A prototype of Chinese Aspirated Consonants Pronunciation Training System Based on Multi-Resolution Cochleagram

Katarzyna A. URBANIEC
AGH University of Science and Technology,
Department of Mechanics and Vibroacoustics,
al. Mickiewicza 30, 30-059 Kraków, urbaniec@agh.edu.pl

Abstract

Many Mandarin Chinese learners, especially those whose mother tongue's phonological system differs significantly from Chinese phonological system, find it challenging to learn pronunciation of Chinese phonemes. Yet pronunciation training in language class settings is limited. It is therefore essential to develop computer-aided training system to help learners practice Chinese pronunciation without teacher's assistance. In this article I introduce a prototype of Chinese pronunciation training system that specifically focuses on phoneme substitution errors related to aspiration of consonants. I describe feature extraction process based on multi-resolution cochleagram (MRCG), a psychoacoustic model of basilar membrane excitation pattern, and architecture of recurrent neural network (RNN) used for mispronounced phonemes detection. The system achieves 96.12% and 98.58% accuracy rate in detecting phoneme substitution errors and determining aspiration length respectively. Proposed system may be particularly useful for learners of Slavic and Romance origin, since in their mother tongues aspiration is not a distinctive feature.

***Keywords*:** computer-aided pronunciation training (CAPT), mispronunciation detection, phoneme substitution

1. Introduction

Since China became a world's largest economy and opened up to the world, encouraging economic cooperation and foreign investment, more and more people have been required or willing to learn Mandarin Chinese as a second language. Because of its phonological and tonal systems, Chinese is found to be difficult to learn, especially by beginner-level learners, whose mother tongue is not a tonal language [1, 2]. Shortage of qualified teachers and limited pronunciation training in formal class settings make it even more challenging. A solution to these problems are computer-aided language learning (CALL) systems incorporating computer-assisted pronunciation training (CAPT), which identify a specific pronunciation error in an utterance and provide a corrective feedback without any human assistance [3].

In most of approaches to CAPT, mispronunciation detection is performed by extending automatic speech recognition (ASR) technologies such as Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs) or hybrid HMM-ANN systems [4]. Many of them employ Log-Likelihood Ratio (LLR) between non-native like and native-like models to detect pronunciation errors. The most representative system using LLR is Goodness Of Pronunciation (GOP) introduced by Witt and Young in 2000. The GOP measure incorporates a set of HMMs trained using mel-frequency cepstral coefficients (MFCC) and provides score for each phone in an utterance. A phone-specific threshold is set based on global GOP statistics and applied to each of the scores to decide, which of the

phones are mispronounced [5]. GOP methodology was used in numerous works, e. g. [6-12]. Above-mentioned LLR-based scoring methods achieve good pronunciation errors detection. Nevertheless, they don't provide user with diagnostic feedback about specific errors made by them such as phoneme insertion, deletion and substitution [13]. This feedback is obtained by using so called extended recognition networks (ERN). ERN is a representation of the canonical pronunciations and possible mispronunciations of a word. Using this representation reduces computational cost of ASR algorithms by avoiding searches in superfluous phone paths [14, 15]. A corrective feedback is also obtained by using ANNs, which map phonemic context information and acoustic features into phonemic posterior probabilities [1, 11, 16, 17].

Difficulties in learning proper Chinese pronunciation encountered by non-English native speakers are rarely discussed. There is only one work mentioning pronunciation problems specific for Mandarin Chinese learners of Slavic origin [18]. These problems are caused predominantly by lack of aspirated consonants in phonological systems of Slavic languages. Since there is no system designed to meet Slavic learners needs available, it is particularly important to develop one.

The aim of this work is to design and implement a prototype pronunciation training system that specifically focuses on pronunciation errors related to aspiration: deaspiration of aspirated consonants and voicing of voiceless not aspirated consonants. The system incorporates multi-resolution cochleagram (MRCG), a psychoacoustic model of basilar membrane excitation pattern [19]. Features extracted from MRCG are fed to recurrent neural network (RNN) detecting mispronounced phonemes and providing feedback about length of aspiration. With this knowledge user can alter the aspiration length adequately and improve their pronunciation skills. Proposed system can also help learners of Romance origin, since in their mother tongues aspiration is not a distinctive feature and aspiration-related errors are common among them [18].

2. System design

Proposed CAPT system is meant to help learners, who do not have access to qualified teachers, practice and obtain native-like pronunciation of Chinese aspirated consonants. Multi-resolution cochleagram (MRCG), composed of four 64-channel cochleagrams at different resolutions (hereafter referred to as CG1, CG2, CG3 and CG4), was used to model frequency selectivity properties of human cochlea and ensure, that extracted features and differences between them may be in fact perceived by human.

2.1. *Speech corpus*

The speech corpus used in this study is a subset of of native speech corpus AISHELL-1 [21]. Based on analysis of Mandarin Chinese phonological system, 12 phonemes were chosen to be included in designed CAPT system: 6 aspirated consonants (k^h, t^h, ts^h, $tʂ^h$, $tɕ^h$, p^h) and their not aspirated equivalents (k, t, ts, $tʂ$, $tɕ$, p) [1]. All above-mentioned consonants may occur only as syllable-initial phonemes and there are no consonant clusters in Mandarin language [20]. Thus, setting the length of analyzed signal to 180 ms ensures, that only one consonant will be included in this signal and the remaining frames not corresponding to any of these consonants will correspond to vowels

or glides. 360 utterances spoken by 15 male speakers and 15 female speakers were used: 300 of them constitute a training and validation data set and remaining 60 constitute a testing data set.

2.2. *MRCG-based classifier*

The mispronunciation detection system was implemented in Python. Speech signals were divided into 20 ms frames overlapping by 50%. For each signal 256-dimensional MRCG features and their deltas and double deltas were calculated using algorithm described in [19]. Kurtosis (Kurt) and spectral moments of 2nd, 3rd and 4th order (M_2, M_3, M_4) were also calculated for each channel of CG1, CG2, CG3, CG4, and their deltas and double deltas. Since MRCG is known to be noise robust [19, 22], signal filtering was omitted. To reduce computational complexity and feature redundancy, dimensionality reduction using recursive feature elimination (RFE) method was performed. 35 most relevant features shown in Table 1 were chosen and used in further analysis.

Table 1. 35 most relevant MRCG-features used in further analysis

Most relevant MRCG-features according to RFE results			
25th CG1 channel	63rd CG3 channel	33rd ΔCG3 channel	32nd ΔΔCG2 channel
26th CG1 channel	64th CG3 channel	34th ΔCG3 channel	10th ΔΔCG3 channel
29th CG1 channel	20th ΔCG1 channel	24th ΔCG4 channel	15th ΔΔCG3 channel
36th CG1 channel	25th ΔCG1 channel	22nd ΔΔCG1 channel	22nd ΔΔCG3 channel
58th CG1 channel	52nd ΔCG1 channel	49th ΔΔCG1 channel	M_2(ΔCG4)
26th CG2 channel	24th ΔCG2 channel	59th ΔΔCG1 channel	M_4(ΔΔCG3)
42nd CG2 channel	31st ΔCG2 channel	64th ΔΔCG1 channel	Kurt(CG2)
58th CG2 channel	60th ΔCG2 channel	4th ΔΔCG2 channel	Kurt(ΔΔCG1)
62nd CG3 channel	32nd ΔCG3 channel	8th ΔΔCG2 channel	

Above-mentioned features were normalized and fed to recurrent neural network (RNN) implemented in Keras and trained for 225 epochs. The RNN's architecture is shown in Figure 1. After each recurrent layer (gated recurrent unit, GRU, or long-short term memory, LSTM), dropout layers were added to avoid model's overfitting. Output layer is a dense layer with softmax activation function, trained under cross-entropy regime. Based on obtained probability distributions, each frame of analyzed signal was classified into one of 13 classes: *k*, *t, ts*, *tʂ*, *tɕ*, *p*, k^h, t^h, ts^h, $tʂ^h$, $tɕ^h$, p^h and *vowel*.

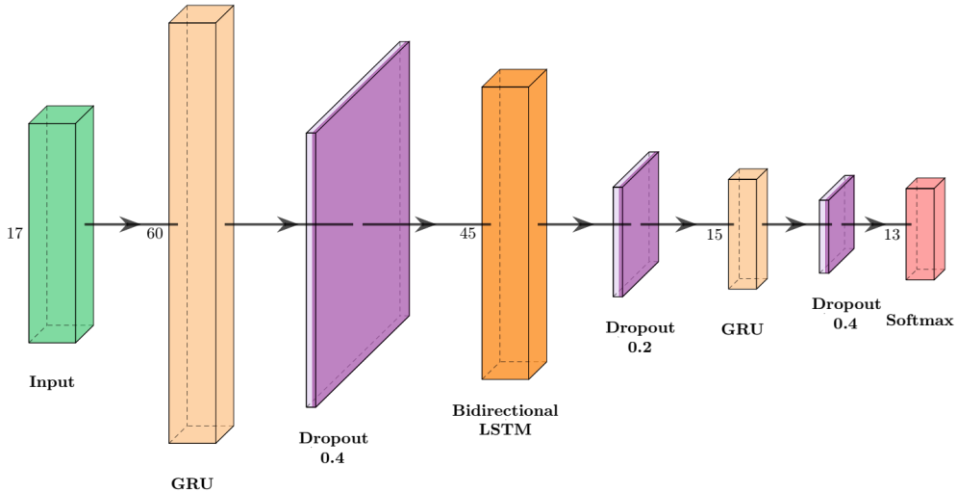

Figure 1. Diagram of RNN architecture

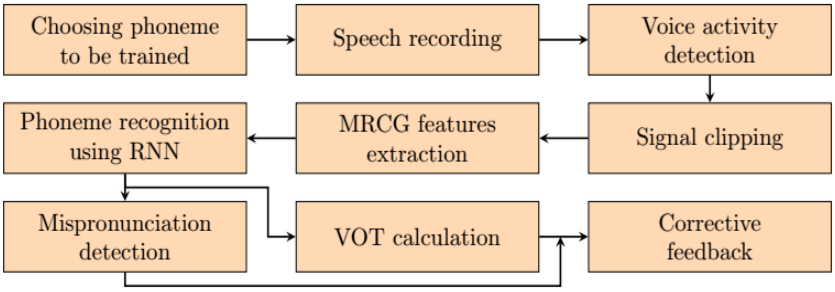

Figure 2. Block diagram of proposed CAPT system

Based on the frame length and the number of frames classified as not a vowel, voice onset time (VOT) was estimated. VOT value combined with information about possible mispronunciations detected by RNN is used to give user a corrective feedback about the reason of incorrect pronunciation (phoneme substitution, wrong aspiration length or both). The block diagram of designed CAPT system is shown in Figure 2.

2.3. *Evaluation metrics*

The performance of proposed MRCG-RNN-based CAPT system is evaluated using three metrics, namely recall, precision and accuracy:

$$Recall = \frac{N_{\mathrm{C}}}{N_{\mathrm{PH}}} \cdot 100\% \tag{1}$$

$$Precision = \frac{N_{\mathrm{C}}}{N_{\mathrm{D}}} \cdot 100\% \tag{2}$$

$$Accuracy = \frac{N_{\mathrm{FC}}}{N} \cdot 100\% \tag{3}$$

where N_{PH} is number of utterances containing analyzed phoneme, N_{C} is number of correctly detected phonemes, N_{D} is number of detected phonemes (both correctly and incorrectly), N_{FC} is number of frames classified correctly, and N is number of frames analyzed.

3. Results and discussion

Table 2 shows the performance of implemented RNN classifier. Precision and recall were calculated for each of 12 consonants incorporated in proposed CAPT system. International Phonetic Alphabet (IPA) is one of the most popular phonetic transcription system, but for Mandarin there is no one accepted IPA system [23]. Thus, in Table 2 consonants are displayed using both IPA and pinyin (romanization system for Mandarin Chinese).

The system achieves high recall and precision rates for recognition of all analyzed phonemes. Its performance is significantly better than performance of state-of-the-art CAPT system described in [2]. RNN classifier used to detect mispronunciations achieves high accuracy rate of 96.12%. The worst results were obtained for ts^h (recall 86.67%) and k^h (precision 85.30%). The reason for this may be that observed VOT differences for ts^h-*ts* and k^h-*k* pairs were relatively small for signals contained in used data sets. Classifier's performance could be improved by increasing number of samples in training data set.

To evaluate system's performance in determining aspiration length, VOT value estimated using classification results of single frames was compared with reference value obtained manually. The results show that proposed system achieves 98.58% accuracy rate in aspiration length determination. It should be noted, however, that the accuracy of approach to VOT estimation used in this work is limited by frame length and using other approach may be worth considering [24].

Table 2. Performance evaluation results of RNN classifier used in proposed CAPT system

Consonant		Precision	Recall
Pinyin	IPA		
b	*p*	100%	96.67%
p	p^h	87.88%	96.67%
d	*t*	100%	96.67%
t	t^h	100%	93.33%
c	ts^h	100%	86.67%
z	*ts*	93.55%	96.67%
j	*tɕ*	93.55%	96.67%
q	$tɕ^h$	100%	96.67%
ch	$tʂ^h$	96.77%	100%
zh	*tʂ*	87.88%	96.67%
g	*k*	100%	100%
k	k^h	85.30%	96.67%

4. Conclusions

I have proposed a prototype CAPT system focusing on pronunciation errors related to aspiration made by Mandarin Chinese learners. It incorporates MRCG, a psychoacoustic model of basilar membrane excitation pattern. Using this model in feature extraction process ensures that the system uses the same features as humans do to differentiate phonemes. MRCG is also known to be noise robust, therefore proposed system is suitable to be used in an environment with relatively low signal to noise ratio.

The system achieves high recall, precision and accuracy rates in phoneme recognition and pronunciation errors detection (86.67-100%, 85.30-100% and 96.12% respectively). Obtained results are significantly better than results described in other works, e. g. [2]. The system also achieves high accuracy rate in aspiration length determination (98.58%). It makes the system particularly useful for Mandarin Chinese learners of Slavic and Romance origin, since proper aspiration is difficult for them.

Although the system's performance is satisfactory, an additional evaluation using non-native speech corpus should be performed. Moreover, increasing the number of signals included in data set used for RNN's training could also improve its performance. Decreasing frame length to 10 ms or using other approach to VOT estimation should also be considered in future work.

References

1. W. Li et al., *Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling*, IEEE International Conference on Acoustics, Speech and Signal Processing, (2016) 6135 – 6139.
2. H.-C. Liao et al., *A prototype of an adaptive Chinese pronunciation training system*, System, **45** (2014) 52 – 66.
3. C. Molina et al., *ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion*, Speech Communication, **51** (2009) 485 – 498.
4. X. Qian, H. Meng, F. Soong, *A Two-Pass Framework of Mispronunciation Detection and Diagnosis for Computer-Aided Pronunciation Training*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **24.6** (2016) 1020 – 1028.
5. S. M. Witt, S. J. Young, *Phone-level pronunciation scoring and assessment for interactive language learning*, Speech Communication, **30** (2000) 95 – 108.
6. B. Mak et al., *PLASER: pronunciation learning via automatic speech recognition*, Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, (2003) 23 – 29.
7. A. Neri, C. Cucchiarini, H. Strik, *ASR corrective feedback on pronunciation: does it really work?*, Proceedings of Interspeech, (2006) 1982 – 1985.
8. J. Zheng et al., *Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation*, IEEE International Conference on Acoustics, Speech, and Signal Processing, (2007) 201 – 204.
9. C. Cucchiarini, A. Neri, H. Strik, *Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback*, Speech Communication, **51** (2009) 853 – 863.
10. H. Strik et al., *Comparing different approaches for automatic pronunciation error detection*, Speech Communication, **51** (2009) 845 – 852.
11. W. Hu et al., *Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers*, Speech Communication, **67** (2015) 154 – 166.
12. G. Huang et al., *English Mispronunciation Detection Based on Improved GOP Methods for Chinese Students*, Proceedings of International Conference on Progress in Informatics and Computing, (2017) 425 – 429.
13. Sh. Mao et al., *Applying Multitask Learning to Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech*, IEEE International Conference on Acoustics, Speech and Signal Processing, (2018) 6254 – 6258.
14. A. M. Harrison et al., *Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training*, Proceedings of ISCA International Workshop on Speech and Language Technology in Education, (2009) 45 – 48.

15. W.-K. Lo, Sh. Zhang, H. Meng, *Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System*, Proceedings of Interspeech, (2010) 765 – 768.
16. Sh. Mao et al., *Unsupervised Discovery of an Extended Phoneme Set in L2 English Speech for Mispronunciation Detection and Diagnosis*, IEEE International Conference on Acoustics, Speech and Signal Processing, (2018) 6244 – 6248.
17. K. Li, X. Qian, H. Meng, *Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **25.1** (2017) 193 – 207.
18. N. F. Chen et al., *Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL*, Speech Communication, **84** (2016) 46 – 56.
19. J. Chen, Y. Wang, D. Wang, *A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios*, IEEE International Conference on Acoustics, Speech and Signal Processing, (2014) 1993 – 2002.
20. L.-H. Wee, M. Li, *Modern Chinese Phonology* In W. S.-Y. Wang, Ch. Sun, *The Oxford handbook of Chinese linguistics*, Oxford University Press, (2015) 474 – 489.
21. H. Bu et al., *AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline*, 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, (2017) 1 – 5.
22. X.-L. Zhang, D. Wang, *Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection,* Proceedings of Interspeech, (2014) 1534 – 1538.
23. K. K. Y. Lam, C. K. S. To, *Speech sound disorders or differences: Insights from bilingual children speaking two Chinese languages*, Journal of Communication Disorders, **70** (2017) 35 – 48.
24. Ch.-Y. Lin, H.-Ch. Wang, *Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection*, Journal of the Acoustical Society of America, **130.1** (2011) 514 – 525.