Evaluating the performance of TDNN-BLSTM on Mandarin read and spontaneous speech

Mihail Chiroşca[1] , **Dr. Siyuan Feng**[1] , **Dr. Odette Scharenborg**[1]
[1]TU Delft

Abstract

A limitation of current ASR systems is the so-called out-of-vocabulary words. The solution to overcome this limitation is to use APR systems. Previous research on Dutch APR systems identified Time Delayed Bidirectional Long-Short Term Memory Neural Network (TDNN-BLSTM) as one of best performing state-of-the-art NN architecture for PR. The goal of this research is to evaluate the performance of the TDNN-BLSTM architecture for phoneme recognition on Mandarin read and spontaneous speech, analyze the differences in performance for the two speech styles as well as compare the results with previous research on Dutch PR.

To achieve this goal 4 different NN models of the TDNN-BLSTM architecture were built and trained on Mandarin read and spontaneous speech. The test results of the NN models were used to calculate the phoneme error rate (PER), decomposed PER, and the contribution of individual phonemes to the overall PER. Based on these findings, conclusions are formulated regarding the impact of different languages, speech styles, and the architectural changes on the performance of the TDNN-BLSTM architecture.

1 Introduction

Automatic speech recognition (ASR) is a perfect example of a system that benefited from the popularity of machine learning and was able to make the step from theory to practical application. Indeed, due to the relatively recent developments of deep learning, ASR performance has improved a lot, to the point that it is used in many different applications including smartphones and smart speakers.

Although current ASR systems work fairly well, they have their limitations. One of these limitations is that an ASR system can only recognize those words that are in its lexicon. If a word is not in its lexicon, it cannot be recognized. To be able to deal with such out-of-vocabulary words, automatic phoneme recognizers (APR) are often employed. Such an APR transcribes the speech signal into a sequence of phonemes, i.e., sounds, instead of words.

In speech recognition phonemes are defined as the smallest unit that will distinguish between words [9]. Thus phonemes are the building block of speech. Therefore phoneme detection is a key step in speech recognition [8]. Although there is some research on APR systems [8], it is typically done with only a small number of well-researched data sets. However a recent study [5], introduced two new APRs systems that were able to outperform the until then best Dutch APR by a wide margin [3]. Based on its findings, the research concluded that the performance of different speech styles was dependent on the architecture of the neural network used in the design of the APR system.

This research builds upon the work done by [5] and aims to assess the performance of TDNN-BLSTM on Mandarin read and spontaneous speech. To be noted that Mandarin is a tonal language, and as such it will give insight on how the performance of the NN changes across different types of languages. Thus this research aims to answer/investigate the following questions:

- What is the performance of the TDNN-BLSTM on Mandarin spontaneous speech?
- What is the performance of the TDNN-BLSTM on Mandarin read speech?
- How does the performance of the TDNN-BLSTM compare between Mandarin read and spontaneous speech?
- How do the results of the TDNN-BLSTM on Mandarin speech compare to previous research conducted on Dutch speech?

This report is structured as follows. First, in Section 2 an overview of the used corpora and tools is given as well as background information about TDNN-BLSTM. Besides that, section 2 also gives an overview on how this research will evaluate its results. Section 3, discusses the initially proposed experimental setup of this research and its variations as well as presents the purpose and the expectations from the corresponding experiments. The issues encountered during the data preparation step and their possible impact on the corresponding results are also introduced in section 3. Section 4 presents the results of the corresponding experiments described in section 3 and their findings. Section 5 is a reflection on responsible research and talks about the reproducibility of the study Thereafter, in Section 6, a summary with the an-

swers to the proposed research questions is presented alongside future research ideas.

2 Methodology

As stated in Section 1, this research is meant to evaluate the performance of TDNN-BLSTM on Mandarin speech. This section explains what TDNN-BLSTM is, its characteristics, as well as the evaluation metrics and procedure. A brief description of the corpora used for NN training concludes this section.

2.1 TDNN-BLSTM and what it is ?

The problem of basic/vanilla Recurrent Neural Networks

- Basic RNNs are unidirectional, which means that a time step *t* is only predicted using the information of the parsed sequence from time steps *1* to *t-1* [5]. However, speech frames are also characterized by their future context.
- The phenomenon of gradient vanishing [4], prevents NN from learning long-term dependencies.
- The important information of a time step *t* lies in a relatively narrow context of that time step [6].

TDNN-BLSTM stands for Time-Delayed Bidirectional Long-Short Term Memory Neural Network, it is a NN that combines in itself different types of architecture. Thus the obtained hybrid-NN can cope with every single issue described above.

First, the LSTM architecture, by introducing cells and gates, can "learn" when and how to update the "internal memory" of a neuron, which solves the vanishing gradient problem. Thus the NN gets the property to differentiate the data patterns/sequences that are important and must be kept (gets a long- and short- memory). In the context of PR, LSTM helps in differentiating between phonemes that share similarities in their pronunciations. With a correct setting, LSTM can enable accents predictions/differentiation thus improving the precision of the System.

Besides getting an "internal memory", that remembers patterns across sequences of data, in speech recognition, the current sounding of a phoneme is directly dependent on not only its previous but also the upcoming phonemes. Unfortunately, as already mentioned simple RNNs are not able to "foresee" the future, but bidirectional RNN can. A situation where a bidirectional NN is useful is trying to predict the next phoneme in the following sentence: "/h/ /e/. . . " Predicting the next phoneme in this sentence is a guess without any other context (e.g. /m/, /l/, /v/, etc.). However, when the part after this time step is predicted as: ". . . /ə/ /ʊ/." predicting the phoneme becomes easier (i.e./l/) [5]. The way BRNN works is by having 2 layers of RNN, one of them being a reverse copy of the other. We then combine the obtained past and future states and can compute/predict the phoneme.

Even though by using LSTM we can learn and memorize the common patterns that add up to a phoneme representation we often do not require the full sequence for this. In phoneme recognition, the distinctive characteristics of a phoneme are often "hidden" in very small chunks (which often overlap) when compared to the actual sequence. Time-Delayed NN are meant to introduce a temporal context when making predictions [10] by also analyzing the *t - r* to *t* in contrast to default RNNs. The strength of TDNNs for PR comes from the characteristic that important information of a time step t lies in a relatively narrow context of that time step [6].

2.2 Evaluation

The performance of the TDNN-BLSTM trained acoustic model is evaluated as follows:

- The Phoneme error rate (PER) of the NN is calculated to find the overall performance of the attempt/experiment for each
- A decomposed PER of a model is to be calculated, thus allowing to identify the strengths/weaknesses of the different models/settings.
- The contribution to total PER is to be calculated.

The Phoneme Error Rate (PER) is based on the "Levenshtein Distance" and is the main metric used to evaluate the performance of attempts. The "Levenshtein Distance" aims to find the minimum number of single-character edits between 2 strings [11]. Similarly, PER calculates the difference in single phonemes edits between, in this case, the ground truth and the predicted phoneme transcripts. The PER metric is calculated by considering 3 types of edits that are needed such that the ground truth and predicted phoneme sequences match, namely:

- deletions - phonemes that are not present in the predicted transcript but are part of the ground truth transcription.
- substitutions - phonemes that must be changed for other phonemes so that they match the ground truth transcription.
- insertions - phonemes that are not present in the ground truth transcript but are present in the predicted sequence.

The formula for calculating PER is as follows:

$$PER = \frac{S_{all} + I_{all} + D_{all}}{N}$$

The above formula represents the sum of all substitutions(S), insertions(I), deletion(D) for all the phonemes divided by the total number of phonemes that occur in the ground truth(N).

Besides that, the decomposed PER represents the individual contributions of all the phoneme substitutions(S), insertions(I) and deletions(D) to the overall PER, thus providing insights into the strengths and weaknesses of the NN. The three formulas for calculating the decomposed PER are presented below:

$$\%Substitutions = \frac{S_{all}}{N}$$

$$\%Insertions = \frac{I_{all}}{N}$$

$$\%Deletions = \frac{D_{all}}{N}$$

The formula for calculating individual contribution to total PER is as follows:

$$ContributionToPER_{phoneme_x} = \frac{S_x + I_x + D_x}{S_{all} + I_{all} + D_{all}}$$

By calculating the contribution to PER, it is possible to identify what percentage of the total PER belongs to certain phoneme and as such its impact on the overall performance.

2.3 Corpora and Tool Set

The corpora used in this research are the Aidatatang_200zh corpus [1] and the Magicdata corpus.

The Aidatatang_200zh corpus is a free Chinese Mandarin read speech corpus containing 200 hours of acoustic data, with 600 speakers from different accent areas in China. The stated accuracy of the corpus transcripts is larger than 98% [1]. This study uses a pre-selected subset of the respective corpus with the training set being composed of 10 speakers with a total of *3.5h* of recorded Mandarin read speech.

The Magicdata corpus is a Chinese Mandarin conversational corpus, which however is not freely available as the previously specified one. This study makes use of a pre-selected subset of the respective corpus with the training set being composed of 26 speakers with a total of *4h* of recorded Mandarin spontaneous speech.

It is worth mentioning that the two corpora that are used, each represent a different speech type, namely the spontaneous and prepared speech. The main difference between these two is, spontaneous speech is a more variable and less well-articulated speech. Moreover, previous research [5], concludes that the performance of the NN is dependent on the speech types it was trained on.

Besides that, it is worth mentioning that the respective training- and test- sets do not contain any overlapping speakers and/or utterances.

This research also makes use of Kaldi [2], a state-of-the-art toolkit written in Shell and C++ used to extract and create feature vectors, build the language model as well as build and train the acoustic model, and perform decoding.

3 Experimental work

This section introduces the experimental work of this research such as:

- the lexicon of the corpora
- the feature vectors used in this research.
- the parameters used for TDNN-BLSTM training and their variations
- the purpose and/or expectations of the conducted experiments
- the issues and the restrictions of this research setup

3.1 Corpora Lexicon

Table 1 gives an overview of the individual phonemes used/present in the selected corpora. Additional entries are:

- *[SPN]* - usually used to describe noise in audio segments and not only (see 3.4)
- *[SIL]* - used to represent silence is audio speech segments
- *[LAU]* - used to represent laughter.

Besides that, due to Mandarin being a tonal language, Table 1 includes phonemes with digit suffixes such as [AA1], [AA2], [AA3], etc. The purpose of including digit suffixed phonemes is for the NN to be able to represent and capture tone information at the phoneme level rather than during the feature vector creation step. The consequences and impact of using phonemes that capture tone variation is described in section 3.3

Z	SIL	AA1	AA2	AA3	AA4	AA5	AE1
AE2	AE3	AE5	AH1	AH2	AH3	AH4	AH5
AO1	AO2	AO3	AO4	AW1	AW2	AW3	AW4
AW5	AY1	AY2	AY3	AY4	AY5	CH	D
EH1	EH2	EH3	EH4	EH5	ER1	ER2	ER3
ER5	EY1	EY2	EY3	EY4	EY5	F	G
HH	IY1	IY3	IY4	IY5	J	JH	K
L	LAU	M	N	N2	N3	N4	N5
NG1	NG2	NG3	NG4	NG5	OW1	OW3	OW4
OW5	P	Q	R	R2	R3	R4	R5
SH	SPN	T	UH1	UH2	UH3	UH4	UH5
UW1	UW2	UW4	UW5	W	X	Y	

Table 1: Corpora Lexicon

3.2 Experimental Settings

This subsection presents the specific settings used for and during the conducted experiments.

This research uses existing Kaldi scripts to extract two different types of features: I-Vectors and Mel-frequency cepstrum features [5]. The MFCC and I-Vector features are then concatenated forming a single feature vector that is used as input for the NN.

Table 2 presents the default parameters [5] as well as their variations that were used for training TDNN-BLSTM on both prepared and spontaneous speech whilst Figure 1 depicts the architecture used for each of the conducted experiments.

Parameter\#Experiment	1	2	3	4
BLSTM layers	3	3	4	4
TDNN layers	7	7	9	9
cells per BLSTM layer	1024	256	256	256
mini-batch	128	128	64	64
initial learning rate	0.001	0.001	0.001	0.01
final learning rate	0.0001	0.0001	0.0001	0.001
epochs	6	8	8	8
L2-regularisation	0.00005			
dropout schedule	0,0@20,0.3@0.50,0			

Table 2: Experimental settings for training TDNN-BLSTM on Aidatatang_200zh and MagicData subset

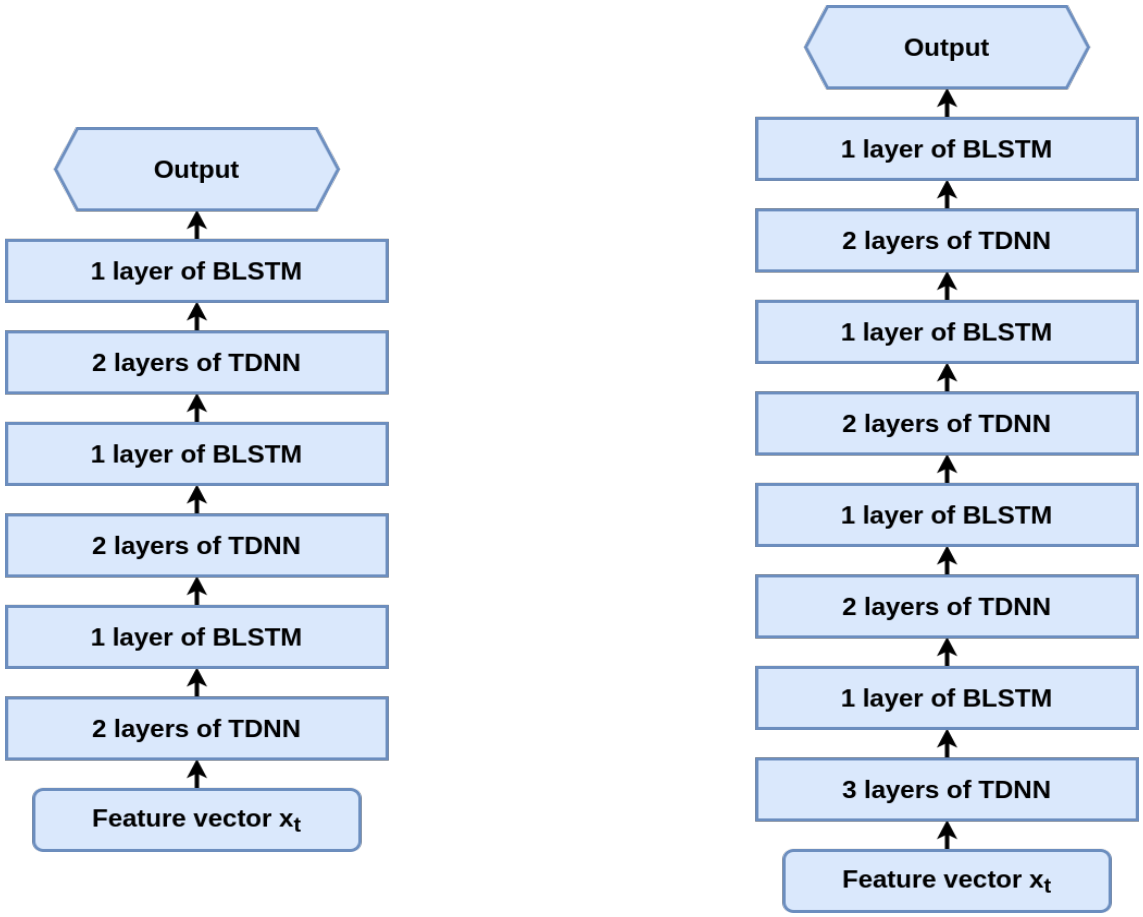

Figure 1: Architecture of the TDNN-BLSTM for experiment 1 & 2 - left, and experiment 3 & 4 - right.

3.3 Experiments: Purpose and Expectations

This subsection explains the purpose of the conducted experiments as well as presents some the expected results.

The first experiment uses the default settings as proposed by [5]. The purpose is to abstract the NN so that the results are only language-dependent, thus allowing for comparisons across different types of languages, in this case a comparison between Dutch and Mandarin. Moreover, analyzing the overall and decomposed PER will give insights into the possible differences and/or origin of the errors for the respective languages.

By considering the reduced size of the training set (see section 2.3), experiments 2, 3 & 4 will aim to test whether it is possible to improve the results obtained in experiment 1 by tweaking different parameters such as the *epoch* parameter which determines the period of evolution/the number of generations of a NN model. Moreover, parameters such as the number of cells per layer, the batch size, and the learning rate are also changed in an attempt to account for the small training sets.

Besides only tweaking the NN parameters, experiments 3 & 4 do introduce additional TDNN and BLSTM layers to the NN model with regards to the default proposed settings from experiment 1. The idea is to identify how additional layers influence the overall and decomposed PER of the system. The correlation between the additional layers and the distribution of phoneme contribution to PER among different speech styles is also worth investigating.

Lastly, as depicted in [5] and is explained in section 2.3, read speech being overall a more consistent and well-articulated speech, results in better scores in the experiments, thus such a pattern is also expected to be observed here.

3.4 Restrictions

One of the starting points of any of the above-mentioned experiments/attempts was the Data preparation step. The selected corpora represents the content of the recordings as sentences/sequences of words, thus they are to be transcribed into phoneme sequences.

The first issue is the incompleteness of the transcripts. The lexicon attached to the Magicdata corpus is incomplete (less than 2%), thus parts of the transcripts that were not possible to be transcribed were marked as spoken noise.

The second restriction is due to Mandarin being a tonal language, which means that the tone of the syllable affects the meaning of the sequence. The provided lexicon for both Magicdata and aidatatang_200zh corpora does, occasionally provide multiple transcriptions for the tone-dependent sylla-

ble that alter the meaning of the words. However, the ability to choose the correct tone-dependent transcription in an automated way as well as being able to prove its veracity, especially for a non-Mandarin speaker, falls out of the scope of this project.

The third restriction is a result of the available data set/corpora size. The article [5] that this research builds upon, used for its training, a subset of 140 hours from the Corpus Gesproken Nederlands(CGN). However, this research uses subsets of the provided corpora, Aidatatang_200zh (read speech) and Magicdata (conversational speech), which are considerably smaller (see section 2.3). Thus even if the preselected subsets would maintain the distribution of the speakers similar to the full sets, due to their reduced size, less accurate models are to be expected.

4 Results

This section introduces the results that were obtained after conducting the experiments specified in section 3.2. First, the overall PER of all the experiments is given (see Section 4.1). Then the errors of the architecture are further analyzed by identifying their origins (see Section 4.2). Lastly in section 4.3 analyzes the distribution of the contributions to the overall PER.

4.1 Overall Phoneme error rates

Tables 3 and 4 present the PER results for all of the conducted experiments according to section 3.2.

Experiment	1	2	3	4
PER	48.89%	50.68%	51.21%	45.31%
subst.	34.16%	35.35%	36.31%	32.99%
delete	11.50%	12.38%	11.89%	9.22%
insert	3.21%	2.93%	3.0%	3.09%

Table 3: Read speech results (Aidatatang_200zh corpus).

Experiment	1	2	3	4
PER	37.88%	39.86%	43.09%	35.38%
subst.	25.25%	26.92%	25.10%	24.09%
delete	9.86%	10.31%	15.44%	8.67%
insert	2.82%	2.95%	2.54%	2.61%

Table 4: Spontaneous speech results (Magicdata corpus).

As can be seen from Table 3 and 4 the settings of experiment 3 achieved the worst overall results for both spontaneous and read speech. Experiment 2 changes its results very little on the read speech in contrast with the spontaneous speech result where it improved by 3%. Experiment 1 with the default parameters showed an improvement of roughly 2% independently of the speech type, whilst experiment 4, succeeded to achieve the lowest overall PER, which means the best-observed performance. When compared to the results of [5], the obtained PER for TDNN-BLSTM are considerably higher (the performance is worse) independently of the experiment. However, the decrease in the performance of NN is to be expected mainly due to the reduced size of the training corpus, as described in section 3.4.

4.2 Decomposed PER

Besides that, Table 3 & 4 also include the decomposed PER, or the individual percentage contribution of substitutions, deletions, and insertions to the overall PER. Thus by inspecting the obtained results, it can be noticed that most of the errors originate from substitutions, followed by deletions, and lastly insertions. Moreover, if we compare the decomposed PER across the two speech types the number of insertions error hardly ever changes settling at around 3%. Similarly, the number of deletions slightly fluctuates with an average of 11% across the 4 experiments for both spontaneous and clear speech. However, when it comes to the number of substitutions between conversational and prepared speech, we get to observe the biggest differences so far, with an average of 25.4% and 34.7% respectively across all 4 experiments. Similarly, it can be seen that TDNN-BLSTM performed better on spontaneous speech rather than on clear speech, despite the expectations formulated in section 3.3. Indeed, this also does not match with results from previous research on Dutch PR. A reason for this unexpected and contradictory result may be the difference in the quality of the recordings between the read and spontaneous speech corpora.

As mentioned above most of the errors originate from substitutions independently of the speech type, which is not the case for the Dutch-based TDNN-BLSTM model from [5]. The reason for this is, as previously explained, the fact that Mandarin is a tonal language. The latter statement is also supported by Figure 2 that pictures the list of substitutions needed so that the predicted sequences match the ground truth. It can be noted that most of the substitution errors are between phonemes with the same basis but different digit suffixes.

4.3 Contribution to PER

The evaluation of the performance of TDNN-BLSTM is further analyzed by inspecting the contribution to PER between different experiments Figure 3 and 4, as well as between different speech styles Figure 5. It is worth mentioning that for the sake of simplicity the phoneme with the same basis was replaced by the respective basis. For example phonemes such as [AA1] and [AA2] were converted to [AA].

By inspecting Figure 3 it can be concluded that the distribution of the contributions to PER is not directly dependent on the parameter and architectural changes of NN across the conducted experiments for read speech.

In Figure 4 the only perceivable changes to the distribution of contributions to PER can be seen in experiment 3, where it succeeds to get a slightly lower contribution to PER for 4 out of 7 top error-prone, namely [R], [Y], [W], [G] which at the same time (excluding [R]) do not have any tone variations of themselves present in Table 1. The previous statement may try to infer that some models of NN with additional layers and/or reduced cell size may favor non-tonal phonemes, but not enough evidence as well as the poor overall PER of experiment 3 support this hypothesis. Thus, similarly to the observations on read speech, from Figure 4, we can infer that

substitution	IY1	IY4	100
substitution	IY4	IY1	99
substitution	IY2	IY4	74
substitution	IY3	IY1	50
substitution	N4	N1	46
substitution	AE4	AE1	44
substitution	IY4	IY2	43
substitution	IY1	IY2	38
substitution	IY2	IY1	37
substitution	NG2	NG1	37
substitution	N	L	36
substitution	D	L	34
substitution	UW1	UW4	34
substitution	AE2	AE3	33
substitution	AE1	AE4	31
substitution	IY4	IY3	31
substitution	L	D	29
substitution	AE2	AE1	28
substitution	IY1	IY3	28
substitution	JH	SH	28
substitution	NG2	NG3	28
substitution	AA4	AA1	26
substitution	IY3	IY4	26
substitution	UW4	UW1	26
substitution	Q	J	25
substitution	CH	SH	24
substitution	OW4	OW3	24
substitution	AE1	AE3	23
substitution	N1	N4	23
substitution	UW3	UW4	23
substitution	Q	X	22

Figure 2: List top of tonal phoneme substitutions in Magicdata. experiment 1.

the changes in experimental settings do not affect in nearly any way the contribution to PER results for the conversational speech.

Lastly Figure 5 compares the contribution to PER for experiment 4 across the two given corpora. By analyzing Figure 5 alongside Table 1 it can be observed that in the case of read speech, most of the non-tonal phonemes (except [D] and [J]) have higher contributions to PER when compared with their spontaneous speech instances. Moreover, it can be noted that nearly all of the tonal phonemes present in spontaneous speech have a higher contribution to PER when compared to the read speech instances. When more thoroughly considered, the just described observations emphasize the difference between read and spontaneous speech, namely the variable nature of the later one. With this observation, it is possible to assume that considering spontaneous speech most of the errors in the overall PER will originate from tonal phonemes, mostly due to the shared basis of the phoneme acoustic structure. Meanwhile, most of the errors that will contribute to the overall PER in the case of read speech will most probably come from non-tonal phonemes due to their unique and less variable acoustic structure thus requiring less clear stable spelling for the tonal phonemes.

5 Responsible Research

This is the first research that aimed to evaluate the performance of phoneme recognition using TDNN-BLSTM on Mandarin read and spontaneous speech thus to ensure the reproducibility of the research section 2 gave a brief description of the evaluation metrics, their respective formulas as well as providing insights into the corpora used for training the NN model. However, it is worth mentioning that the Magicdata corpus is not freely available which may or may not be a problem for future research.

Besides that, the experimental settings such as the composition of the feature vector and the NN parameters as well as the actual architecture of the TDNN-BLSTM were presented (see figure 1 and section 3.2). Furthermore, section 3.4 talked about possible restrictions of the experimental setup that may have influenced the final results. Lastly, previous and similar setup researches that were used to compare and support the findings could be found in the references.

6 Conclusions and Future Work

This research aimed to test and evaluate the performance of one of the proposed state-of-art APR systems on Mandarin read and spontaneous speech. This section presents a summary of this study's findings and compares them to other relevant research with similar setup [5] [7], as well as gives an overview of possible future work.

The overall results of TDNN-BLSTM on the two Mandarin speech styles were presented in section 4.1, the results however are not very promising and the reason for this (see section 3.4) was the small size of the subset of the corpora used for training.

At the same time, the results indicate that the variations in the structure of the NN hardly ever affected the results of the overall and individual phonemes independently of the speech style. Furthermore, the obtained results, as well as those of [7], suggest that the only parameter variations that could improve the performance of the NN are the number of cells per NN layer and the learning rates. Thus future research could take into consideration these findings when optimizing their NN parameters.

Comparing the results of the two speech styles side by side indicates that there exists a difference in performance between tonal and non-tonal phonemes in Mandarin speech. Section 4.3 demonstrates that tonal phonemes resulted in more errors, whilst non-tonal phonemes had fewer errors in the case of spontaneous speech when compared to read speech. Additional research of other tonal languages could be done to test if this observation holds as it may indicate the required distribution of tonal and non-tonal phonemes in the training set for achieving better results.

Lastly, this research compared the performance of TDNN-BLSTM on Mandarin speech and Dutch speech from previous research [5]. This comparison reveals the origins of the performance difference of TDNN-BLSTM across different types of languages, namely tonal and non-tonal languages. Indeed the results of section 4.2 point out that the presence of tones notably affects the precision with which the NN correctly predicts phonemes. This can be seen by the significant rise in the percentage of substitutions errors which are mostly between tonal phonemes that differ only in their digit suffixes for example [AA1] and [AA2]. Thus future work should consider using different feature vectors that could capture the tonal information of Mandarin speech. Additionally larger corpora with more accurate phoneme transcripts and higher distribution of tonal phonemes could be considered.

To conclude the goal of this research was to test the performance of the TDNN-BLSTM architecture on Mandarin read and spontaneous speech. It has done so in several ways.

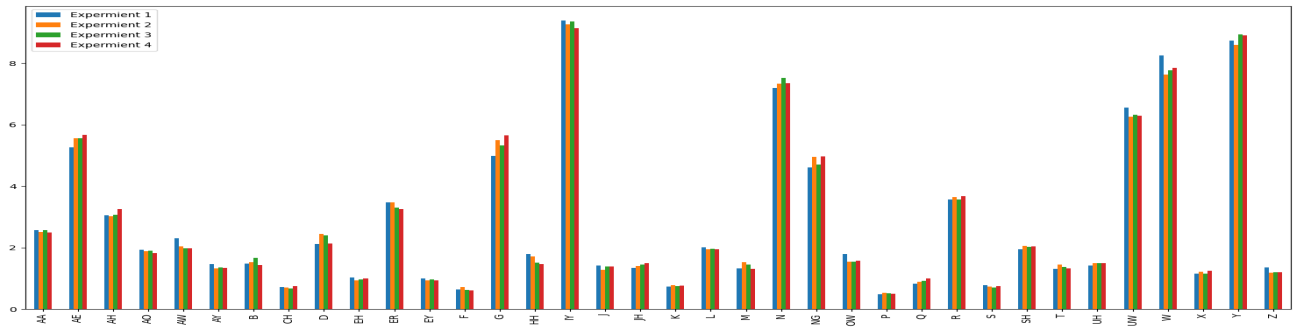

Figure 3: $\%ContributionToPER_{phoneme_x}$ Aidatatang_200zh.

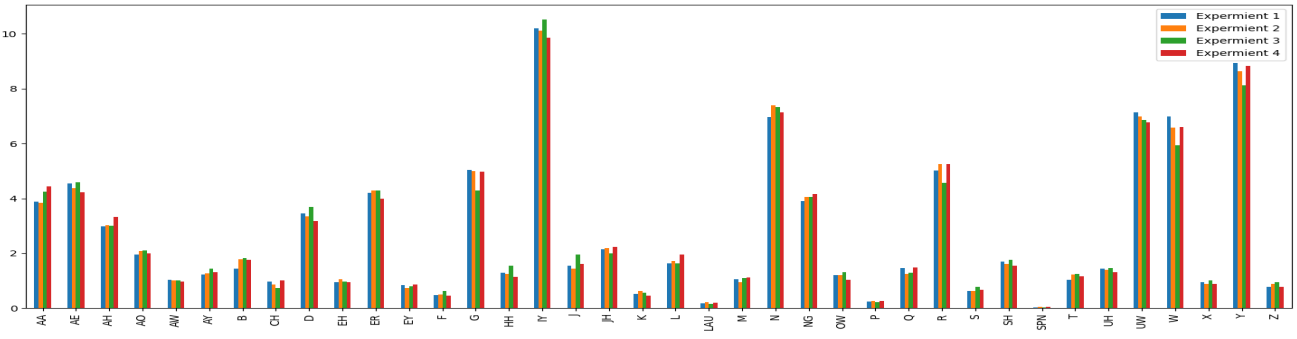

Figure 4: $\%ContributionToPER_{phoneme_x}$ Magicdata.

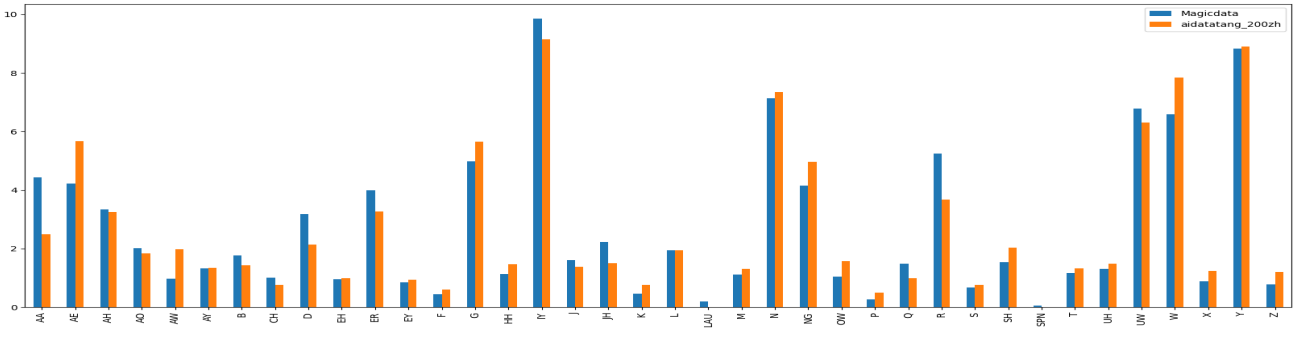

Figure 5: $\%ContributionToPER_{phoneme_x}$ for Experiment 4.

This work has analysed and compared how phoneme recognition using the TDNN-BLSTM architecture performed on the two different speech styles as well as different languages. It also analyzed whether the obtained results are dependent of the TDNN-BLSTM architectural variations as well as identified the necessary settings for model optimization. Hopefully this research enable future improvement and popularity of the APR systems.

References

[1] Aidatatang_200zh. https://openslr.org/62/. Accessed on 2021-Jun-07.

[2] Kaldi ASR. https://kaldi-asr.org. Accessed on 2021-Jun-07.

[3] D. H. Beun, L. C. W. Pols, and H. Kloosterman. Phoneme-based automatic speech recognition: towards a demonstrator for information retrieval, using dutch hi-fi speech. *IFA-Proceedings*, (19):126–134, 1995.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv*, (1406.1078), 2014.

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

[6] B. Liu, W. Zhang, X. Xu, and D. Chen. Time delay recurrent neural network for speech recognition. *Journal of Physics: Conference Series*, 1229(1), 2019.

[7] J. van der Tang. Evaluation of phoneme recognition through tdnn-opgru on mandarin speech. *BSc thesis, Delft University of Technology.*, 2021.

[8] H. van Geffen, M. Smit, A. Warners, F. Warners, and T. Yarally. A review of deep neural network-based phoneme recognition systems. *BSc group project, Delft University of Technology*, 2019.

[9] R. J. J. H. Van Son and L. C. Pols. Phoneme recognition as a function of task and context. *Proc. Institute of Phonetic Sciences University of Amsterdam*, 24:27–38, 2001.

[10] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, (37(3)):328–339, 1989.

[11] Wikipedia. Levenshtein distance. https://en.wikipedia.org/wiki/Levenshtein_distance. Accessed on 2021-Jun-07.