*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# Deep feature transfer learning for automatic pronunciation assessment

*Binghuai Lin*$^\star$, *Liyuan Wang*$^\star$

Smart Platform Product Department, Tencent Technology Co., Ltd, China
{binghuailin, sumerlywang}@tencent.com

## Abstract

Automatic pronunciation assessment is commonly developed to evaluate pronunciation quality of second language (L2) learners. Traditional methods for automatic pronunciation assessment normally utilize speech features such as Goodness of pronunciation (GOP), which may not provide sufficient information for the pronunciation proficiency assessment [1]. In this paper, we propose a transfer learning method for automatic pronunciation assessment. We directly utilize the deep features from the acoustic model instead of traditional features such as GOP, and transfer the acoustic knowledge from ASR to a specific scoring module. The scoring module is designed to consider the relationship among different granularities in an utterance based on an attention mechanism. Only this module is updated for faster transfer and adaptation of various pronunciation assessment tasks. Experimental results based on the dataset recorded by Chinese English-as-second-language (ESL) learners and the Speechocean762 [1] dataset demonstrate that the proposed method outperforms the traditional GOP-based baselines in Pearson correlation coefficient (PCC) and yields parameter-efficient transfer for different pronunciation assessment tasks.

**Index Terms**: Pronunciation assessment, transfer learning, scoring module, pre-train, fine-tune

## 1. Introduction

Non-native speakers are heavily influenced by their native tongue (L1) when learning the target language (L2). The typical approach to tackle this problem in computer-assisted language learning is through Computer-Assisted Pronunciation Training (CAPT). A system for CAPT can provide overall assessment of pronunciation for language learners.

The common feature used for automatic pronunciation assessment is GOP, which is defined as the posterior probability of the uttered phoneme given the corresponding acoustic segment [3]. GOP was first computed using Gaussian mixture model-hidden Markov model (GMM-HMM)-based acoustic models [3], and it has prevailed in the related studies ever since [4, 5, 6]. With the development of deep neural network (DNN), it has been demonstrated that DNN-HMM-based ASR can reduce word error rate (WER) significantly compared with GMM-HMM models [7]. As a result, GOP was further optimized based on the DNN-HMM-based acoustic model to improve the performance of Computer-Aided Language Learning (CALL) [8][9]. It has been found there are many limitations in the current DNN-HMM-based GOP computation. As DNN-HMM-based formulations didn't consider transition probabilities, previous work combined HMM state transition probabilities (STPs) with the sub-phonemic (senone) posterior to compute GOP scores [10]. Two major limitations were pointed out in GOP computation and two factors called the transition factor and the duration factor were introduced to get context-aware GOP scores [11].

Based on GOP scores extracted from ASR, many tasks such as automatic pronunciation assessment and mispronunciation detection are conducted. Phone-specific GOP thresholds were predefined to determine the pronunciation error, and it was proved to agree with human scores well [3][11]. Some previous work created a classifier to determine the mispronunciation automatically, and it achieved high accuracy in detecting pronunciation errors [12][13]. Pronunciation scoring at the word or sentence level is obtained by averaging the phone level GOP scores, and it achieved high correlation with the expert ratings [8][10]. A multi-granularity scoring network based on GOP scores was proposed to obtain pronunciation scores at the phone, word, and sentence levels simultaneously [14]. Some other methods were proposed to update acoustic model and phone GOP parameters simultaneously to ensure objective improvement by maximum F1-score discriminative criterion (MFC) training [15][13]. These methods for pronunciation assessment or mispronunciation detection depend heavily on GOP scores extracted from the ASR systems.

Pronunciation assessment based on the acoustic model can also be treated as a problem of domain adaptation. Optimization of the acoustic model is a task from the source domain, while pronunciation assessment is in the target domain. Domain adaptation can be solved by transfer learning, which refers to extracting knowledge from the source domain and applying the knowledge to the target domain [16]. The typical way to conduct transfer learning is fine-tuning the model trained on the source task using data from the target domain [17]. Also, it has been found the strictness of labeling varies in different pronunciation tasks [3]. Thus, a fast adaptation to various pronunciation assessment tasks is necessary.

In this paper, we propose a transfer learning method for automatic pronunciation assessment. Different from traditional GOP-based methods, we directly utilize the deep features of the acoustic model and transfer the acoustic knowledge for pronunciation assessment. A scoring module following the acoustic model of ASR is designed to evaluate the proficiency of the whole utterance considering the relationship among phonemes, words, and sentences based on the attention mechanism. The network training is composed of three stages. First, the DNN-HMM-based acoustic model is pre-trained in a normal ASR training procedure to obtain deep audio features. Second, deep features from the acoustic model are fed into the scoring module to conduct a second-stage pre-training based on low-quality scoring data, which is necessary for better adaptation for pronunciation assessment. Finally, the network is fine-tuned with a small amount of task-specific pronunciation assessment data for efficient adaptation to various tasks. Experimental results based on the dataset recorded by Chinese ESL learners and Speechocean762 dataset [2] show the superiority of the proposed

[*] equal contribution

[1] https://www.openslr.org/101/

http://dx.doi.org/10.21437/Interspeech.2021-931

network to the traditional GOP-based methods and the efficiency of transfer learning for various pronunciation assessment tasks. In section 2, we will introduce the proposed method. The experiments are conducted in section 3. We will draw the conclusion and future suggestions in section 4.

# 2. Proposed method

We propose a transfer learning method for automatic pronunciation assessment based on the deep features of the ASR acoustic model without intermediate feature computation. Figure 1 shows the comparison between the traditional GOP-based methods and our proposed method. The whole network is shown in Figure 2, and the detailed scoring module is shown in Figure 3 . The network consists of two parts: the DNN-HMM-based acoustic model and the attention-based scoring module. The acoustic model comes from a pre-trained ASR model to extract acoustic features for pronunciation assessment. The attention-based scoring module is used to model the relationship among frames, phonemes, words, and sentences as well as to score the utterance.

## 2.1. DNN-HMM-based acoustic model

The acoustic model of DNN-HMM-based ASR has multiple layers between input and output. It takes acoustic features as input and outputs the posterior probability of the senone given observation with acoustic frames associated defined as Eq. (1) [18][19], where $p_j$ is the output class probability, $x_j$ is $jth$ output before the SoftMax layer, $k$ is an index over all classes, $o_t$ is the observation vector, and $s_t$ is the senone state at time $t$. It is optimized by cross-entropy, whose frame-level labels are obtained based on forced alignment using a GMM-HMM system. The posterior probability of the senone can be converted back into the HMM's emission likelihood defined as Eq. (2) [20].

$$P(s_t|o_t) = p_j = \frac{e^{x_j}}{\sum_k e^{x_k}} \tag{1}$$

$$P(o_t|s_t) = \frac{P(s_t|o_t) \times P(o_t)}{P(s_t)} \tag{2}$$

## 2.2. Scoring module for pronunciation assessment

The scoring module takes the deep features from the acoustic model and the alignment information from the pre-trained ASR as input to score the utterances as shown in Figure 3. As an utterance is composed of multiple words, phonemes, and frames, we exploit the intrinsic relationship among them based on a hierarchical network. Given the alignment information, we can obtain the acoustic representations for each phoneme by averaging the deep features of the corresponding frames in the phoneme. As each phone type in the phone set shows distinct characteristics [3], we employ independent numerical representations for each phoneme, which is called phoneme embedding. We sum the phoneme embedding to the phoneme-level acoustic representations to obtain the final phoneme representations. As different phonemes in an utterance may make different contributions to the final scores, we assign different weights to phonemes in an utterance. Specifically, we apply the multi-head self-attention mechanism to the phoneme-level features, which has prevailed in the natural language processing (NLP) applications [21]. The self-attention mechanism is used for computing the representation of a sequence by paying different attention to different positions in the sequence. We use the self-attention mechanism to calculate the weighted representations for each phone in a word, and then obtain the corresponding word representations by averaging these phone representations followed by a nonlinear transformation. Similarly, we consider different words in an utterance contribute differently to the utterance score [22]. The self-attention mechanism is applied to the word-level representations as well. Finally, we calculate the utterance-level representations by averaging the weighted word representations in an utterance. Based on the utterance-level representations, we apply a non-linear transformation followed by a sigmoid activation function to obtain a score ranging from 0 to 1.

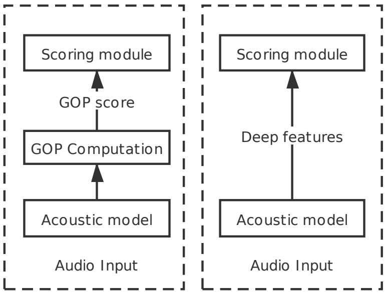

Figure 1: *Comparison between GOP-based methods (left) and ours (right)*

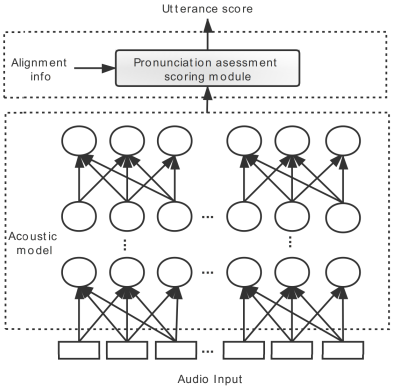

Figure 2: *Proposed network structure*

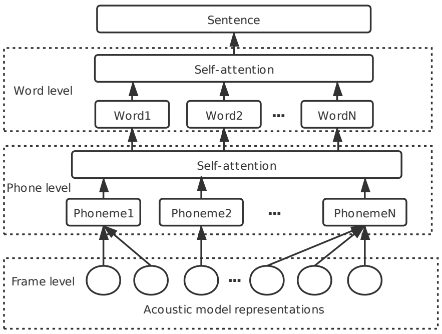

Figure 3: *The detailed scoring module*

## 2.3. Training scheme

We conduct a three-stage training scheme for the whole network. First, the DNN-HMM-based acoustic model is pretrained in a standard training procedure of the ASR system before outputting deep features to the scoring module. Second, to better adapt the network to assessment tasks as well as facilitate faster transfer to various tasks, we further pre-train the network utilizing a significant amount of synthetic scoring data. The scores are generated from a simple GOP-based scoring model, which may be of low-quality. We experiment with both frozen and unfrozen weights of the acoustic model in this training stage. Finally, based on the transferred network, we update the scoring module with a small amount of high-quality data labeled by human experts. The last two training stages are both optimized by a mean square error loss between the predicted scores $p$ and their corresponding gold-standard scores $y$ defined as Eq. (3), where $i$ denotes $ith$ utterance and $n$ means the number of utterances.

$$L_{score} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2 \tag{3}$$

# 3. Experiments

## 3.1. Corpus

The corpus consists of data for ASR training and pronunciation scoring. The ASR training data includes the 960-hour native LibriSpeech corpus [23], and the 1000-hour non-native corpus recorded by Chinese teenagers with age evenly distributed from 16 to 20 years. The male and female speakers are distributed evenly as well. The pronunciation scoring data include the synthetic data for the second-stage pre-training and human-labelled data for fine-tuning. The synthetic data consists of $50,000$ of the recorded Chinese teenagers' utterances and the corresponding scores generated from a GOP-based regressor.

For the last-stage fine-tuning, we utilize two different human-labelled datasets to demonstrate the efficiency and effectiveness of fast adaptation of the proposed scoring module. The first dataset consists of $11,000$ English utterances read by $1,000$ Chinese speakers with ages evenly distributed from 16 to 20 years. The average number of words in the sentences is 13. Three experts rated sentence pronunciation on a scale of 1-5, with 1 representing hardly understandable pronunciation and 5 representing native-like pronunciation. The averaged inter-rater correlation at the sentence level, calculated by Pearson correlation coefficient (PCC) between scores of one rater and average scores of the other two, is $0.78$. By averaging scores from three experts, we obtain continuous sentence scores ranging from 1 to 5. The whole dataset is divided into $10,000$ sentences for training and $1,000$ for testing. The score distribution of the test set is shown in Figure 4. The x-axis represents the average scores of human raters, and the y-axis represents the number of occurrences within the corresponding score range. The second dataset comes from a public dataset called Speechocean762. It is recorded by 250 non-native speakers, where half of the speakers are children. It consists of $5,000$ English sentences with multi-granularity labeling - each phoneme, words, and sentence is labeled, and with an overall score assigned to each sentence. Half of the dataset is used for training while the other half is used for testing. We scale the scores to the range of 0 and 1.

## 3.2. Experimental setup

The DNN-HMM-based acoustic model consists of a time delay neural network with layers number of 11 [24]. We utilize deep features of the acoustic model as the input for the scoring module. The deep features and phoneme embedding have a dimensionality of 256 and 32, respectively. Non-linear transformation with $256 \times 32$ parameters has been applied to the deep features before being added to the phoneme embedding. The head number of the self-attention modules at the phone and word levels is 4. The parameters of phone-level, word-level and sentence-level non-linear transformation have dimensionality of $32 \times 32$, $32 \times 32$, and $32 \times 1$, respectively. Three layers with layer normalization (LayerNorm) are added to the frame, phone, and sentence layers, respectively, for faster convergence [25]. The number of parameters is shown in Table 1, indicating the scoring module has only $0.2\%$ of parameters compared with the acoustic model.

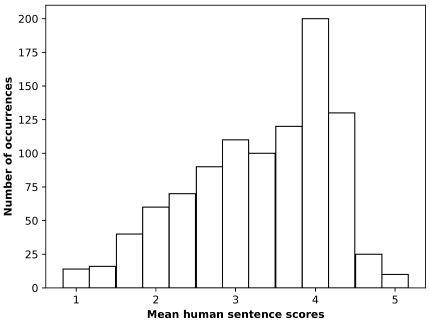

Figure 4: *The distribution of human-labelled scores*

Table 1: *Parameters of the network*

| Model | Parameters |
|---|---|
| Acoustic model | 8165047 |
| scoring module | 20289 |

## 3.3. Comparative study

We evaluate the proposed method by calculating PCC of the sentence scoring. Three experimental settings are presented to demonstrate its effectiveness: (1) comparison with GOP-based baselines for pronunciation assessment; (2) performance with different settings of pre-training and fine-tuning; (3) ablation studies about the rationality of the proposed scoring structure. The human-labelled datasets introduced in 3.3.3 are denoted as L2 set and Speechocean762 set, respectively.

### *3.3.1. Comparison with GOP-based baselines*

We conduct two experiments in this part: (1) comparison with the previous state of the art GOP-based methods; (2) comparison to our proposed scoring module with the traditional or improved GOP as input instead of the deep features.

First, we compare our method with two traditional baselines, including a neural network (NN)-based model and Gradient Boosting Tree (GBT)-based model with phone-level GOP scores as input. The GOP score is derived from the previous work [3]. The NN-based model consists of two Bidirectional Long Short-Term Memory (BLSTM) layers followed by a multilayer perceptron (MLP) layer with the standard sigmoid activation function (2BLSTM + MLP) [26]. The first BLSTM layer takes the phone GOP as well as the phoneme embeddings as input. The outputs of the last hidden units are concatenated as input for the next BLSTM layer. The MLP layer is then applied over the concatenated representations from the second BLSTM to obtain the sentence score. Features for the GBT-based model include the average GOP score of each phoneme, the average position of each phoneme in a sentence, which is the beginning (1), middle (2), ending position (3) in a word as well as the position representing a single-phoneme word (4), and the total number of phoneme in a sentence. Results of the comparison are shown in Table 2.

Table 2: *Comparison between the GOP-based baselines and ours*

| Model | L2 set | Speechocean762 set |
|---|---|---|
| GOP (2BLSTM+MLP) [26] | 0.83 | 0.67 |
| GOP (GBT) | 0.80 | 0.65 |
| Ours | 0.86 | 0.72 |

From the results, we can see the GBT-based method is inferior to the BLSTM-based method by nearly 2% in PCC, indicating it is beneficial utilizing a complex neural network-based structure. Our proposed method outperforms the GOP-based baselines in PCC by $4\%$ at least, indicating our results correlate better with human scores.

Second, we compare our method to the same scoring module with GOP as input instead. We replace the deep features from the acoustic model with the traditional GOP scores [3] or an improved GOP considering STPs (STPs GOP) [10]. The results in Table 3 demonstrate the superiority of taking the deep features as input compared with GOP scores.

Table 3: *Comparison with different inputs for the proposed scoring module*

| Model | L2 set | Speechocean762 set |
|---|---|---|
| GOP [3] | 0.81 | 0.64 |
| STPs GOP[10] | 0.84 | 0.66 |
| Ours | 0.86 | 0.72 |

### *3.3.2. Performance with different settings of the second-stage pre-training and fine-tuning*

We do not experiment with changing the first-stage pre-training of the acoustic model. However, we do experiment with variants of the second pre-training stage for transferring acoustic knowledge from the acoustic model and with the fine-tuning stage for adaptation to different pronunciation assessment tasks. The experiments are conducted with the following settings: training the scoring module with the frozen or unfrozen weights of the acoustic model, respectively, and fine-tuning and not fine-tuning the scoring module when conducting new tasks. The results are shown in Table 4 and 5.

From the results, we can see the experimental setting with frozen acoustic model and fine-tuning achieves the best performance with a PCC of $0.86$ for L2 testing set and $0.72$ for the Speechocean762 set. The necessity of the second-stage pre-training is clear when compared with the results without pre-training. During the second stage, freezing the acoustic model performs better than the unfrozen one by $3\%$ in PCC based on these two datasets. Moreover, the process of fine-tuning can further improve the performance for specific assessment tasks. Given a particular task, only $0.2\%$ of parameters need to be fine-tuned and thus it proves to be parameter-efficient.

Table 4: *Performance of different settings of the second-stage pre-training and fine-tuning based on L2 set*

| Pre-train | Fine-tune | Training | Testing |
|---|---|---|---|
| frozen | frozen | 0.87 | **0.86** |
| unfrozen | unfrozen | 0.87 | 0.84 |
| unfrozen | frozen | 0.84 | 0.83 |
| unfrozen | – | – | 0.80 |
| – | frozen | 0.80 | 0.81 |

Table 5: *Performance of different settings of the second-stage pre-training and fine-tuning based on the Speechocean762 set*

| Pre-train | Fine-tune | Training | Testing |
|---|---|---|---|
| frozen | frozen | 0.82 | **0.72** |
| unfrozen | unfrozen | 0.84 | 0.71 |
| unfrozen | frozen | 0.80 | 0.69 |
| unfrozen | – | – | 0.65 |
| – | frozen | 0.80 | 0.68 |

### *3.3.3. Ablation studies for the scoring module*

The scoring module takes deep features at the frame level as input and outputs scores at the utterance level. The self-attention mechanism exploits the relationship among phones, words, and sentences. We conduct experiments to replace the self-attention module with a simple averaging module, where word and sentence representations are obtained by averaging phone and word representations. The results are shown in Table 6. From the results, we can see that the self-attention-based scoring correlates better with the human raters than simply averaging the information based on the sentence structure.

Table 6: *Models with and without self-attention modules*

| Model | L2 set | Speechocean762 set |
|---|---|---|
| Without self-attention | 0.84 | 0.69 |
| With self-attention | 0.86 | 0.72 |

## 4. Conclusion

In this paper, we propose a transfer learning method for automatic pronunciation assessment. Instead of relying on complex and approximate feature computation, such as GOP scores, we directly utilize the deep features from the acoustic model. A scoring module followed by deep features is designed to model the human scoring based on the self-attention mechanism. For efficient adaptation to different pronunciation assessment tasks, we conduct a second-stage pre-training with a significant amount of synthetic scoring data, and fine-tune the scoring module with a small amount of task-specific scoring data. Results based on the datasets of Chinese ESL learners and Speechocean762 show the proposed method outperforms the traditional GOP-based baselines. In the future, we will focus on combining mispronunciation detection with the proposed scoring method to obtain more detailed pronunciation feedback for L2 learners.

# 5. References

[1] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "ASR-free pronunciation assessment," *arXiv preprint arXiv:2005.11902*, 2020.

[2] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[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," in *SLaTE*, 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] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, "Pronunciation assessment of japanese learners of french with GOP scores and phonetic information," *Proc. Interspeech 2016*, 2016.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] 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.

[9] ——, "An improved DNN-based approach to mispronunciation detection and diagnosis of l2 learners' speech." in *SLaTE*, 2015, pp. 71–76.

[10] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities," *Proc. Interspeech 2019*, 2019.

[11] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *arXiv preprint arXiv:2008.08647*, 2020.

[12] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[13] H. Huang, H. Xu, Y. Hu, and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3165–3177, 2017.

[14] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for l2 pronunciation," *Proc. Interspeech 2020*, pp. 3022–3026, 2020.

[15] H. Huang, J. Wang, and H. Abudureyimu, "Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[17] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: transfer learning through adaptive fine-tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4805–4814.

[18] A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, 2012.

[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[20] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 24–29.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

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

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] 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.