Interspeech 2018
2-6 September 2018, Hyderabad

Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment

Yujia Xiao[1,2*], *Frank K. Soong*[2], *Wenping Hu*[2]

[1] South China University of Technology, Guangzhou, China
[2] Microsoft Research Asia, Beijing, China

xiao.yujia@mail.scut.edu.cn, frankkps@microsoft.com, wenh@microsoft.com

Abstract

This work proposes to incorporate paired phone-posteriors as input features into a neural net (NN) model for assessing ESL learner's pronunciation quality. In this work, posteriors of forty phones, instead of several thousand sub-phonemic senones, are used to circumvent the sparsity issues in NN training. Phone posteriors are assembled with their corresponding senone posteriors estimated via a speaker-independent, DNN-based acoustic model, trained with standard American English speech data (i.e., Wall Street Journal database). Phone posteriors of both reference (standard American English speaker) and test speaker are paired together as augmented input feature vectors to train an NN based, 2-class, i.e., native vs nonnative speaker, classifier. The Goodness of Pronunciation (GOP), a proven effective measure, is used as the baseline for comparison. The binary NN classifier trained with such features achieves a high classification accuracy of 89.6% on native and non-native speakers' data. The classifier also shows a better equal error rate (EER) than the GOP-based baseline classifier in either phone or word level pronunciation, i.e., at phone level from 18.3% to 6.2%, and at word level from 12.98% to 2.54%.

Index Terms: Computer-Aided Language Learning, Deep Nerual Network, Pronunciation Quality Evaluation

1. Introduction

Learning English, the de facto international language, is critical in cross-cultural or business communications in today's world. To overcome the short supply of qualified human teachers, Computer-Aided Language Learning (CALL) is useful with effective computational assessment for improving the efficiency in both learning and teaching of a second language like English. An important module of CALL, Computer-Aided Pronunciation Training (CAPT), is to evaluate the pronunciation quality of a learner, detecting mispronunciation or deficiency, and providing timely and focused feedback to the learner.

Most research work on CAPT are based upon speech recognition algorithms. Speech recognition systems can perform forced alignment of a learner's read-after-prompted-sentence input with the corresponding prompted text and provide appropriate scores at different unit levels, e.g., phone, syllable, word, phrase, or sentence. The recognition accuracy can be considered as an important scoring criterion for pronunciation evaluation. For example, Kim *et al.* [1] proposes three probabilistic models to produce pronunciation scores based on the phonetic time alignments, *i.e.*, HMM-based log-likelihood scores, HMM-based log posterior scores and segment duration scores. Kawai *et al.* [2] also showed that using log-likelihood scores in forced alignment mode is helpful for teaching Japanese pronunciation. Franco *et al.* [3] proposed to use a log-likelihood ratio (LLR) score based on two different acoustic models (i.e., trained with authentic, native pronunciations and accented, non-native pronunciations, respectively), and found that the LLR based method achieved better performance, compared with the posterior based methods. However, this LLR based approach needs data from target non-native speakers, which may not be available for the corresponding model training. The Goodness of Pronunciation (GOP) based method, which was proposed by Witt and Young [4] is based on the posterior probability. Many follow-up works extend the GOP method by considering scaled log-posterior probability [5], generalized segment posterior probability [6] and log-likelihood ratios [7].

Furthermore, some discriminative training algorithms were proposed to improve the performance of the GMM-HMM based speech recognition system [8, 9, 10], which has been applied to CAPT system but with limited improvement [11, 12]. More recently, in light of the advancement of deep neural network (DNN) technology on speech processing [13], significant performance improvement was achieved by HMM-DNN based systems over the GMM-HMM based approach in CAPT [14, 15, 16].

As the prominent and typical pronunciation evaluation method based upon posterior probability, GOP scores are usually calculated by the posteriors corresponding to its aligned phonetic (sub-phonemic) units. For speech uttered by non-native, beginning learners, such posteriors estimated from the acoustic model (usually trained with native speaker' data) has a flatter, *i.e.*, not salient and distinctive, distribution at the target unit, compared with those of a native speaker. The divergence between the posterior distributions of native speaker and non-native speaker is investigated in this study (Section 3). The correlation between scores of system and human experts is used as a check to evaluate the performance of CAPT systems. Based on the metric, a human labelling process on all of the experiment data is needed. However, this process is time-consuming and subjective. In this paper, we proposed a DNN-based pronunciation assessment system by using the "native" or "non-native" as the speaker's label. In addition, we constructed our input features as paired phone-posteriors, which can be found in Section 4. The effectiveness of the proposed method is tested both on native and nonnative corpus and compared with DNN-GOP algorithm in Section 5.

2. Goodness of Pronunciation

In this section, we give a brief review of our GOP-based baseline system. In a traditional GMM-HMM based system [4], the phone-level GOP score is defined as follows:

$$GOP_1(p) = \frac{|log(\frac{p(o^{(p)}|p)\times p(p)}{\sum_{q\in Q} p(o^p|q)\times p(q)})|}{NF(p)} \tag{1}$$

[*] Work performed as an intern in Microsoft Research Asia.

10.21437/Interspeech.2018-1270

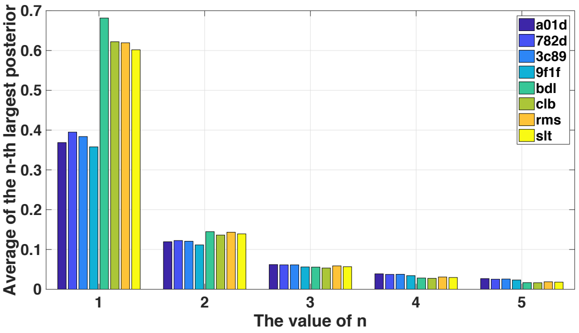

Figure 1: *Average of the n-th largest posterior of different speakers' SPD ($n = 1 \sim 5$)*

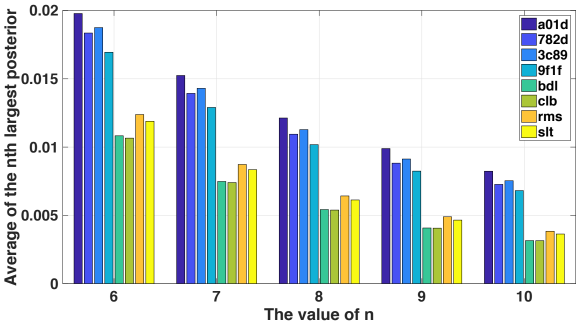

Figure 2: *Average of the n-th largest posterior of different speakers' SPD ($n = 6 \sim 10$)*

where p and Q represents, the target phone and the phone set, respectively. The acoustic observations of the phone p is denoted as $o^{(p)}$. Additionally, $p(o^{(p)}|p)$ is computed by HMMs and $NF(p)$ indicates the duration (frames) of $o^{(p)}$.

In light of significant progress achieved by deep learning techniques on speech recognition, Hu *et al.* [16] extended the GOP evaluation from GMM-HMM to DNN-HMM systems with different score estimated ways. Based on the evaluation results, the best experimental result for assessing GOP is achieved by considering the averaged frame-level posteriors. Formally, the refined GOP score is defined as follows:

$$GOP_2(p) = p(p|t_s, t_e; o^{(p)}) = \frac{1}{t_e - t_s} \sum_{t=t_s}^{t_e} P(s_t|o_t^{(p)}) \quad (2)$$

where t_s and t_e represents, the starting and ending frames of phone p, respectively. s_t denotes, the senone label at frame t obtained after alignment. $P(s_t|o_t^{(p)})$ is the softmax output, or the estimated posterior, of the DNN-based acoustic model.

3. Feature Analysis

In this section, we conduct statistical analysis to show the motivations of feature extraction in our model. As shown in Eq (2), DNN-GOP scores are calculated by the aligned senone's posterior for each frame t. In this work, we assumed that the divergence between the pronunciation of native and nonnative speakers is represented by posteriors of all senones. The DNN-based native acoustic model trained has 6 hidden layers, 2k nodes for each hidden layer, 2,754 output "senone" states, trained with the Wall Street Journal (WSJ) CSR Corpus [16].

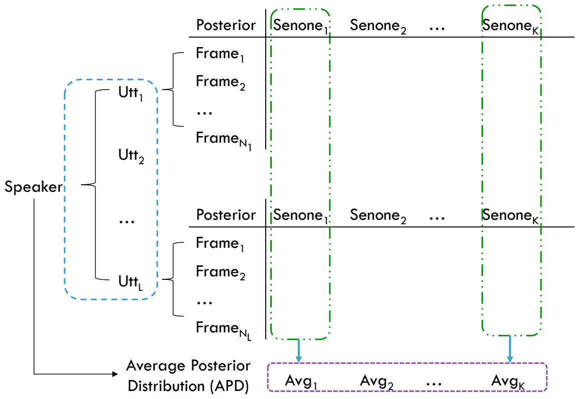

Figure 3: *The construction way of each speaker's APD*

3.1. Corpus

Two different datasets are used in our experiments:

- CMU-Arctic database [17] where 4 US native English speakers, 2 female (slt, clb) and 2 male (bdl, rms). Approximately 1,200 phonetically balanced English utterances were recorded in a quier studio.
- mTutor-User database [1], users of Microsoft English learning project "mTutor", 3c89 (female), 2,332 utterances; a01d (female), 859 utterances; 782d (male), 1,288 utterances, 9f1f (male), 1,597 utterances. They are L2 English learners with Mandarin as L1 and they read after the standard recordings from a female native speaker.

3.2. Divergence between Senone Posterior distributions

We obtained the senone posterior distributions (SPDs) from the DNN-based acoustic model of the aforementioned two datasets. In each SPD, we ranked the component posteriors in descending order and selected the n-th ($n = 1 \sim 10$) largest posteriors of each SPD. Given a specific value of n, we then take the average of the selected posteriors across different SPDs for each speaker. Fig 1 and Fig 2 presents the statistical results of different speakers, *i.e.*, native speakers (bdl, clb, rms, slt) and nonnative speakers (a01d, 782d, 3c89, 9f1f), for $n = 1 \sim 5$ and $n = 6 \sim 10$, respectively.

From these figs, speakers in the same group, native or nonnative, have similar distributions. For example, in Figure 1, where n equals to 1 (*i.e.*, the average of the maximum posterior of each speaker's SPDs), the values of native speakers (the 4 bars on the right side) are much larger than that of nonnative speakers (the 4 bars on the left side). The pronunciations of nonnative speakers are not as authentic as the native speakers which match better with the acoustic model trained with the WSJ CSR corpus recorded by native speakers. The average posterior will not only be assembled in one single senone, but also in other potential candidate senones. In Figure 1, we can observe that the difference between native speakers and nonnative speakers decreases gradually as n increases. In Figure 2, where n is from 6 to 10, average posteriors of nonnative speakers are larger than those of native speakers. Inspired by the above observations, we propose to consider the posterior information with $n > 1$ which can also be helpful to differentiate native and nonnative speakers.

[1] http://www.engkoo.com/

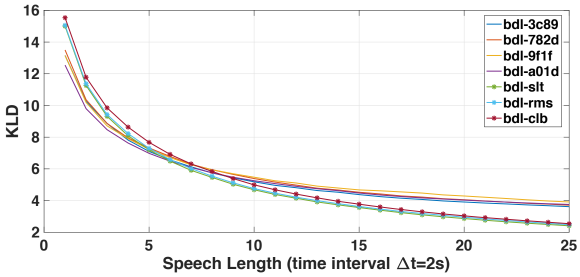

Figure 4: *KLD between different speakers' APD in different speech length*

3.3. Averaged posterior distribution based KL-divergence

We used a symmetric Kullback-Leibler divergence (KLD) [18, 19] to measure the difference between averaged posterior distribution (APD) of different speakers. The block diagram for extracting each speaker's APD is showed in Figure 3. For each speaker, we obtained the SPDs by utilizing our developed acoustic model. Notice that the silence frames in the head and tail of each utterance have been removed, to avoid interfering by the irrelevant background noise.

In specific, we calculated the KLD between different speakers' APDs based on Eq (3), where A_i or A_j are the APD of speaker i and j, respectively. k is the senone index.

$$D_{KL}(A_i||A_j) = \sum_k \{[A_i(k) - A_j(k)] \times log \frac{A_i(k)}{A_j(k)}\} \quad (3)$$

Figure 4 shows how the KLD between the variation of speaker bdl's APD and other speakers' APD with the increasing of speech length used to construct the APD (using other native speakers' yielded similar result). When the speech utterance is short, the APD is affected by the content. In that case, posterior distributions of native speakers output from the acoustic model are sharper than that of nonnative speakers. Therefore, the KLDs between native-native speaker pairs are larger than that between native-nonnative speaker pairs. With the increasing of test token length, the influence from content reduces gradually and the pronunciation information becomes the predominant factor. Therefore, all of the KLDs from different groups decrease. When it comes to a length that is large enough, the KLDs between native-native speaker pairs will be smaller than that between native-nonnative speaker pairs. In conclusion, a speaker's pronunciation quality can be assessed by the APD of test token. When the data used to calculate the APD is not long enough, phonetic content can affect the classification results.

4. Pronunciation Assessment

In this work, we propose a DNN-based binary classifier to perform pronunciation assessment. The framework is shown in Figure 5. For each test utterance in a read-after-a-prompt trial, a reference utterance spoken by a native speaker is also provided. The acoustic features extracted from both test and reference utterances are converted to senone or phone APDs. The constructed APDs are combined into a final augmented vector, which serve as the input feature of our assessment model. The output probability of the trained model will be used as the senone-level pronunciation assessment score. Details are explained in the remainder of this section.

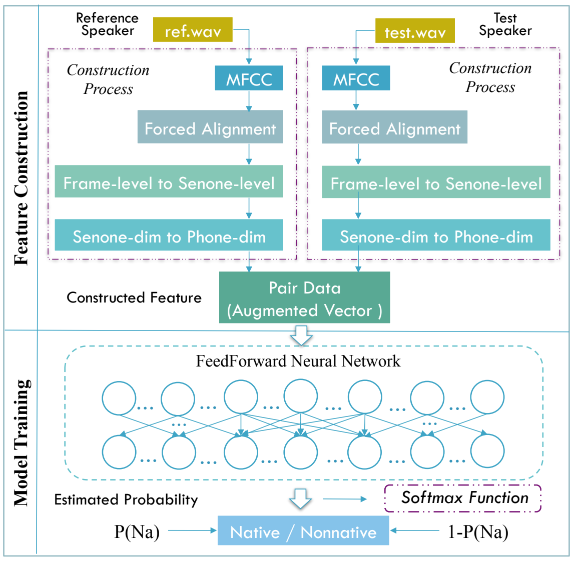

Figure 5: *The framework of a Senone-level Posterior based Binary Classifier*

4.1. Senone-level Posterior Distribution

Each constructed feature is corresponding to a senone-level acoustic segment. For each utterance, we first obtained the frame-level posterior distribution from the acoustic model. Given the transcription, each frame is aligned with the text down to the senone level. Therefore, we can calculate the senone-level APDs by averaging the frame-level posterior distributions. Therefore, the output probability of our model for each input feature is a senone-level pronunciation assessment score, which can be used to generate a high-level segment's score, such as phone, word and utterance.

4.2. Phone-dimension Posterior Distribution

Up to now, the posterior distributions we process are vectors of 2754 senones. With a large number of senones, most cells in the posterior vector are with small values, which leads to the sparsity of the input feature. Also, training a DNN model with such high-dimensional input vectors can be time consuming. Since each senone can be matched to the corresponding phoneme, we shrink the dimension of the posterior distribution from senone to phone as 40 (by summing up the senone posteriors which corresponds to the same phone). Phonetic posterior features have also been considered in other task, such as voice conversion [20, 21]. In our work, the utilization of phone-dimension posterior distribution makes the training process more efficiently and the feature more compactly.

4.3. Paired Posterior Features

Suggested by the experimental results in Section 3.3, we notice that when test data is phonetically rich enough and its duration is long enough, the pronunciation quality of a speaker can be well assessed, in terms of native vs nonnative proficiency. However, a senone-level APD is in general not enough for assessing the pronunciation reliably. To improve the assessment, we propose to use paired phone-posteriors from both test speaker and native (reference) speaker to take advantage of the contrasting nature of the paired features. A reference utterance spoken by a native speaker is provided for pairing with each test utterance.

Table 1: *Classification Accuracy of our pronunciation assessment model.*

Model Variants	Dimension of Input Features	Classification Accuracy (%)
A	5508	77.5
B	40	74.9
C	80	79.9
C-Log	80	89.6

In our corpus, the native speaker in CMU Arctic database uses another native speaker in the same dataset as the reference one. For the nonnative speaker in mTutor user, the reference speaker is the native speaker they read after.

4.4. Log Transformation of Posteriors

The input feature vector of our model is a senone-level, phone-dim, paired posterior distribution. The value of each feature is between 0 and 1 and the sum of the whole vector components is 1. For the dynamic range of the posteriors in a linear scale can be pretty large and makes the training process numerically difficult, we use logarithm to compress it into a smaller dynamic range.

4.5. DNN-based Binary Classification Model

We train a feedforward network to perform a native vs non-native speaker classification. Our model is trained as a 4 layer network, consisting of 1 input layer (an augmented feature vector with 80 dimensions), 2 hidden layers (each layer with 16 units) and 1 output layer (2 classes, native and non-native). In this network, sigmoid function is used as the activation function; a softmax function is used to convert the output to an actual descent probability of each class. We use stochastic gradient (SGD) [22] to minimize the loss function (cross-entropy). The sample-level learning rate is 0.0001 and the epoch number is 20.

5. Evaluation

5.1. Classification Accuracy

The two speech databases were randomly divided into 3 subgroups with the same size for cross validation, where 2 groups were considered for training and the remaining one for testing. Three different variants of our model, A, B and C, described in Section 4, are trained to examine their effectiveness in differentiating native US English speaker from non-native, ESL learners. The experimental results of classification accuracy are shown in Table 1.

The difference between model A and model C is the dimension of model's input features. Particularly, model A uses the features with paired posteriors of *senones* of a dimension 5508, while model C uses the features with paired posteriors of *phones* of a dimension 80. From the result in Table 1, mapping posteriors from senones to phonemes improves the classification performance, i.e., classification accuracy improved from 77.5% to 79.9%.

Furthermore, we investigate the effectiveness of *paired data* construction by comparing the performance of model B and model C. Their used features are all senone-level and phone-dimension posterior distributions. The difference is that model C uses paired posteriors while model B does not. The experiment result shows that using paired data improves the classification accuracy by 5% , i.e., from 74.9% to 79.9%.

By taking logarithm of the input features in model C, we can further improve the performance significantly. The classification accuracy is boosted by 9.7%, i.e., from 79.9% to 89.6%, which suggests the effectiveness of the logarithm operation by considering a smaller dynamic range of the input features.

Table 2: *Comparing our model with DNN-GOP by EER(%).*

Segmental Level	DNN-GOP	Model C-Log
Phone	18.30	6.24
Word	12.98	2.54
Utterance	0.33	0.00

5.2. Comparison with the baseline DNN-GOP

5.2.1. Performance Measure: Equal Error Rate

We use equal error rate (EER) as the performance measure for comparing the proposed model with the DNN-GOP based baseline system. Note that false acceptance (FA) error rate and false rejection (FR) error rate are equal in EER. In this study, an FA error is made when we misclassify a test token from a nonnative speaker as a native speaker, and an FR error is made if we misclassify a test token from a native speaker as a nonnative speaker.

5.2.2. Performance Measure: Equal Error Rate

For each utterance, two scores are obtained: a) forced alignment for calculating the phone-level DNN-GOP score in Eq (2); b) the model (Model C with log) proposed in Section 4 is used to output the probability of a native speaker as the pronunciation score (in senone level). Given the scores in phone-level (or senone-level), a higher segmental level score can be obtained by averaging the corresponding sub-level segments' scores. In Table 2, the EERs of our model are compared with DNN-GOP for different units. Our proposed model shows a better performance than DNN-GOP, i.e., the EERs at phone, word and utterance, are improved from 18.30%, 12.98%, 0.33% to 6.24%, 2.54%, 0.0%, respectively.

6. Conclusion

We propose to incorporate paired (augmented) phone posteriors of both the reference (standard American English) speaker and ESL learner as input features into a DNN-based binary classifier in read-after-me, oral practice for assessing a test speaker's pronunciation quality. The proposed approach achieves a classification accuracy of 89.6%. Specifically, the experimental results show that the new model trained with paired phone-posteriors outperforms a DNN-GOP baseline by 12.06% reduction of EER, from 18.3% down to 6.24%, at phone level; and by 10.44% reduction of EER, from 12.98% down to 2.54%, at word level.

7. References

[1] Y. Kim, H. Franco, and L. Neumeyer, “Automatic pronunciation scoring of specific phone segments for language instruction,” in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997.

[2] G. Kawai and K. Hirose, “A call system using speech recognition to train the pronunciation of japanese long vowels, the mora nasal and mora obstruents,” in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997.

[3] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,” in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999.

[4] 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.

[5] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, “Automatic mispronunciation detection for mandarin,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2008, pp. 5077–5080.

[6] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W.-p. Ye, “Generalized segment posterior probability for automatic mandarin pronunciation evaluation,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. IEEE, 2007, pp. IV–201.

[7] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, “A new method for mispronunciation detection using support vector machine based on pronunciation space models,” *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[8] L. Bahl, P. Brown, P. De Souza, and R. Mercer, “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11. IEEE, 1986, pp. 49–52.

[9] B.-H. Juang, W. Hou, and C.-H. Lee, “Minimum classification error rate methods for speech recognition,” *Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[10] D. Povey and P. C. Woodland, “Minimum phone error and i-smoothing for improved discriminative training,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–105.

[11] K. Yan and S. Gong, “Pronunciation proficiency evaluation based on discriminatively refined acoustic models,” *International Journal of Information Technology and Computer Science*, vol. 3, no. 2, pp. 17–23, 2011.

[12] X. Qian, F. K. Soong, and H. Meng, “Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (capt),” in *Annual Conference of the International Speech Communication Association (ISCA)*, 2010.

[13] 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,” *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[14] W. Hu, Y. Qian, and F. K. Soong, “An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech.” in *Symposium on Languages, Applications and Technologies (SLATE)*, 2015.

[15] X. Qian, H. Meng, and F. K. Soong, “The use of dbn-hmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training,” in *Annual Conference of the International Speech Communication Association (ISCA)*, 2012.

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

[17] J. Kominek and A. W. Black, “The cmu arctic speech databases,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[18] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[19] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.

[20] F.-L. Xie, F. K. Soong, and H. Li, “A kl divergence and dnn-based approach to voice conversion without parallel training sentences.” in *International Speech Communication Association (Interspeech)*, 2016, pp. 287–291.

[21] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[22] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.