INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Automatic scoring at multi-granularity for L2 pronunciation

Binghuai Lin[1], *Liyuan Wang*[1], *Xiaoli Feng*[2], *Jinsong Zhang*[2]

[1]Smart Platform Product Department, Tencent Technology Co., Ltd, China
[2]College of Information Science, Beijing Language and Culture University, China
{binghuailin, sumerlywang}@tencent.com, fengxiaoli314@163.com, jinsong.zhang@blcu.edu.cn

Abstract

Automatic pronunciation assessment and error detection play an important part of Computer-Assisted Pronunciation Training (CAPT). Traditional approaches normally focus on scoring of sentences, words or mispronunciation detection of phonemes independently without considering the hierarchical and contextual relationships among them. In this paper, we develop a hierarchical network which combines scoring at the granularity of phoneme, word and sentence jointly. Specifically, we achieve the phoneme scores by a semi-supervised phoneme mispronunciation detection method, the words scores by an attention mechanism, and the sentence scores by a non-linear regression method. To further model the correlation between the sentence and phoneme, we optimize the network by a multitask learning framework (MTL). The proposed framework relies on a few sentence-level labeled data and a majority of unlabeled data. We evaluate the network performance on a multi-granular dataset consisting of sentences, words and phonemes which was recorded by 1,000 Chinese speakers and labeled by three experts. Experimental results show that the proposed method is well correlated with human raters with a Pearson correlation coefficient (PCC) of 0.88 at sentence level and 0.77 at word level. Furthermore, the semi-supervised phoneme mispronunciation detection achieves a comparable result by F1-measure with our supervised baseline.

Index Terms: Pronunciation assessment and error detection, multi-granular, multitask learning, hierarchical network, attention

1. Introduction

Non-native speakers are heavily influenced by their own native tongue (L1) when learning the target language (L2). The common approach to tackle this problem is through Computer-Assisted Pronunciation Training (CAPT). A system for CAPT should not only provide an overall assessment of pronunciation but also detailed feedback on pronunciation error for language learners [1].

Features used for pronunciation assessment and error detection are usually extracted from the hidden Markov model (HMM) of an automatic speech recognizer. HMM likelihood, posterior probability and pronunciation duration features were proposed for pronunciation assessment in [2]. A variation of the posterior probability ratio called the Goodness of Pronunciation (GOP) [3] was proposed for pronunciation evaluation and error detection [4, 5]. GOP was further optimized based on deep neural network (DNN) to improve the accuracy of phoneme mispronunciation detection [6]. With the development of deep neural network, the feature learning and pronunciation assessment or error detection can be optimized jointly. Long short-term memory recurrent network (LSTM) was adopted to extract feature representations for features such as speech attributes and phone features for mispronunciation detection and pronunciation assessment [7, 8]. A convolution neural network (CNN) was used to extract features for pronunciation error detection based on the MLP classifier [9]. A Siamese network was developed to extract features of distance metrics at the phone instance level for pronunciation assessment [10]. Other methods such as APM (acoustic-phonemic model) calculated the phone-state posterior probabilities from acoustic features and phonetic information based on DNN to generate the recognized phoneme sequence for phoneme mispronunciation detection [11, 12].

These systems focus on improving either the pronunciation evaluation or error detection independently. To provide L2 learners an overall pronunciation assessment, methods for scoring at multi-granularity levels are required. Some systems combined global evaluation and detailed feedback together [13]. It provided multi-granular pronunciation feedback of word and sentence independently. However, phoneme, word and sentence are not independent of each other and the larger granularity such as sentence or word contributes to contextual facilitation effects for the smaller one such as word or phoneme [14, 15].

In this paper, we propose a hierarchical network to score at multi-granularity jointly considering the structure of both sentence and word. The contextual dependence among phoneme, word and sentence is modeled by particular mechanisms between layers of the hierarchical network. To further capture the correlations of sentence and phoneme, an MTL framework is proposed to combine the semi-supervised phoneme mispronunciation detection and sentence scoring. In section 2, we will introduce the proposed network. The experiments are conducted in section 3. We will draw the conclusions and future suggestions in section 4.

2. Proposed method

We propose a hierarchical network for L2 language learners' pronunciation error detection and evaluation. The network is composed of three layers: phoneme, word and sentence layer. The phoneme layer accepts phoneme features and provides outputs for the word layer as well as an auxiliary phoneme mispronunciation detector. The word layer utilizes the information from the phoneme layer to calculate word scores based on individual phoneme contributions. The word scores are then fed into the third sentence layer to obtain the final sentence score. The whole network shares the phoneme feature representation and is optimized by MTL which combines the main scoring task and the auxiliary phoneme mispronunciation detection task. Figure 1 shows the proposed network. We demonstrate the whole system in the following sections in detail.

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-1282

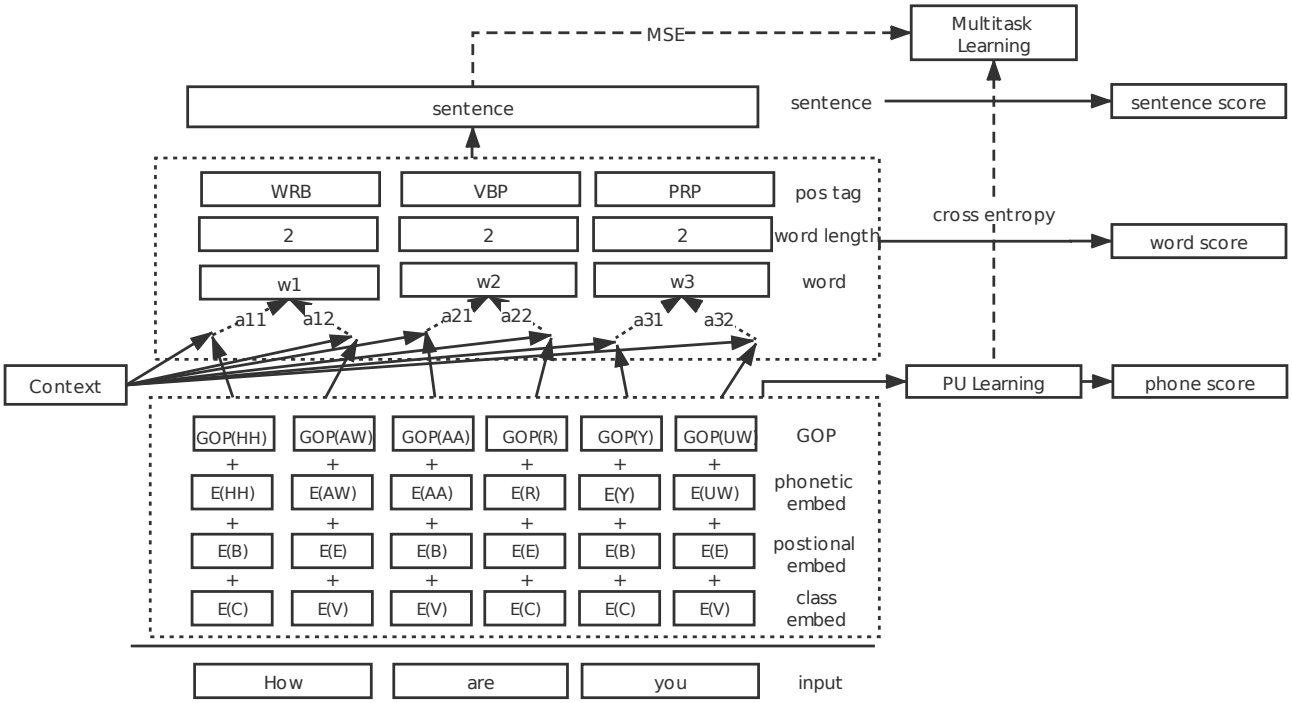

Figure 1: *An attention based semi-supervised hierarchical network with multitask learning*

2.1. Feature representation

GOP is used as one feature for our proposed network. GOP score is defined as Eq. (1) [3]:

$$\mathrm{GOP}(p) = \frac{|\log(P(p|o^p)|}{\mathrm{NF}(p))} = \frac{|\log(\frac{P(o^p|p)P(p)}{\sum_{q\in Q} P(o^q|q)P(q)})|}{\mathrm{NF}(p)} \quad (1)$$

where $P(p|o^p)$ is the posterior probability of phoneme p given pronunciation o and Q represents all possible phonemes corresponding pronunciation o. $\mathrm{NF}(p)$ is the pronunciation frames of phoneme p and $P(p)$ is prior probability of phoneme p. To calculate GOP, the phonemes of utterances are forced-aligned first by a Kaldi-based Automatic Speech recognition (ASR) system [16].

Other features affecting pronunciation are taken into account. First, the phoneme pronounces differently depending on its positions in the word [17]. We denote phoneme positions by 'B', 'I', 'E', 'S', which represent the beginning, middle, ending positions in a word as well as single-phoneme words. Second, as vowels and consonants are of different importance in a word [14], we use 'C' and 'V' to represent phoneme classes of consonants and vowels separately. We also employ independent numerical representations for each phoneme. We encode these phoneme properties into numerical vectors which are called positional, phonetic class and phonetic embedding. We combine these feature representations with phoneme GOP as input of phoneme layer.

2.2. Semi-supervised learning on phoneme layer

As labeling phoneme mispronunciation is a time-consuming and labor-intensive task [18], we conduct phoneme mispronunciation detection based on a semi-supervised learning mechanism. Assuming native speakers' pronunciation as perfect, most native speakers' pronunciation can be regarded as positive samples. Phoneme mispronunciation detection problem can be converted into positive and unlabeled learning problem. Some studies are related to positive and unlabeled learning problem. There is extensive research related to this particular positive and unlabeled learning problem. The method [19] converted the problem into a positive negative learning problem. It proved these two problems can be convertible with a constant factor difference. The method [20] took unlabeled data as a combination of positive samples and negative samples.

In an L2 learners' dataset, the GOP usually varies within a specific range for each phone and lower GOP values can be regarded as negative samples. The problem can then be converted to a negative and unlabeled learning problem. The proposed method [21] combined positive unlabeled (PU) and negative unlabeled (NU) learning and proved the combined method can reduce the generalization error bound and risk variance. The PUNU learning loss function can be defined in Eq. (2):

$$R^{\gamma}_{\mathrm{PUNU}}(g) = (1-\gamma)R_{\mathrm{PU}}(g) + \gamma R_{\mathrm{NU}}(g) \quad (2)$$

where

$$\begin{aligned} R_{\mathrm{PU}}(g) = \theta_{\mathrm{P}} E_{\mathrm{P}}[l(g(x),1)] + E_{\mathrm{U}}[l(g(x),-1)] \\ -\theta_{\mathrm{P}} E_{\mathrm{P}}[l(g(x),-1)] \end{aligned} \quad (3)$$

and

$$\begin{aligned} R_{\mathrm{NU}}(g) = \theta_{\mathrm{N}} E_{\mathrm{N}}[l(g(x),-1)] + E_{\mathrm{U}}[l(g(x),1)] \\ -\theta_{\mathrm{N}} E_{\mathrm{N}}[l(g(x),1)] \end{aligned} \quad (4)$$

where g is an arbitrary decision function and l is the loss function where the value $l(t,y)$ means the loss incurred by predicting an output t when the ground truth is y. E_{U}, E_{P} and E_{N} denote the loss expectation over unlabeled, positive, negative class marginals, respectively. θ_{P} and θ_{N} represent class-prior probability of positive and negative samples. γ varies between

0 and 1 and denotes balanced weights between positive unlabeled and negative unlabeled loss. In our mispronunciation detection task, positive and negative samples indicate phonemes with correct and wrong pronunciation.

In our PUNU learning framework, we take native pronunciation as positive samples, L2 learners' pronunciation with low GOP values as negative samples, and the rest L2 pronunciation as unlabeled data. The final mispronunciation detection subnetwork is optimized by the aforementioned PUNU learning loss function.

2.3. Attention based learning on word layer

The word layer takes results from phoneme layer to calculate word scores. Each word consists of one or more phonemes, and each phoneme in a particular word makes different contribution to the final word score. For example, phoneme 'o' mispronounced in the word dog usually contributes more to the word score than the phoneme 'g' depending on the context. This concept can be implemented by the attention mechanism which has gained popularity in training neural network [22, 23]. In our word scoring layer, we assign different weights for each phoneme in the word. This yields,

$$U_{\mathrm{p}} = \tanh(w * O_{\mathrm{p}} + b) \tag{5}$$

$$\alpha_{\mathrm{p}} = \frac{\exp(U_{\mathrm{p}}^T U_{\mathrm{w}})}{\sum_{q \in w} \exp(U_{\mathrm{q}}^T U_{\mathrm{w}})} \tag{6}$$

$$S_{\mathrm{w}} = \sum_{p \in w} \alpha_{\mathrm{p}} O_{\mathrm{p}} \tag{7}$$

where O_{p} is the score of phoneme p. w, b are the trainable weights and biases of word layer. U_{w} is a randomly-initialized vector which can be taken as a memory unit of the word context. We measure the importance of the phoneme in a word based on the similarity between U_{w} and the transformed phoneme score, and then normalize the results as shown in Eq. (6). Finally, we compute the word score as a weighted sum of phoneme scores in Eq. (7).

2.4. Multitask learning on sentence layer

The sentence layer takes output from the word layer to calculate the final sentence score. As sentence is composed of words, each word in a sentence can make different contributions to the sentence score depending on word attributes such as the part-of-speech (POS) tagging and the number of phonemes in a word [24]. Word layer outputs, word lengths, and POS tagging are combined as the input features to fit expert scores by a non-linear regression method.

To further model the relationship between the phoneme and sentence, we apply an MTL framework to the pronunciation scoring task by combining the phoneme mispronunciation detection and sentence pronunciation scoring tasks. The MTL is also a good solution when there are multiple related tasks with limited training samples for each [25]. Taking phoneme mispronunciation detection as an auxiliary task, we combine these two tasks by the MTL framework. Specifically,

$$L_{\mathrm{total}} = (1 - w) \times L_{\mathrm{sent}} + w \times L_{\mathrm{phoneme}} \tag{8}$$

where L_{sent} is the mean square error loss of the sentence scoring and L_{phoneme} is the aforementioned PUNU loss. w is a constant value balancing these two tasks.

3. Experiments

3.1. Introduction of datasets

The corpus consists of 22,998 English utterances read by 1000 Chinese speakers with ages evenly distributed from 16 to 20 years and the Timit dataset [26]. The total numbers of sentence, word scoring and phoneme mispronunciation labeling are 8998, 4000, and 10000 utterances, respectively. The average number of words in sentence scoring is 13, and the total number of phonemes in mispronunciation labeling is 99568. Three experts rated word and sentence score on a scale of 1-5 with 1 representing hardly understandable pronunciation and 5 representing native-like pronunciation. The averaged inter-rater correlations at the sentence and word levels, which are calculated by PCC between scores of one rater and average scores of the rest raters, are 0.78 and 0.76. By averaging scores from three experts, we obtain sentence and word scores ranging from 1 to 5. Phoneme mispronunciation was voted by three experts, and a phoneme is treated as mispronounced with two or three votes. The inter-rater labeling consistency of phoneme mispronunciation is evaluated at Kappa, which is calculated by averaging any two raters based on 1000 sentences randomly chosen, and the final value is 0.65 with the 95% confidence interval $(0.647, 0.653)$ and p-value less than 0.1%, indicating a fairly good quality of labeling.

The training data of our experiments is composed of 7998 sentences of non-native speakers with labeled scores and 5000 sentences of native speakers without scores. The testing data composed of three parts: 4000 utterances with 39808 phonemes labeled, 1000 scored words and 1000 scored sentences. The testing data of phoneme mispronunciation is based on Carnegie Mellon University (CMU) Pronouncing Dictionary composed of 39 different phonemes [27]. The percentage of phonemes labeled as mispronounced is about 14%. The distribution of phoneme mispronunciation is shown in the Figure 2. The x-axis shows CMU phonemes and the y-axis represents the number of the corresponding phoneme. The upper white bar is the number of mispronounced phoneme while the lower black bar is the number of phonemes with correct pronunciation. The number shown above each bar is the mispronounced ratio of the particular phoneme.

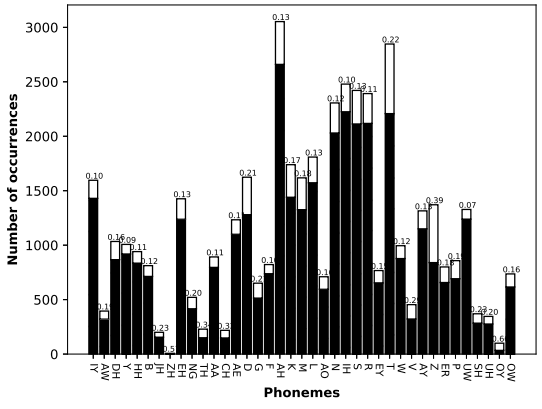

Figure 2: *The distibution of manual phoneme mispronunciation*

The distribution of human sentence scores and word scores is shown in the Figure 3. The x-axis is the averaged scores of human raters and the y-axis represents the occurrence of the

corresponding score range.

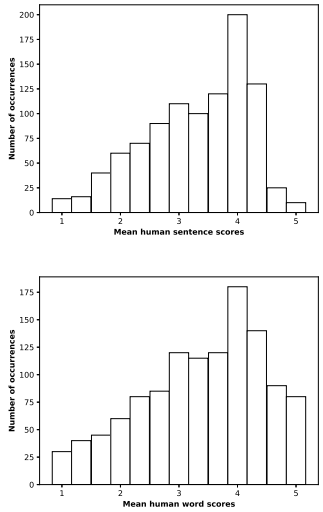

Figure 3: *The distibution of human sentence and word scores*

3.2. Performance of overall network

We evaluate the overall network from three perspectives: sentence, word and phoneme. The performance of sentence score and word scores is evaluated by PCC. Following previous work [8, 12], the phoneme mispronunciation detection performance is evaluated by F-measure, false rejection rate (FRR) and false acceptance rate (FAR).

3.2.1. Performance of sentence score

The baseline model consists of two BLSTMs followed by a multilayer perceptron (MLP) layer with the standard logistic sigmoid activation function (2BLSTM + MLP) [7]. The first BLSTM takes the phone GOP as well as the phonetic positioning, phonetic class (vowel or consonant) and phonetic embeddings same as our model's input. The outputs of the last hidden units are concatenated as input for the next BLSTM. Additionally, word properties including word POS tagging and lengths are fed to the second BLSTM. Then the MLP is applied over the concatenated representations from the second BLSTM to obtain the sentence score. To demonstrate the effect of multitask learning for sentence scoring, we compare results of MTL with the same proposed network optimized only by the sentence evaluation task (Ours (STL)).

We compare results from two perspectives: the mean squared error (MSE) and PCC. The sentence results are shown in Table 1. Though the BLSTM model obtains comparable MSE as our proposed model, it has an PCC lower by 2%, indicating our system correlated better with human ratings. Furthermore, the 3% gap between STL and MTL show that MTL can improve the performance of sentence scoring.

3.2.2. Performance of word score

We compare the result with the same proposed network without sentence scoring layer which is trained with 3000 labeled word data in a supervised way (Ours (SL)). The result is further compared with the aforementioned BLSTM network with the second BLSTM removed (i.e., BLSTM + MLP). The word

Table 1: *Sentence comparison between the BLSTM and ours*

Model	MSE	PCC
2BLSTM+MLP	0.033	0.83
Ours (STL)	0.031	0.85
Ours	0.030	0.88

PCC and MSE comparison results are shown as Table 2. The results indicate our model also performs well at the word level, given only the sentence labeled data. Despite the higher MSE of our model, the higher PCC compared with the BLTSM demonstrates that our system can capture human scoring trend well even with noisy labeled data. The 2% gap between BLSTM and Ours (SL) show the superiority of attention mechanism over BLSTM.

Table 2: *Word comparision result*

Model	MSE	PCC
BLSTM+MLP	0.033	0.72
Ours (SL)	0.033	0.74
Ours	0.042	0.77

3.2.3. Performance of phoneme mispronunciation detection

The baseline is achieved by utilizing the phoneme layer in our proposed network optimized in a supervised way (Ours (SL)) based on 6000 utterances with 59760 phonemes mispronunciation labeled. The results are shown in Table 3. The proposed method based on semi-supervised learning is only a little inferior to the supervised by 2% in F-measure and FAR and 0.1% in FRR, given no additional phoneme labeled data. The results also show 35% absolute gap between FRR and FAR, which is caused by low recall of mispronunciation detection. Nevertheless, we should pay more attention to FRR than FAR as high FRR may discourage L2 learners.

Table 3: *Comparison between the baseline and ours*

Model	Precision	Recall	F1	FRR	FAR
Ours (SL)	0.65	0.62	0.63	0.0169	0.378
Ours	0.64	0.59	0.61	0.0171	0.40

4. Conclusion

In this paper, we propose an automatic scoring method at multi-granularity levels for L2 learners. A hierarchical network is developed with the consideration of hierarchical and contextual influence among phoneme, word and sentence. The sentence scoring and semi-supervised mispronunciation detection of phoneme are combined by an MTL framework. Experimental results show the system correlates well with human raters at sentence and word levels. Meanwhile, a comparable performance is achieved in phoneme mispronunciation detection even without a large amount of labeled data. In the future, we will focus on the improvement of phoneme detection task to obtain better pronunciation error feedback for L2 learners.

5. References

[1] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: When technology meets pedagogy," 2002.

[2] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1471–1474.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," 2009.

[5] H. Ryu, H. Hong, S. Kim, and M. Chung, "Automatic pronunciation assessment of korean spoken by l2 learners using best feature set selection," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[6] W. Hu, Y. Qian, and F. K. Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call)." in *Interspeech*, 2013, pp. 1886–1890.

[7] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 338–345.

[8] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection for non-native learners with multisource information and lstm-based deep models." in *INTERSPEECH*, 2017, pp. 2759–2763.

[9] A. Lee *et al.*, "Language-independent methods for computer-assisted pronunciation training," Ph.D. dissertation, Massachusetts Institute of Technology, 2016.

[10] K. Kyriakopoulos, K. Knill, and M. Gales, "A deep learning approach to assessing non-native pronunciation of english using phone distances," 2018.

[11] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2016.

[12] S. Mao, Z. Wu, R. Li, X. Li, H. Meng, and L. Cai, "Applying Multitask Learning to Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6254–6258.

[13] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[14] H. Al-Barhamtoshy, S. Abdou, and K. Jambi, "Pronunciation evaluation model for none native english speakers," *Life Science Journal*, vol. 11, no. 9, pp. 216–226, 2014.

[15] G. B. Simpson, R. R. Peterson, M. A. Casteel, and C. Burgess, "Lexical and sentence context effects in word recognition." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 15, no. 1, p. 88, 1989.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[17] J. S. Bowers, N. Kazanina, and N. Andermane, "Spoken word identification involves accessing position invariant phoneme representations," *Journal of Memory and Language*, vol. 87, pp. 71–83, 2016.

[18] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, "Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall," *Speech Communication*, vol. 84, pp. 46–56, 2016.

[19] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 213–220.

[20] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *Advances in neural information processing systems*, 2017, pp. 1675–1685.

[21] T. Sakai, M. C. du Plessis, G. Niu, and M. Sugiyama, "Semi-supervised classification based on classification from positive and unlabeled data," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2998–3006.

[22] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[23] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[24] R. Kingdon, *The groundwork of English stress*. Longmans, 1958.

[25] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[27] R. L. Weide, "The cmu pronouncing dictionary," *URL: http://www.speech.cs.cmu.edu/cgibin/cmudict*, 1998.