INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An Analysis of Goodness of Pronunciation for Child Speech

Xinwei Cao[1], *Zijian Fan*[1], *Torbjørn Svendsen*[1], *Giampiero Salvi*[1,2]

[1]Department of Electronic Systems, NTNU, Norway
[2]KTH Royal Institute of Technology, EECS, Sweden

{xinwei.cao, zijian.fan, torbjorn.svendsen, giampiero.salvi}@ntnu.no

Abstract

In this paper, we study the use of goodness of pronunciation (GOP) on child speech. We first compare the distributions of GOP scores on several open datasets representing various dimensions of speech variability. We show that the GOP distribution over CMU Kids, corresponding to young age, has larger spread than those on datasets representing other dimensions, i.e., accent, dialect, spontaneity and environmental conditions. We hypothesize that the increased variability of pronunciation in young age may impair the use of traditional mispronunciation detection methods for children. To support this hypothesis, we perform simulated mispronunciation experiments both for children and adults using different variants of the GOP algorithm. We also compare the results to real-case mispronunciations for native children showing that GOP is less effective for child speech than for adult speech.

Index Terms: speech assessment, mispronunciation detection and diagnosis, data scarcity, child speech, ASR, GOP

1. Introduction

Computer-Aided Language Learning (CALL) is known to be a useful tool for learning new languages due to its ubiquitous availability and high degree of self-controlled manner of study. The CALL system can be equipped with a Computer Assisted Pronunciation Training (CAPT) component aiming at Mispronunciation detection and/or diagnosis (MDD) where detecting and correcting mispronunciations by the learner is vital. In this respect, systems vary depending on the detail of feedback they are able to provide (word or phoneme level).

One intuitive approach of MDD is to use speech from the "teacher" and utilizes pattern matching mechanisms to judge how different the input speech ("the student") is from the gold standard (the "teacher"). Within this direction, Lee et al. [1] experimented with dynamic time warping (DTW) using MFCC and Gaussian posteriorgram (based on a pre-trained universal background model) as features for the alignment between the teacher and the student. However, this kind of approach suffers from the limited number of stored audio templates from the teacher and it is often not reasonable to rely on one single piece of teacher's speech given that the variability in speech is huge, e.g, age, gender, accent, noise, etc.

This problem can be mitigated by utilizing well-trained acoustic models from an ASR system. The goodness of pronunciation (GOP) [2] is one of those approaches which requires only the canonical phoneme level transcription together with the student's speech, and a well trained ASR model, without the need of teacher's voice. Many variations of GOP were proposed to increase the accuracy, e.g., the Weighted-GOP (WGOP) [3], the context-aware-GOP (CaGOP) [4], the Lattice-based GOP (LGOP) [5], the Force-aligned GOP (FGOP) [6]. Neumeyer et al. [7] also investigated to use the likelihood-based scores from the Gaussian mixture model based hidden Markov model (GMM-HMM) for ASR, but unlike the GOP which was computed over phoneme segments, the likelihood ratio here was computed at frame-level and duration or the speed of the speech were considered at both word and sentence level to improve the accuracy. Maximum likelihood linear regression (MLLR), a widely used speech adaptation method for ASR, was studied for GOP in [6]. Hu et al. [8] compared three different approaches to implement DNN-based GOP and showed that the "senone average posterior" approach outperforms the other methods. Lin et al. proposed a transfer learning approach for training a hierarchical deep neural MDD system [9]. More recently researchers [10, 11] have explored to use a generally pretrained model as feature extractor, together with an end-to-end speech recognizer as fine-tuning task, and finally adapted to the target data for MDD. The advent of these approaches mitigated the problem of limited availability of annotated material for certain target groups of speakers, e.g., L2 learners or children with speech disorders, and achieved good performance. However, due to the end-to-end nature of the models, feedback was in these cases limited to a general pronunciation score, and no detailed phonetic information was provided by the system. The prior linguistic knowledge of mispronunciation for the target users can be injected to ASR systems by means of extended recognition network (ERN) [12]. Similarly, Dudy et al. [13] studied the data from children with speech disorders and explicitly modeled the typical error patterns in their proposed GOP (GOP-CI).

The motivation of this work is to investigate the potential of applying GOP-alike algorithms to child speech. In spite of being a slightly older technique, GOP has the advantage of giving detailed information about pronunciation quality at the phonetic level. This may constitute very valuable information for an L2 learner. The main challenges we face in this work are the large variability of child pronunciation and the scarcity of phonetically annotated material for second language (L2) learners of young age. To overcome the scarcity of material, we perform our analysis based on a number of openly available speech datasets spanning different dimensions of speech variability, including age, dialect or foreign accent, spontaneity and environment conditions. We firstly analyze the spread of the distributions of GOP scores (including possibly correct and incorrect pronunciations) on the different datasets. Secondly, we simulate mispronunciation detection by introducing artificial errors similarly to [14], but with systematic phoneme replacements that include all possible phoneme pairs. We report area under the curve results for both child and adult data. We also compare the statistics of artificial errors to real-case errors by human an-

10.21437/Interspeech.2023-743

notations for the child data. Finally, the above analysis was performed using several variants of the GOP score to show how the different methods compare when applied to real-case pronunciation errors for child speech.

2. Methods

We will first discuss the different GOP definitions used in this paper, and then the method and the assumptions we use to perform our analysis.

2.1. GOP definitions

There exist a large number of definitions and variations of GOP-alike methods. We start by introducing the original version from [14] and then explain the GOP methods that we use in our experiments. The original GOP [14] uses ASR methods to align the correct pronunciation to the speech recording by means of forced alignment. It then scores deviations in pronunciation of each phoneme with the following equation:

$$\text{GOP}(p) = \log\left(\frac{p(O|p)P(p)}{\sum_{q\in Q} p(O|q)P(q)}\right)/N_p \qquad (1)$$

where p represents the correct phoneme, $O = \{o_1, \ldots, o_{N_p}\}$ is the sequence of observations corresponding to the segment for phoneme p and is derived from forced alignment. Q is the set of phonemes of the target language and N_p is the length of the segment. $P(p)$ is the prior probability for phoneme p which is conventionally neglected. We will remove this term in the following discussion as it is not part of the implementation. GOP relies on the HMM-based ASR models to compute the probabilities in Eq. 1. In these models, each phoneme is often modeled as three successive states or senones. The numerator in Eq. 1 should be in principle computed with the forward algorithm. However, this likelihood is often approximated by the likelihood of the Viterbi path using the correct phoneme model. Similarly, the denominator is approximated by the likelihood of the best path through the segment O, obtained from a phone loop decoding of the entire utterance.

Some variations of GOP (e.g. LGOP) were motivated by the observation that the likelihood of the denominator may be under-estimated because running the phoneme loop over the entire utterance may introduce misalignments of the phoneme best path with respect of the forced alignment in the numerator. In other variants (e.g. FGOP), instead, it was observed that stronger constraints could improve the accuracy. To explore these two extremes, we use the following two versions of GOP throughout our experiments. The first version denoted as GOP-align is computed by:

$$\text{GOP-align}(p) = \log\left(\frac{p(O|\text{SS}(p))}{\max_{q\in Q} p(O|\text{SS}(q))}\right)/N_p \qquad (2)$$

where $\text{SS}(p)$ is phoneme p's senone sequence derived from full utterance forced alignment and restricted within the observation sequence O. For the denominator, for each phoneme in the inventory, we compute the Viterbi likelihood for the observations O given the corresponding phoneme model. Again, $\text{SS}(q)$ denotes the corresponding senone level best path through the model. GOP-align is considered to have the strongest constraint on the denominator because it forces the alignment for each phone model p to start at the beginning of the sequence O and end at its end. This method is similar to FGOP ([6]). The second method, GOP-frame, is defined by:

$$\text{GOP-frame}(p) = \frac{1}{N_p}\sum_{i=1}^{N_p} \log\left(\frac{p(o_i|\text{ss}(p,i)}{\max_{s\in S} p(o_i|s)}\right) \qquad (3)$$

where S is the set of senones for all the phonemes in our model, $\text{ss}(p,i)$ is the ith senone from the senone sequence of phoneme p derived from the forced alignment. The denominator is calculated by searching for the best senone in a frame-by-frame manner, i.e., the best state in the pool of models to explain the current observation o_i. In this case, we disregard the transition model in the HMM for the denominator, and allow any state of any model to contribute to the likelihood for each observation. This method is similar to the idea of "normalized likelihood metric" proposed by Neumeyer et al. in [7] and could be considered as the other extreme of GOP-align which has only minor constraint.

2.2. Analysis Methods

The goal of this paper is to perform an analysis of the performance of different GOP scores on child speech, in spite of the lack of well annotated data. Our analysis is limited to English, and is performed with three methods. The first method, in the lack of mispronunciation annotations, uses an indirect observation to infer the potential performance of GOP for child speech. The assumption is that, if GOP is a good indicator of pronunciation quality, it would be desirable that the spread of GOP scores be smaller for native adult and child speakers than for non-native adult speakers. To test this we estimate the distributions of GOP scores on different datasets with varying age, degree of spontaneity, and language proficiency and compare the spreads of these distributions.

The second method makes up for the lack of annotated non-native child speech recordings by simulating mispronunciations. The method is similar to [15], but we perform phoneme substitutions systematically considering any possible pair of phonemes and picking the worst case scenario (the least separable substitution). In this case, we use the area under the receiver operating characteristic curve (AUC-ROC) to evaluate the performance of the GOP score in detecting simulated mispronunciations. This score is independent of the threshold chosen to perform the binary classification.

In this formulation, the number of positive and negative examples is, in general, unbalanced. To verify that the AUC score is not strongly affected by this, we also tried to sub-sample the class with the most samples. Because the results were similar, we report results obtained on the full data, here.

The third method uses real mispronunciations (or rather phoneme substitutions) by children to validate the results from the simulated experiment. In this case, because we need phonetically annotated material, the amount of data is more limited than in the simulated case, but it is more representative of real mispronunciations. To detect what kind of substitutions are present in the spoken utterance we realign the canonical pronunciation from the lexicon to the phonetic annotations in the dataset.

In our analysis, we test different GOP definitions and different models to estimate the probabilities described in Eqns. 1–3 amd further explained in Section 3.

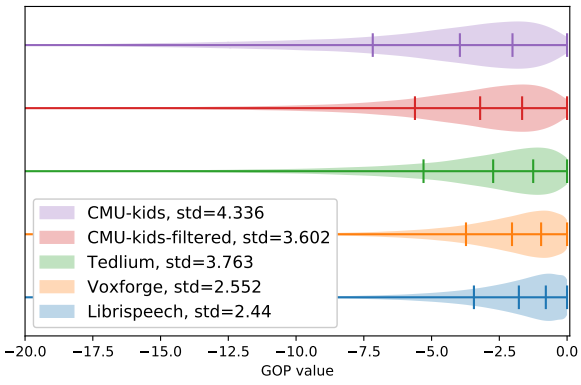

Figure 1: *GOP distributions per dataset using the DNN-HMM-tri-frame method.*

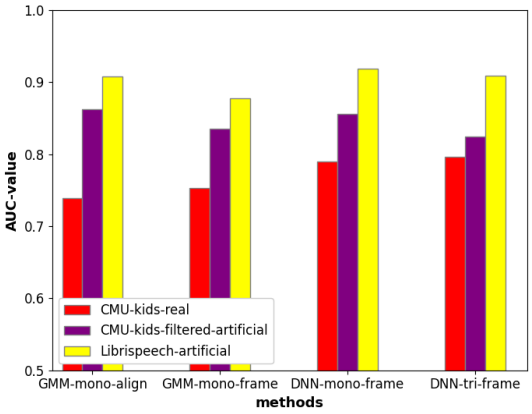

Figure 2: *AUC-average per GOP method on the Librispeech, complete CMU-kids and filtered CMU-kids datasets.*

3. Experiments

3.1. Data

We select four English datasets containing speech with different characteristics, both in terms of speaker and speaking style:

The LibriSpeech ASR corpus [16] is a well preprocessed dataset designed for ASR training in the domain of audio books. We use the part "train-100-clean" to train the teacher's ASR models. It contains 100 hours of book reading from 251 gender balanced adult speakers. For testing the GOP methods, we combine the "test-clean", "dev-clean", "test-other" and "dev-other" for a total of around 21 hours of audio read by 146 speakers. We include the dev parts in the test set because we do not use them for parameter optimization. Whereas the training speakers are mostly native speakers with US accent, the test set may contain some accented speech.

The second dataset is Voxforge [17] that contains recording submitted by volunteers reading prompted text. In our experiment, we randomly select 500 speakers and allow all possible dialects of English with different nationalities. There are around 27.5 hours speech in total. The content domain of Voxforge overlaps slightly with LibriSpeech but with a wider variability in terms of accents and dialects. In addition, since the recordings are performed by volunteers in an uncontrolled environment, this data also reflects a certain degree of environment noise, different microphones etc.

The third dataset is TEDLIUM [18]. The data is recorded by 1862 speakers with multi culture backgrounds and accents giving TED talks in English. We randomly select 10,000 utterances from various speakers which add up to 3.2 hours in total. In our view, this dataset represents stronger accent and spontaneity compared the the previous datasets. The environmental noise is also an issue here.

Finally, for the child speech, we used the full CMU Kids dataset [19]. This data is collected from children of age six to eleven in the US's local schools. CMU Kids has 9.1 hours of audio in 5180 utterances spoken by 24 male and 54 female speakers. Part of the speakers are at risk of becoming bad readers. The dataset is equipped with annotations for the erroneous utterances. In our tests we divided this data according to whether it contains erroneous utterances or not. The complete dataset is referred to as CMU-kids, and the dataset where incorrect pronunciations have been filtered out as CMU-kids-filtered.

3.2. Training the ASR teacher models

We use Kaldi [20] to train the ASR models used for the alignment and the probability estimation in the GOP scores. We trained four models, depending on if we include phonetic context in the modeling (mono vs tri) and if we use GMMs or DNNs as state output probability estimators. We use the Librispeech lexicon [16] covering 206495 English word pronunciations using a set of 39 phonemes. Each phoneme is further categorized with positional and stress information (if it is a vowel) plus some other affiliated tokens (silence, noise, disambiguate symbols) which sum up to 364 phonemes for our HMM in total. The number of independent states is decided during the training with a decision tree while the number of Gaussian in each GMM is dynamically adjusted iteratively. For the DNN, we use the standard feed-forward-structure with four hidden layers with p-norm activation [21] as in the standard Kaldi recipe. 13 dimensional MFCC features with cepstral mean normalization are used with all the models. For the GMM-based models, delta features are computed to increase the context awareness whereas for DNN-based models, a window of size 9 (± 4) is applied to the MFCC features and fed into a fixed Linear Discriminant Analysis (LDA) affine-transformation without dimension reduction. We use the same data (train-100-clean) for training all the models.

3.3. GOP implementations

We implement the GOPs as described in Section 2 using Kaldi[1]. The phoneme alignment is always performed using the context independent GMM-HMM model, to make sure all the GOP evaluations are performed on the same set of observations O (see Eqns. 1–3). The alignment not only provides information of segmentation of the phonemes, it also generates the senone-sequence as needed in the numerators of Eqn. 2 and Eqn. 3. Note that, for DNN models, the acoustic likelihood is calculated by $p(s_i|o_i)/P(s_i)$ where $p(s_i|o_i)$ is the softmax output and $P(s_i)$ is the prior for each senone estimated when training the model. This is equivalent to computing $p(o_i|s_i)$ because the missing term $p(o_i)$ is the same on the numerator and denominator for the GOP calculation, and cancels out.

In the following we will refer to the GOP evaluation method by combining the method for state output probability estimation (GMM vs DNN), the context dependency (mono vs tri) and the

[1]https://github.com/frank613/GOPs.git

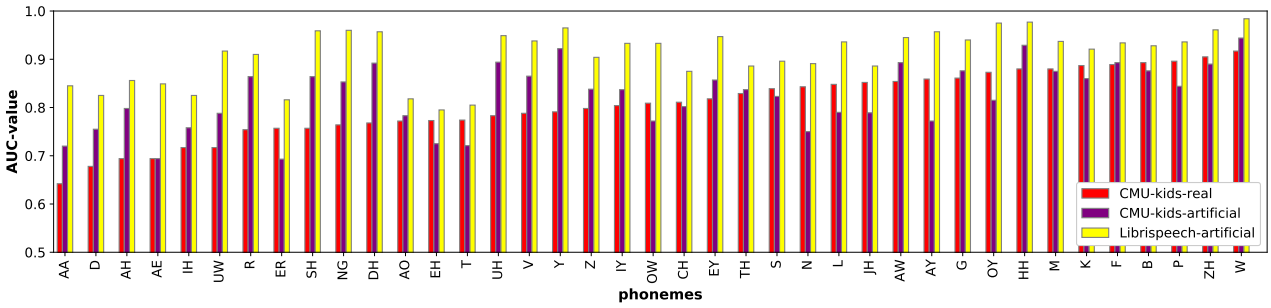

Figure 3: *AUC comparison for each phoneme, sorted by the AUC values for real errors of CMU-kids*

method for computing the GOP (align vs frame). For example GMM-mono-frame corresponds to the GOP computed by Eq. 3 and using the context independent GMM-HMM model as probability estimator.

4. Results and discussion

The results of the first analysis are reported in Figure 1 which plots the distribution of DNN-tri-frame GOPs over different datasets. We also produced similar plots with the other GOP methods obtaining similar results and decided to omit them for space constraints. A similar spread of the GOP distribution is observed for Librispeech and Voxforge. Both datasets include read speech (low spontaneity) from adult speakers. The larger variability of Voxforge in terms of native and non-native accents and recording conditions does not seem to have a strong impact on the GOP values. Moving on, the CMU-kids-filtered, containing native child speech with correct pronunciations, shows a higher spread of GOP scores compared to Librispeech and Voxforge, which was expected due to the larger variability and higher spontaneity of child speech. More surprisingly, however, the spread for CMU-kids-filtered is also higher than for the TEDLIUM data. The latter contains recordings by non-native adult speakers with a higher degree of spontaneity and should correspond to much larger GOP variability than the first. This seems to imply that age has a stronger impact on the GOP scores than possible mispronunciations or strong non-native accents. Finally the non-filtered version of CMU-kids which includes mispronunciations by children displays the widest spread. Although this analysis is not conclusive on the performance of GOP for each dataset, it should indicate a potential problem with using this pronunciation detection technique with children.

Motivated by the above results, we analyze the discriminative ability of GOP in terms of AUC values, using both simulated and real mispronunciations. In this case we focus on Librispeech and CMU-kids as representative of adult and child data. The corresponding results are shown in Figures 3 and 2.

Figure 2 compares the the AUC values when using different ASR models and different GOP definitions over the Librispeech and CMU-kids data. Results are reported for three cases: simulated mispronunciations both for adult speech (Librispeech-artificial) and child speech (CMU-kids-filtered-artificial) and real mispronunciation for children (CMU-kids-real). The results show that, regardless the model and the GOP definition, the adult simulated substitutions are easier to detect. Artificial substitutions for children are harder, but not as hard to detect than the real mispronunciations. This is expected because the main difference between the simulated and real errors is that the first assume a substitution between two completely different phonemes from the phoneme inventory (although we consider the closest substitution). In the case of real errors, instead, we consider any possible deviation from the correct pronunciation. The different methods do not report large differences for adult speech. However, for child speech results are slightly contradictory. For real mispronunciations, DNN-tri-frame is the best method, followed by DNN-mono-frame, GMM-mono-frame and GMM-mono-align. On the contrary, for simulated mispronunciations, GMM-mono-align seems to outperform the other methods. These results may be explained considering the different kind of errors (simulated vs real), however, we would need a more detailed analysis to confirm this hypothesis.

In Figure 3 the AUC values obtained with the DNN-HMM-tri-frame method are reported for each phoneme. Again simulated and real errors are used similarly to Figure 2. From the figure it can be seen that results are strongly dependent on the phoneme. However, in all cases, the AUC for adult speech is better than for child speech. Also, in the majority of cases, for the children real mispronunciations seem to be more difficult to discriminate than artificial ones, as was the case for the global results in Figure 2. The exceptions are S, AY, N, T, K, L, B, ER, P M, EH, and OW.

5. Conclusions

In this work, we present the analysis of goodness of pronunciation (GOP) scores applied to child data. Because of the scarcity of phonetically annotated data for second language learners of young age, we use a number of methods to infer the potential performance of GOP for mis-pronunciation detection. We first compared the distributions of GOP scores for different dimensions of speech variability observing a larger variability for child data as opposed to adult data, even when non-native speakers were included in the analysis. We then ran mis-pronunciation detection (MDD) on both simulated and real mispronunciations both for child and adult data. We show that MDD is always harder for child data than for adult speech. We also show that real mispronunciations are on average harder to detect than simulated ones, with some exceptions. This indicates that simulated mispronunciations could be a good method to investigate MDD in the case of data sparsity. We believe that our results indicate the need for further research in this area, both to produce more reliable data, and to propose methods that are more robust with respect to the age of the language learner.

6. Acknowledgements

This research has been funded by the TEFLON NordForsk project nr. 103893 and by the SCRIBE Research Council of Norway project nr. 322964.

7. References

[1] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 382–387.

[2] 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.

[3] J. Doremalen, C. Cucchiarini, and H. Strik, "Using non-native error patterns to improve pronunciation verification," 09 2010, pp. 590–593.

[4] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," 2020. [Online]. Available: https://arxiv.org/abs/2008.08647

[5] Y. Song, W. Liang, and R. Liu, "Lattice-based gop in automatic pronunciation evaluation," *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, vol. 3, pp. 598–602, 2010.

[6] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation," in *Tenth annual conference of the international speech communication association*, 2009.

[7] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, no. 2, pp. 83–93, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639399000461

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

[9] B. Lin and L. Wang, "Deep Feature Transfer Learning for Automatic Pronunciation Assessment," in *Proc. Interspeech 2021*, 2021, pp. 4438–4442.

[10] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection." in *Interspeech*, 2021, pp. 4428–4432.

[11] Y. Getman, R. Al-Ghezi, E. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, S. Strömbergsson *et al.*, "wav2vec2-based speech rating system for children with speech sound disorder," in *Proc. Interspeech*, 2022.

[12] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[13] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer Speech & Language*, vol. 50, pp. 62–84, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230816301322

[14] S. Witt, "Use of speech recognition in computer-assisted language learning. phd, dissertation," 12 1999.

[15] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," in *Proc. Speech and Language Technology in Education (SLaTE 2009)*, 2009, pp. 49–52.

[16] 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)*, 2015, pp. 5206–5210.

[17] Voxforge.org, "Free speech... recognition (linux, windows and mac) - voxforge.org," accessed 03/07/2023. [Online]. Available: http://www.voxforge.org/

[18] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an automatic speech recognition dedicated corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 125–129. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/698_Paper.pdf

[19] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus LDC97S63. Web Download. Philadelphia: Linguistic Data Consortium," 1997, accessed 03/07/2023. [Online]. Available: https://catalog.ldc.upenn.edu/LDC97S63

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[21] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 215–219.