The mini-batch size parameter is set to 64 as opposed to 128 used in [5] because of gpu memory constraints. Other variables such as the learning rate, layer size and layer count will be varied because their influence on the performance can vary with the size of the training data [11]. These parameters can be found in 2.

Parameter	Value
L2 Regularization	0.00005
Epochs	6
Mini-batch size	64
dropout schedule	0,0@0.20,0.3@0.50,0

Table 1: Fixed parameters

4.4 Decoding and scoring

For decoding and scoring some specific Kaldi algorithms will be used. For decoding the nnet3 decode.sh script will be used [12]. For scoring the result, Kaldi's score_wer_kaldi.sh script will be used [12].

Parameter	Minimum Value	Maximum Value
Initial Learning Rate	0.0005	0.05
Final Learning Rate	0.00005	0.005
Layer size	256	1024
Recurrent projection size	64	256
TDNN layer amount	4	7
OPGRU layer amount	2	3

Table 2: Variable parameters

5 Results

In this section the results of different inputs and training parameters will be presented. First the difference of the Tri3 and Tri5a GMM-HMM models will be shown. Next, the performance of the TDNN-OPGRU network with different parameters will be shown. Lastly, the differences between prepared and spontaneous speech, as well as the impact of tones will be shown based on selected parameters.

5.1 GMM-HMM

The PER that is achieved by the Tri3 and Tri5a GMM-HMM models on Aidatatang is shown in Table 3. This table shows that the Tri5a model does not offer an improved PER over Tri3, therefore Tri3 will be used to generate the forced alignments from now on.

Name	PER	
	test	dev
Tri3	43.88	43.58
Tri5a	44.54	44.19

Table 3: GMM-HMM results on Aidatatang

5.2 TDNN-OPGRU optimization

For determining what parameters are optimal several iterations of the TDNN-OPGRU architecture with varying parameters were performed on the Aidatatang dev set, the resulting PER of these iterations can be seen in Table 4.

Table 4 shows that the best results with tone were achieved with an Initial Learning Rate of 0.05 and Final Learning Rate of 0.005. The other parameters such as layer size and recurrent layers ended up being most optimal with values as specified in [5]. The need for a high learning rate, combined with a high layer count and size shows that the network was under trained on the smaller training data, but did require the bigger layer size and count to model the complexity of the language.

The difference in performance between the set including tone information and the set without tone information shows that the smaller difference in pronunciation between phonemes and the higher amount of phonemes requires a more complex network; The difference in PER between a layer size of 1024 and 256 is smaller for the tests with tone data removed. This also shows that since with tones, Mandarin has more phonemes than English, perhaps a layer size larger than 1024 might improve the PER of TDNN-OPGRU.

Initial learning rate	Final learning rate	Layer size	Rec layer size	TDNN layers	OPGRU layers	PER	
						tones	no tones
0.0005	0.00005	256	64	7	3	46.28	34.04
0.005	0.0005	256	64	7	3	41.55	31.61
0.05	0.005	256	64	7	3	42.63	30.89
0.0005	0.00005	1024	256	7	3	44.16	33.56
0.005	0.0005	1024	256	7	3	39.22	29.66
0.05	0.005	1024	256	7	3	**39.07**	**29.10**
0.005	0.0005	1024	256	4	2	42.15	31.51
0.005	0.0005	512	256	7	3	40.70	30.71

Table 4: TDNN-OPGRU PER on Aidatatang dev

5.3 PER in prepared and spontaneous speech

Based on the results in the previous section, the choice was made to go with the following settings: Initial Learning Rate = 0.05, Final learning rate = 0.005, Layer size = 1024, Recursive layer size = 256, 7 TDNN layers, 3 OPGRU layers. When evaluating the network on the test sets of Aidatatang and Magic Data

data set	PER
Aidatatang_200zh	39.99
Aidatatang_200zh no tones	29.34
Magic Data	30.76
Magic Data no tones	23.27

Table 5: Achieved PER on Aidatatang and Magic Data test sets

Table 5 shows that the TDNN-OPGRU network performs better on the Magic Data set than on the Aidatatang set. As a reminder, the Aidatatang_200zh data set is prepared speech, and the Magic Data set is spontaneous speech. This is unexpected because the general expectation is that prepared speech has a lower error rate than spontaneous speech [4].

There are several possible explanations for the difference in performance that are not caused by the language or network itself. Firstly, even though a similar amount of training material was used in both cases, a smaller amount of speakers was used for the Aidatatang set, with more audio material per speaker. Secondly, the Aidatatang data set was transcribed on a word level, where the Magic Data data set was transcribed on a character level. This possibly influences the correctness of the phoneme transcripts, as for Aidatatang a choice had to be made out of several options without being convinced of the correct one.

Lastly, the gender distribution of the training set for Aidatatang is 9 males and 1 female. Since the general set indicates a better balance of males and females, the higher PER could be caused by worse performance on female speakers. The gender distribution of the Magic Data training and test sets is unknown, so it is unclear how much influence this has.

Both data sets indicate that the participants were from different regions across Mainland China [15] [14], as such a diversity or lack thereof in this area does not appear to be the cause for this discrepancy.

Table 5 also shows that the removal of tone information before training has a positive impact on the PER for both prepared and spontaneous speech. Since the amount of possible

phonemes is smaller without tonal information there are less opportunities for confusion, and less complexity in the language. Lastly, the table shows that the difference in PER between data sets with and without tone information is similar between prepared and spontaneous speech. This might be an indication that the amount of errors made in decoding caused specifically by cases where only the tone is incorrect is similar between prepared and spontaneous speech.

Phoneme	AA	AE	AE5	AH	A0	AW5	AY5	EH	EH5
PER	1.0	1.0	0.86	0.99	1.0	1.0	0.95	9.83	0.97
Contribution to PER	0.006	0.004	0.19	0.016	0.001	0.003	0.054	0.021	0.023
Phoneme	IY	IY5	NG5	OW5	R5	UH	UH5	UW	UW5
PER	0.97	0.85	0.86	0.88	0.96	1.0	1.0	0.99	0.92
Contribution to PER	0.047	0.403	0.128	0.148	0.05	0.0	0.001	0.009	0.173

Table 6: PER of phonemes present in Aidatatang but missing in Magic Data

Table 6 shows the set of phonemes that are present in Aidatatang, but not in Magic Data. These phonemes almost exclusively exist of phonemes without tone that originate from English sections in the transcripts, as well as tones with the 5th neutral tone. The PER indicates that the network failed to recognize these phonemes well. The contribution to PER indicates that these phonemes did not occur a lot in the Aidatatang test set. In total they only account for 1.28% of the PER on the Aidatatang set, however the increased amount of phonemes still increases the complexity of the model.

5.4 Error-prone phonemes

For the evaluation of error-prone phonemes, tone information will be ignored. This means that substitutions where only the tone is decoded incorrectly will be considered as correct. Table 7 displays all phonemes which have an above average average contribution to the PER.

Looking at the results of Table 7, it shows that IY does not have an above average PER, but does still have the highest contribution to PER, this is caused by the high amount of occurrences of this phoneme, rather than the PER of the phoneme. The phonemes N, UW, NG, ER and AH can be considered error-prone phonemes for both prepared and spontaneous speech. AW and AE are error-prone only for prepared speech, D and AA are error-prone only for spontaneous speech. This shows that these are phonemes that are pronounced less clearly between the different types of speech. Several of the phonemes that are error-prone, such as N, ER, AA have a PER that is less than 2% higher than the average PER. This indicates that the main reason for the higher contribution to PER is the how often the phoneme occurs in the transcripts.

Another aspect that can be deduced from Table 7 is that the average PER of the sets where tone information was stripped before training is lower than the PER of the sets where the tone information was not stripped. This indicates that the introduction of tone information has a negative impact on ability of the TDNN-OPGRU architecture to correctly decode the base phonemes.

5.5 Tone

There are several ways in which the influence of tone on the performance of the TDNN-OPGRU can be examined. The direct approach is by looking at the error rate in tone

Aidatatang		
Phoneme	PER	cont
average	30.7	2.78
IY	22.1	9.5
N	31.2	8.6
UW	37.6	6.9
NG	41.0	6.0
AE	32.5	5.8
ER	31.7	4.4
AH	40.0	3.5
AA	30.6	3.0
AW	34.8	2.9

Aidatatang no tones		
Phoneme	PER	cont
average	28.5	2.78
N	29.0	9.1
IY	16.2	7.8
AE	33.4	7.3
UW	34.3	7.3
NG	36.7	6.1
ER	30.2	4.6
AH	40.7	4.1
AA	27.7	3.1
D	22.8	2.9
AW	31.4	2.9
Y	23.9	2.8

Magic Data		
Phoneme	PER	cont
average	23.4	2.78
IY	18.5	10.2
N	25.2	8.2
UW	30.5	7.9
ER	24.4	5.0
AA	25.2	4.9
D	24.5	4.5
AE	20.0	4.3
NG	24.3	4.1
AH	31.9	3.4
JH	33.1	2.9

Magic Data no tones		
Phoneme	PER	cont
average	22.8	2.78
N	25.8	9.2
IY	14.8	8.9
UW	27.4	7.7
ER	27.2	6.1
AA	24.0	5.1
D	24.8	5.0
AE	19.6	4.8
NG	22.4	4.1
AH	33.6	4.0
R	44.7	3.2
Y	22.3	3.1
JH	32.0	3.0
AO	25.1	3.0

Table 7: The phonemes with an above average contribution to PER, sorted by contribution to PER, the cont column represents contribution to PER.

recognition, this can be seen in Table 8. The influence of tone on the performance of the TDNN-OPGRU network between prepared and spontaneous speech can be analysed by looking at what percentage of the substitution errors is caused by tone-only errors. These are errors in decoding where the base phoneme is correctly recognized, but the tone is incorrectly recognized. This contribution to substitution is shown in Table 9.

data set	Tone error rate
Aidatatang	27.48
Magic Data	20.42

Table 8: Tone Error rate per data set.

data set	substitution %
Aidatatang	25.94
Magic Data	25.99

Table 9: Percentage of substitutions caused by tone errors.

Table 8 shows that a significantly higher tone error rate Table 9 demonstrates that even though a lower PER is achieved on Magic Data, a similar percentage of the substitutions made are caused by tone-only errors. Indicating that the TDNN-OPGRU network does not have more issues with identifying tone between prepared and spontaneous speech.

Tone	1	2	3	4	5
1	77.4	5.5	4.5	11.6	1.0
2	11.2	68.5	10.6	8.1	1.6
3	8.3	6.8	73.5	9.8	1.5
4	12.4	4.9	6.7	74.2	1.8
5	10.0	8.5	9.6	13.5	58.4

Table 10: Aidatatang tone confusion matrix

Tone	1	2	3	4	5
1	81.5	4.9	3.7	8.7	1.2
2	9.0	75.7	6.9	7.2	1.2
3	6.3	6.6	79.2	7.2	0.8
4	9.2	5.6	3.8	80.6	0.8
5	4.7	5.9	3.4	8.0	77.9

Table 11: Magic Data tone confusion matrix

Tone	Tone percentage	
	Aidatatang	Magic Data
1	23.2	27.2
2	17.9	19.0
3	22.6	15.6
4	28.4	33.5
5	7.8	4.8

Table 12: Distribution of tones in the test sets

Tables 10 and 11 are confusion matrices. Each row shows how often a tone, together with the distribution of tones in Table 12. Show some similarities and some differences between the different data sets regarding tone recognition.

Tables 10 and 11 show that in both Aidatatang and Magic Data sets, tone 1 and tone 4 are mistakenly recognized the most often. Tone 5 is mistakenly recognized very little. In both the Aidatatang and Magic Data test set the occurrence of tone 1 and 4 is the highest, this might also be present in the training data, causing the network to be biased towards tone 1 and 4. However, Tone 5 is poorly recognized in the Aidatatang test set, with a correct recognition of only 58.5%, compared to a recognition of 80.1% in Magic Data. This can be partially explained by the phonemes present in Table 6, since the total amount of phonemes with tone 5 in the transcript is small, and there are several phonemes with tone 5 present in that table,

6 Responsible Research

This section will discuss the ethical aspects of the research and the repoducibility of the methods used. It will go into detail on the general ethical implications of phoneme recognitions, as well as mention some aspects that could be investigated further. After this it will discuss how it has been ensured that the results obtained in this research are reproducible.

6.1 Ethical implications

As mentioned in the introduction of this paper, APR can benefit people who are learning a new language or benefit people with a speech impediment [3]. In this regard APR research has a societal benefits that it can increase equality by helping people communicate better, both in helping people to learn languages, and increasing the performance of ASR for people who have trouble speaking.

This specific research does not focus on the bias of APR or ASR systems between people of different genders, ethnicity or geographical locations. This is something that is worth researching, and is covered for example in [18]. This bias is actually a potential issue in this research, because there is an imbalance in the gender distribution of one of the training sets. However, despite this the research can still provide insight into the Performance of TDNN-OPGRU as an APR system for Mandarin.

6.2 Reproducibility and integrity

There are several ways in which this paper attempts to allow reproducibility of the results. The main way this is achieved is by including as much information as possible when it comes

to the parameters that were used to train the network. Another way in which this was done was by disclosing choices that were made, such as the choice to select the first occurrence of a character or word in the lexicon. This situation is not ideal, but the choice that is made was described and can thus be followed by future research, or changed if deemed necessary or possible.

Some issues, such as the imbalance in gender distribution in the Aidatatang training set are not ideal. However, while they influence the result of the research, they are disclosed and discussed. This allows readers of this paper to draw their own conclusions and perform future research. The code used for preparing the data and running the networks will be available on GitHub [19] as well. This will ensure that other people are able to reproduce the results and further investigate these issues.

This research aimed to take the parameters selected in [5]. But when implementing and training the network it appeared that some parameters, such as the learning rate, were missing from the information. It is important to disclose these possible deviations from that research since some the results will be compared.

7 Discussion

In this section the results be reflected on and placed in a broader context by comparing them to other research on PER. The results will also be compared with other papers focusing on tone recognition.

7.1 Phoneme Error Rate

The difference in PER between prepared and spontaneous speech is unexpected as mentioned in the results section. A better result is achieved on spontaneous speech than on prepared speech. This counters the results of [5], where prepared speech is easier to decode than spontaneous speech, especially by a large margin as shown in Table 5. However, results obtained by [17] which used the same setup but implements the TDNN-BLSTM architecture, obtain similar results. The performance on prepared speech is lower than on spontaneous speech.

As mentioned in the results section, there are several reasons that can cause the higher PER on prepared speech that are not due to the Mandarin language or the TDNN-OPGRU architecture, such as the amount of speakers, gender distribution an differences in transcription. The parameters that were used for the comparison between prepared and spontaneous speech were determined on the prepared speech dev set. This gives a possible bias in the performance of the architecture towards prepared speech, since different parameters might have resulted in a higher performance on the spontaneous speech set, but this was not evaluated. Despite this, the achieved PER of the TDNN-OPGRU architecture is still higher on spontaneous speech than on prepared speech.

The PER obtained by TDNN-BLSTM in [17] is higher than that of TDNN-OPGRU for both prepared and spontaneous speech. TDNN-BLSTM obtained a PER of 45.31 and 35.38 on prepared and spontaneous speech respectively, compared to 39.99 and 30.76 for TDNN-OPGRU. Together with the results obtained in this research this indicates that TDNN-OPGRU is better suited for phoneme recognition in Mandarin than TDNN-BLSTM.

Based on the optimization results a higher layer size is beneficial for an improved PER with the TDNN-OPGRU network. However, no layer size above 1024 cells was evaluated. It is possible that a large layer size will yield an even lower PER.

Two researches that evaluated TDNN-BLSTM [20] and TDNN-OPGRU [16] on English prepared and spontaneous speech had a similar result. There TDNN-OPGRU obtains a lower PER than TDNN-BLSTM on both prepared and spontaneous speech as well. This indicates that TDNN-OPGRU might be a better performing network in general, despite outperforming the TDNN-BLSTM network on dutch prepared speech [5].

7.2 Error-prone phonemes

In the results section the phonemes N, UW, NG, ER and AH have been marked as error-prone phonemes for both prepared and spontaneous speech, it is important to repeat that these are not complete phonemes in Mandarin, as the tonal aspect has been stripped. These error-prone phonemes are quite different from the error-prone phonemes occurring in dutch according to [5]. IH, AA, K, S, and AX are error-prone in Dutch, which indicates that there is no direct overlap in error-prone phonemes between these languages, AX does not exist in the Mandarin phonemes at all. Some similarities occur with the phoneme AA, it is error-prone for spontaneous speech in Mandarin, but not for prepared speech. These differences in error-prone phonemes demonstrate how different the Mandarin and Dutch language are.

7.3 Tone

The results demonstrate that an equal amount of the PER is caused by tone-only errors in both prepared and spontaneous speech. Despite this, tone errors do make up 29 % of the substitution errors made on both prepared and spontaneous speech. TDNN-OPGRU achieved a tone-error of 27.48% on prepared speech and 20.42% on spontaneous speech. These results are better than the achieved PER on base phonemes without tone information. However there were 36 phonemes present and only 5 tones, indicating that the tone recognition is harder in comparison to the base phoneme recognition for the TDNN-OPGRU network.

This achieved tone error is also worse than the results obtained in [7], where the best tone error achieved is 19.82% by a DMN network. They are also worse than the results obtained in [8], which indicates a tone error of 9.7% on read news speech. This shows that the TDNN-OPGRU network in its current implementation is still behind in regards to tone recognition by alternative methods.

8 Conclusions and Future Work

The goal of this research was to investigate the performance of the TDNN-OPGRU network when decoding phonemes in Mandarin prepared and spontaneous speech. Several research questions were formed to aid this goal, these questions will now be answered. Recommendations for future research will be made

What is the PER when recognizing phonemes in Mandarin speech with a TDNN-OPGRU architecture? The best achieved PER is 39.99% on prepared speech and 30.76% on spontaneous speech. This is achieved with an intial and final learning rate of 0.05 and 0.005 respectively. The other parameters for this are identical to those defined in [5], except for the mini-batch size which is set to 64 because of computational limitations. A higher PER might be achievable with bigger data sets and computational power. This is something that should be examined in future work.

What is the difference in PER between prepared and spontaneous speech when recognizing phonemes in Mandarin speech with a TDNN-OPGRU architecture? The difference

in PER is 7.23% in the favor of spontaneous speech. This is different from existing research [5], the general expectation is that spontaneous speech is harder to decode [4]. There are several aspects of the research that could have contributed to this result. A smaller amount of speakers in the prepared speech training data, an in balance in the gender distribution of the prepared data training set for prepared speech and a difference in the transcripts of the two data sets. The fact that TDNN-OPGRU performs better on spontaneous speech is unexpected, and should be investigated further.

What phonemes have an above average PER when decoding phonemes in Mandarin speech with a TDNN-OPGRU architecture? The phonemes N, UW, NG, ER and AH are error-prone when decoding both prepared and spontaneous speech with TDNN-OPGRU. The phonemes AW and AE are error-prone only when decoding prepared speech, and the phonemes D and AA are error-prone when decoding spontaneous speech with a TDNN-OPGRU architecture. This difference in error-prone phonemes highlights areas of the language where the TDNN-OPGRU architecture performs worse between the speaking styles. Further research investigating the cause of this difference in error-prone phonemes is recommended.

What influence does the presence of tones have on the PER when decoding phonemes in Mandarin speech with a TDNN-OPGRU architecture? A similarly sized TDNN-OPGRU network will achieve a worse PER when tone information in present. It achieves a PER of 39.99% versus 29.34% on prepared speech, and a PER of 30.76% vs 23.27% on spontaneous speech. This shows that the added complexity and increased amount of phonemes caused by tonal information makes it more difficult for the TDNN-OPGRU architecture to correctly recognize phonemes. However, in both prepared and spontaneous speech, mistakes where only the tone is recognized incorrectly contribute to a similar percentage of the substitution errors. Indicating that the difference in the PER that is obtained by TDNN-OPGRU is not caused by errors where only the tone is incorrect. It is also shown that the PER on base phonemes is worse when tone information is included during training. This indicates that the inclusion of tone information negatively affects the ability of the TDNN-OPGRU architecture to recognize the base phonemes.

In this research the main focus has been on a separate evaluation of tone recognition and base phoneme recognition. However as is stated in [8] the tone and base phoneme are context dependent and the decision for tone can affect the decision for the base phoneme during decoding. Due to this more insights and improvements can be gained in future research by evaluating the tone and phoneme results of TDNN-OPGRU as a whole.

References

[1] Britannica, T. Editors of Encyclopaedia, *Phoneme. encyclopedia britannica*, https://www.britannica.com/topic/phoneme, Accessed:2021-06-04, 2009.

[2] S. J. Witt S. M. & Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30(2-3), pp. 95–108, 2000.

[3] B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild," in *Proc. Interspeech 2020*, 2020, pp. 4826–4830. DOI: 10.21437/Interspeech.2020-1598. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1598.

[4] R. Dufour, "From prepared speech to spontaneous speech recognition system: A comparative study applied to french language," in *Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, ser. CSTST '08, Cergy-Pontoise, France: Association for Computing Machinery, 2008, pp. 595–599, ISBN: 9781605580463. DOI: 10.1145/1456223.1456345. [Online]. Available: https://doi.org/10.1145/1456223.1456345.

[5] R. Levenbach, "Phon times: Improving dutch phoneme recognition," M.S. thesis, Delft University of Technology, Delft, 2021.

[6] Mustgo, *Mandarin (Chinese)*, https://www.mustgo.com/worldlanguages/mandarin/, Accessed: 2021-04-21.

[7] Z. Chen Mingming. Yang and W.-J. Liu, "Deep neural networks for mandarin tone recognition," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1154–1158, 2014.

[8] X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for mandarin broadcast news speech recognition," *INTERSPEECH-2006*, paper 1752–Tue3A2O.4. 2006.

[9] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with lstms," in *Proc. Interspeech 2017*, 2017, pp. 1586–1590. DOI: 10.21437/Interspeech.2017-129. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-129.

[10] W. Zaremba, I. Sutskever, and O. Vinyals, *Recurrent neural network regularization*, 2014. [Online]. Available: https://arxiv.org/abs/1409.2329.

[11] T. van Niedek T. amd Heskes and D. van Leeuwen, "Phonetic classification in tensorflow," 2016.

[12] *Kaldi ASR*, https://kaldi-asr.org/, Accessed:2021-04-21.

[13] OpenSLR, *Aidatatang-200zh*, http://www.openslr.org/62/, Accessed:2021-04-20.

[14] L. D. Consortium, *Magic data chinese mandarin conversational speech*, https://catalog.ldc.upenn.edu/LDC2019S23, Accessed:2021-04-20, 2019.

[15] *HPC cluster | Intelligent Systems Department*, http://insy.ewi.tudelft.nl/content/hpc-cluster, Accessed:2021-05-20.

[16] G. Genkov, "Training and testing the tdnn-opgru acoustic model on english read and spontaneous speech," B.S. Thesis, Delft University of Technology, Delft, 2021.

[17] M. Chiroşca, "Evaluation of phoneme recognition through tdnn-blsm in mandarin speech," B.S. Thesis, Delft University of Technology, Delft, 2021.

[18] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, *Quantifying bias in automatic speech recognition*, 2021. arXiv: 2103.15122 [eess.AS].

[19] J. van der Tang, *Tdnn-opgru-mandarin git repository*, https://github.com/jordyjordy/TDNN-OPGRU-Mandarin.

[20] I. Klom, "Assessing the performance of the tdnn-blstm architecture for phoneme recognition of english speech," B.S. Thesis, Delft University of Technology, Delft, 2021.