# A New DNN-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL)

*Wenping Hu*[1,2], *Yao Qian*[1], *Frank K. Soong*[1]

[1]Microsoft Research Asia, Beijing, P.R.C.
[2]University of Science and Technology of China, Hefei, P.R.C.
{v-wenh,yaoqian,frankkps}@microsoft.com

## Abstract

In this paper, we propose to use Deep Neural Net (DNN), which has been recently shown to reduce speech recognition errors significantly, in Computer-Aided Language Learning (CALL) to evaluate English learners' pronunciations. Multi-layer, stacked Restricted Boltzman Machines (RBMs), are first trained as nonlinear basis functions to represent speech signals succinctly, and the output layer is discriminatively trained to optimize the posterior probabilities of correct, sub-phonemic "senone" states. Three Goodness of Pronunciation (GOP) scores, including: the likelihood-based posterior probability, averaged frame-level posteriors of the DNN output layer "senone" nodes, and log likelihood ratio of correct and competing models, are tested with recordings of both native and non-native speakers, along with manual grading of pronunciation quality. The experimental results show that the GOP estimated by averaged frame-level posteriors of "senones" correlate with human scores the best. Comparing with GOPs estimated with non-DNN, i.e. GMM-HMM, based models, the new approach can improve the correlations relatively by $22.0\%$ or $15.6\%$, at word or sentence levels, respectively. In addition, the frame-level posteriors, which doesn't need a decoding lattice and its corresponding forward-backward computations, is suitable for supporting fast, on-line, multi-channel applications.

**Index Terms**: Conputer-Aided Language Learning, Deep Neural Network, Pronunciation Quality Evaluation

## 1. Introduction

The current globalization of people in regions of different languages has accelerated the demand of foreign language proficiency. While English has widely been considered to be the lingua franca of business [1], it is estimated that 2 billion people are learning English [2]. For an English-as-Second Language (ESL) learner, Computer-Aided Language Learning (CALL) can be very helpful in term of availability and its interactiveness. With the popularity of smart phone, tablet and laptop computers, more language learners prefer to use CALL for language learning. However, as an indispensable part of CALL, Computer-Aided Pronunciation Training (CAPT), which aims at evaluating a learner's pronunciation proficiency and detecting or identifying pronunciation errors or deficiency in a high precision, is still a challenging research.

There are a great deal of research work on automatic pronunciation scoring over the last few decades. An in-depth review of CAPT is given by Witt [3]. Features, for assessing the pronunciation quality, are mostly extracted from the output of an HMM based speech recognizer, e.g. the force alignment result, recognition accuracy and generated lattice. Franco [4] investigated three most important features: HMM-based phone log posterior probability, segment duration and timing scores, in pronunciation evaluation. It was found in the same study that posterior achieves the highest sentence level correlation with human scores. Tobias [5] also compared the performance of some individual pronunciation features in pronunciation scoring, including the likelihood ratio, duration, phoneme accuracy, etc. The result showed that likelihood calculated from a GMM-HMM model has the highest correlation, compared with others. GOP in posterior probability, was first proposed by Witt and Young [6] and then extended by other researchers to different scenarios and applications [7][8]. The posterior probability based scoring has become the predominant technique for the pronunciation quality evaluation in CAPT [3].

Given the current performance of CAPT system, scores generated by the traditional HMM-based speech recognizer need to be further improved. The HMM-based acoustic models can be refined by discriminative training, e.g. MMIE [9][10], MCE [11] and MPE/MWE [12]. However, the improvement on pronunciation evaluation or mispronunciation detection by such discriminative training was limited [13][14]. Additionally, to improve the correlation with human scores or the accuracy of pronunciation error detection, the knowledge of L1 (the native language), e.g. the set of common pronunciation errors made by non-native speakers, is used to contrast L1-L2 phone confusion pairs to build a more custom-made CAPT system. However, it is usually preferable to build an L1-independent system, not only for commercial implementation convenience, but also for native language with many dialectical variabilities, like Chinese.

Recently, a new machine learning algorithm named Deep Neural Network (DNN) has improved speech recognition performance significantly [15]. Application of using DBN-HMMs to mis-pronunciation detection and diagnosis in L2 English has been tried by Qian [16], and a significant improvement on word pronunciation relative error rate was obtained. We want to extend it to a pronunciation quality scoring task in this study. Multi-layer neural networks are trained as nonlinear basis functions to represent speech signals while the top layer of the network is trained discriminatively to sharpen the posterior probabilities of sub-phones ("senones"). In this paper, we investigate the advantages of using posterior probabilities from DNN to pronunciation quality scoring in an L1-independent scenarios, i.e. regardless of the non-native leaner's first language.

## 2. Deep Neural Network Training

A Deep Neural Network (DNN) is a feed-forward, artificial neural network with multiple hidden layers between its input and output. For each hidden unit $j$, a function, typically a logistic

one, is used to map all input from the lower layer, $x_j$, to a scalar state, $y_j$, which is then fed to the upper layer.

$$y_j = logistic(x_j) = \frac{1}{1+e^{-x_j}} \quad (1)$$

where

$$x_j = b_j + \sum_i y_i w_{ij} \quad (2)$$

and $b_j$ is the bias of unit $j$, $i$ is the unit index of lower layer, $w_{ij}$ is the weight on the connection between unit $j$ and unit $i$ in the layer below. For multi-class classification, a "softmax" nonlinear function is used to converts the inputs, $x_j$, into a class probability, $p_j$, where $k$ is an index over all classes.

$$p_j = \frac{exp(x_j)}{\sum_k exp(x_k)} \quad (3)$$

All weights and bias are initialized in pre-training [17], and then discriminatively trained by optimizing a cost function which measures the discrepancy between target label and the predicted output with the Back-Propagation procedure (BP) [18]. When "softmax" output is used, the natural cost function is the cross entropy between the target probability, $d_t$, and actual output of "softmax", $p_t$. When binary, one or zero, target probability, $d_t$, is taken, the cost function $C$ can be simplified as:

$$C = -\sum_{t=1}^{T} d_t log(p_t) = -\sum_{t=1}^{T} logP(s_t|o_t) \quad (4)$$

where $o_t$ is the training sample; $s_t$, the corresponding ground truth label; and $T$ is the size of training samples. It is optimized by a "mini-batch" based stochastic gradient ascent algorithm. More training details, considerations and the integration of DNN with CD-HMMs can be found in [17][19][20].

# 3. Goodness Of Pronunciation (GOP) Estimation

The Goodness of Pronunciation (GOP) scores are estimated in three different ways as shown in the following subsections.

### 3.1. Likelihood-based Posterior Probability

Posterior Probability (PP) estimated from a lattice is generally used as GOP to evaluate the pronunciation quality. It is defined as the conditional probability of a linguistic label, given the acoustic observations. If we use phone as the linguistic representation, the PP of a phone, $p$, is then defined as

$$GOP(p) = P(p|O) = \frac{P(O|p)P(p)}{\sum_{q\in Q} P(O|q)P(q)} \quad (5)$$

where O is the given observation containing all acoustic frames associated and Q is the whole phone set. $P(O|p)$ and $P(p)$ are Acoustic Model (AM) likelihood and Language Model (LM) probability, respectively. PP is calculated in a recognition lattice via the forward-backward algorithm.

We use the Generalized Posterior Probability (GPP) [21], which addresses the uncertainties of modeling and other numerical issues, e.g. adjusting LM scale to compensate for the dynamic range difference between AM likelihoods and LM probabilities, and relaxing unit boundary to avoid underestimating posterior probability in a reduced search space, i.e. a lattice, in PP calculation.

As introduced in Section 2, CD-DNN-HMM outputs the posterior probability $P(s|o)$ of a tied triphone state ("senone"), $s$, given the observation vector, $o$. For the decoding and lattice generation in speech recognition, the posterior of state $s$ is converted into the HMM's emission likelihood $P(o|s)$ by dividing its prior $P(s)$ [19]:

$$P(o|s) = \frac{P(s|o)}{P(s)} \cdot P(o) \quad (6)$$

where $P(o)$ is unknown and approximated by a constant.

### 3.2. Averaged Frame-level Posteriors

Unlike the likelihood proposed in Section 3.1, we can also directly use the posterior probability of "senone" given the argument observations, instead of converting it back to HMM's emission likelihood. Here the phone posterior is approximated by

$$GOP(p) = p(p|t_s, t_e; O) = \frac{1}{t_e - t_s} \sum_{t=t_s}^{t_e} P(s_t|o_t) \quad (7)$$

where $P(s_t|o_t)$ is the output of the last "softmax" layer of our DNN model, $o_t$ is the argument input observations of the frame $t$, $t_s$ or $t_e$ is the start or end frame index of phone $p$, respectively, $s_t$ is the "senone" label of the frame $t$ generated by force alignment. The frame posterior is a probability, where $\sum_{s\in S} P(s|o_t) = 1$, $S$ is the set of all "senones".

### 3.3. Likelihood Ratio between Correct and Competing Models

In statistics, a likelihood ratio test is a statistical hypothesis testing to compare data fitting between two competing models. We use the Log Likelihood Ratio (LLR) between the correct phone model, $p$, and the competing phone model, $q$, as GOP. It is defined as:

$$GOP(p) = \frac{1}{t_e - t_s} \cdot \{\sum_{t=t_s}^{t_e} logP(o_t|p) - \max_{\{q\in Q, q\neq p\}} \sum_{t=t_s}^{t_e} logP(o_t|q)\} \quad (8)$$

This score expresses how likely the observation $o_t$, $t$ from $t_s$ to $t_e$, are under correct phone model $p$ than the competing model $q$. Here the competing phone model $q$ is phone model with highest likelihood, excluding the correct one.

# 4. Experiments and Results

In our experiments, we refine acoustic model, trained first conventionally in the maximum likelihood (ML) sense, by DNN training. Three GOP scores, introduced in the above section, are then tested in a language learning corpus against manual grading of pronunciation quality.

### 4.1. Acoustic Models

Wall Street Journal CSR Corpus is used to train the acoustic models for scoring pronunciation quality in this study. Training set (SI-284) contains $\sim 36K$ utterances ($\sim$ 78 hours) recorded by 284 native American English speakers. Two testing sets (SI-DT-S6 and SI-DT-05), which consists of 202 and 248 utterances, respectively, are used to evaluate the speech recognition

performance of acoustic models. There are no non-native speakers' data used for training or adapting the acoustic models since the ultimate goal of our system is to evaluate the pronunciation quality for any non-native English learners.

Baseline acoustic models are first trained as context-dependent GMM-HMM models (CD-GMM-HMM) in the ML sense. All speech signals are sampled in 16KHz. The acoustic features, extracted by a 25ms hamming window with a 10ms time shift, consist of 13 MFCC and their first and second-order time derivatives. The cepstral mean normalization is performed for each utterance. Three-states, left-to-right HMM triphone models, each state with 16 Gaussian components, diagonal covariance output distribution, are adopted. The phone set is constructed by grouping 61 TIMIT phonemes into 40 phonemes. The total number of "senones" after state-tying is ($\sim 7.3K$). The configurations of training are almost the same as those provided by Vertanen [22].

Acoustic models are then enhanced by DNN training [19]. Our DNN model (CD-DNN-HMM) is an 8 layer network, consisting of 1 input layer, 6 hidden layers, each layer with 2K units, and 1 output layer, with the same number of "senones" output as in CD-GMM-HMM. The input of DNN is an MFCC feature vector, which contains 5 left frames, the current frame and 5 right frames (429 dimensions). Each dimension is normalized to zero mean and unit variance. Our DNN is initialized with the Deep Belief Network (DBN) pre-training procedure [17], which treats each consecutive pair of layers in the MLP as a Restricted Boltzmann Machine(RBM) [17]. Since input feature vector is a continuous variable, the first two layers (i.e. input and the 1st hidden layers) are modeled as a Gaussian-Bernoulli RBM with 26 epochs and other hidden layer pairs are modeled as Bernoulli-Bernoulli RBM with 13 epochs, where the epoch size is set as 24 hours data. All the weights and bias are then discriminatively tuned using about 100 epochs in the BP phase and the learning rates and size of mini-batch are also set dependently for each RBM and each training phase. The state transition parameters are obtained from original CD-GMM-HMM training.

To assess the quality of the two trained acoustic models, we first test them in speech recognition with two WSJ testing sets, SI-DT-05 and SI-DT-S6. The dictionary and language model we used are WSJ 5K, non-verbalized, closed vocabulary set and its corresponding bi-gram language model, respectively. The recognition Word Error Rate (WER) are given in Table 1, which shows the good performance of CD-DNN-HMM, i.e. less than 5% of WER, in an LVCSR task. It achieves up to relative $45.0\%$ and $58.2\%$ WER reduction on SI-DT-05 and SI-DT-S6, respectively, over the conventional CD-GMM-HMM.

Table 1: *Recognition (WER%) of GMM and DNN trained acoustic models*

| Test Sets | SI-DT-05 | SI-DT-S6 |
|---|---|---|
| CD-GMM-HMM | $9.14\%$ | $8.49\%$ |
| CD-DNN-HMM | $4.94\%$ | $3.49\%$ |

### 4.2. Language Learning Corpus

A reading corpus, named "English-900", with manual scoring, is used to evaluate the performance of pronunciation quality of sentences. The "English-900" sentence set contains 900 sentences used in conversational English. Most sentences are colloquial and conversational. The corpus is recorded by 8 non-native English learners with different English proficiency levels, based upon their TOEFL[1] or IELTS[2] scores, and 2 native American English speakers. Each speaker read 300 sentences, randomly selected from the "English-900" sentence set.

The "ground truth" assessment of pronunciation were obtained by three human expert linguistic raters, one native rater who has English teaching experience in China and two non-native English speaking raters, both living in English speaking countries for many years, to grade the pronunciation quality at both word and sentence levels. All raters have a good sense of assessing spoken English quality and they are educated in English phonetics. Standard American English pronunciation is used as a criterion for scoring. All speakers' proficiencies are bracketed into five grades, ranging from 1 (best) to 5 (worst). The criteria for scoring is given in Table 2. To provide a more detail and specific feedback of pronunciation mistakes, sub-word or phone level assessment are also useful, but the reliability of pronunciation quality scoring by human, even for linguistics experts, at the phone level is not high [23]. Therefore, the manual scoring is only evaluated at the levels of word and sentence.

After scoring, we calculated the averaged correlation for measuring inter-rater agreement. Each correlation is calculated between one rater's scores with the average of two other raters. Table 3 gives the averaged correlation of three raters at word, sentence and speaker levels. The score at speaker level is the averaged score at sentence level per speaker.

Table 2: *Scoring standard*

| Grade | Standard |
|---|---|
| 1 | Native-like pronunciation and intonation, natural and intelligible |
| 2 | Good and fluent English pronunciation of non-native speakers |
| 3 | Clear pronunciation but still with some common pronunciation errors and unnatural intonations |
| 4 | Not clear pronunciation or frequent pauses of words and consistent pronunciation mistakes |
| 5 | Unintelligible pronunciations and awkward intonation |

Table 3: *Inter-rater agreement in scoring*

| | Word | Sentence | Speaker |
|---|---|---|---|
| Correlation | 0.62 | 0.61 | 0.86 |

Since the ultimate goal of our CAPT system is to identify the pronunciation error within a word, i.e. at syllable or phone level, the GOP score by our system is estimated at the phone level and then accumulated across word or sentence. In addition, to avoid Out-Of-Vocabulary (OOV) problem in evaluating pronunciation quality, we generally use phone or syllable-based ASR. We carry out a phone recognition with refined acoustic models trained with DNN. The acoustic feature extraction is the same as that in DNN training. A bi-gram language model, which is trained from a corpus (TIMIT) with manually labeled, balance designed phone level script, is adopted.

[1] Test Of English as a Foreign Language

[2] International English Language Testing System

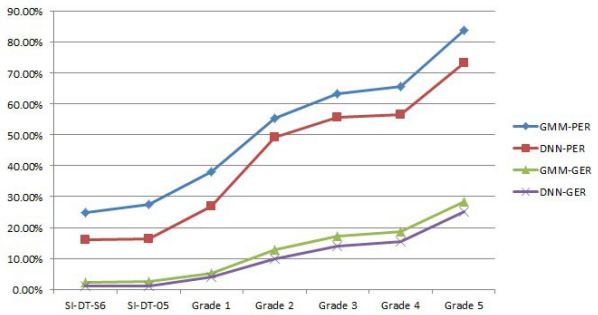


Figure 1: *PER and GER for different sets and models.*

Figure 1 shows Phone Error Rate(PER) and Graph Error Rate (GER) for WSJ testing sets and the corpus of "English-900", which is bracketed into 5 sets by different manual scores at sentence level. GER is computed by determining that sentence through the phone graph that best matches the spoken sentence in terms of phone errors. It is clearly that DNN-HMM model has a consistently better discriminative ability than GMM-HMM on all sets. However, PER and GER for "English-900", even the data with grade 1, are higher than those for WSJ testing sets. This is due to the fact that sentences of "English-900" are uttered in conversation style while those of WSJ are uttered in news broadcasting style. In addition, there are about $13.3\%$ of OOV in "English-900" if using the word dictionary provided by WSJ. Figure 1 also shows the PER and GER are continuously increased with the manually grades ranging from 1 to 5, which matches our labeling criteria listed in Table 2.

### 4.3. Evaluation Results and Analysis

The GOP computed in our new CAPT are correlated with human ratings. The machine score at the word level is the average phone score over its phone constituents and the machine score at the sentence level is the average word score over its word constituents. Four systems are used in our CAPT experiment. They are listed as the following:

- System 1 (baseline): GOP is computed as the conventional likelihood-based posterior probability, as described in section 3.1. The phone lattice is generated by the acoustic model trained by CD-GMM-HMM and the bi-gram LM of phone transitions.
- System 2: GOP is computed as the conventional likelihood-based posterior probability, as described in section 3.1. The phone lattice is generated by the acoustic model trained by CD-DNN-HMM and the bi-gram LM of phone transitions.
- System 3: GOP is computed as averaged frame-level posteriors of the DNN output layer "senone" nodes, as described in section 3.2.
- System 4: GOP is computed as the log likelihood ratio between the correct and competing models, as described in section 3.3.

The correlation coefficients between the machine human ratings for all systems are shown in Table 4.

Both system 1 (baseline) and system 2 use the likelihood-based posterior probability as GOP score. Likelihood of phone and phone lattice, used for approximating all decoding space, in baseline are generated by acoustic model CD-GMM-HMM

Table 4: *Correlation between each automatic system and human expert scoring*

| | System 1 | System 2 | System 3 | System 4 |
|---|---|---|---|---|
| Word | 0.41 | 0.44 | **0.50** | **0.50** |
| Sentence | 0.45 | 0.46 | **0.52** | 0.49 |
| Speaker | 0.93 | 0.94 | **0.95** | **0.95** |

while those of system 2 are generated by acoustic model CD-DNN-HMM. In our experiment, we use GPP as GOP and the phone LM scales for GPP calculation, with CD-GMM-HMM and CD-DNN-HMM, are optimized through a grid search, respectively. Although CD-DNN-HMM can significantly reduces the PER for all data sets, e.g. the maximum reduction is on grade 1 by $11.3\%$ and the minimum reduction is on grade 2 by $6.4\%$, the improvement on correlations between GOP score and manual score are not as distinctive as the recognition performance improvement.

Compared with the traditional lattice based GOP algorithm, system 3 and system 4 perform consistently better at different phonetic units. The GOP in system 3, which is estimated by averaged frame-level posteriors of "senones", correlates with human scores the best. It can improve the correlations relatively by $22.0\%$ or $15.6\%$, at word or sentence levels, respectively. In addition, the frame-level posterior GOP, which doesn't need a decoding lattice and its corresponding forward-backward computations, is very suitable for supporting fast, on-line, multi-channel applications. The GOP in system 4, which is estimated by log likelihood ratio of the correct and competing models, also achieves the highest correlations to manual scores at both word and speaker levels. System 3 and System 4 can achieve almost the same performance in term of the correlations between manual grading scores and machine GOP scores at both word and speaker levels. However, the computational complexity of System 3 is much lower than that of System 4. The GOP in System 3 can be implemented by force-alignment with the corresponding given phone sequence, while the GOP in system 4 need to consider all possible phone candidates.

## 5. Conclusions and Future Work

We propose a new DNN-based, high quality pronunciation assessment for computer-aided language learning in this paper. The acoustic models, which are trained by 284 native American English speakers' recordings in the ML sense, are refined by DNN training. The refined acoustic models are then used to generate the evaluation GOP scores and tested in a language learning corpus with manual grading of pronunciation quality. The GOP scores, generated from the refined acoustic models, can correlate with human scores better than those scores generated from the conventional approaches. The new proposed approach, using the averaged frame-level posteriors of the DNN output as GOP score, achieves the best performance. It is also very suitable for supporting fast, on-line, multi-channel application due to its low computational complexity.

Besides the pronunciation quality, prosodic features, like intonation, and stress, and fluency, are also important for evaluating a learner's English proficiency. In the future, we will embed those features with the DNN based GOP scores into our assessment to make it even more comprehensive.

# 6. References


[1] B. Seidlhofer, "Common ground and different realities: world englishes and english as a lingua franca," *World Englishes*, vol. 28, no. 2, pp. 236–245, 2009.

[2] D. Graddol, *Why global English may mean the end of English as a Foreign Language*. British Council, 2006.

[3] S. M. Witt, "Automatic error detection in pronunciation training:where we are and where we need to go," 2012.

[4] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, vol. 30, no. 2-3, pp. 121–130, 2000.

[5] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[6] 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.

[7] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R.-H. Wang, "Automatic mispronunciation detection for mandarin," pp. 5077–5080, 2008.

[8] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation," pp. 608–611, 2009.

[9] V. Valtchev, J. J. Odell, P. Woodland, and S. Young, "MMIE training of large vocabulary recognition systems," 1997.

[10] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 25–47, Jan. 2002.

[11] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[12] D. Povey, P. C. Woodland, and M. J. F. Gales, "Discriminative map for acoustic model adaptation," *2003 IEEE International Conference on Acoustics Speech and Signal Processing 2003 Proceedings ICASSP 03*, vol. 1, pp. I–312–I–315, 2003.

[13] S. G. Ke Yan, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," pp. 17–23, 2011.

[14] X. Qian, F. K. Soong, and H. M. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," pp. 757–760, 2010.

[15] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, 2012.

[16] X. Qian, H. M. Meng, and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," 2012.

[17] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. Oct, pp. 533–536, 1986.

[19] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," pp. 24–29, 2011.

[20] 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 & Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[21] F. K. Soong, W. kit Lo, and S. Nakamura, "Generalized word posterior probability (GWPP) for measuring reliability of recognized words," 2004.

[22] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," 2006.

[23] P. Mueller, F. de Wet, C. van der Walt, and T. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers."