A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection

Huang, Hao; Xu, Haihua; Hu, Ying; Zhou, Gang

2017

Huang, H., Xu, H., Hu, Y., & Zhou, G. (2017). A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection. *The Journal of the Acoustical Society of America*, 142(5), 3165-3177.

<https://hdl.handle.net/10356/86625>

<https://doi.org/10.1121/1.5011159>

© 2017 Acoustical Society of America. This paper was published in *Journal of the Acoustical Society of America* and is made available as an electronic reprint (preprint) with permission of Acoustical Society of America. The published version is available at: [<http://dx.doi.org/10.1121/1.5011159>]. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper is prohibited and is subject to penalties under law.

Downloaded on 26 Apr 2023 20:57:20 SGT

A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection

Hao Huang,^{1,a)} Haihua Xu,² Ying Hu,¹ and Gang Zhou¹

¹*School of Information Science and Engineering, Xinjiang University, Shangli Road, Urumqi 830046, China*

²*Temasek Laboratories, Nanyang Technological University, 50 Nanyang Drive, Singapore 637553, Singapore*

(Received 16 February 2017; revised 2 October 2017; accepted 27 October 2017; published online 21 November 2017)

Goodness of pronunciation (GOP) is the most widely used method for automatic mispronunciation detection. In this paper, a transfer learning approach to GOP based mispronunciation detection when applying maximum F1-score criterion (MFC) training to deep neural network (DNN)-hidden Markov model based acoustic models is proposed. Rather than train the whole network using MFC, a DNN is used, whose hidden layers are borrowed from native speech recognition with only the softmax layer trained according to the MFC objective function. As a result, significant mispronunciation detection improvement is obtained. In light of this, the two-stage transfer learning based GOP is investigated in depth. The first stage exploits the hidden layer(s) to extract phonetic-discriminating features. The second stage uses a trainable softmax layer to learn the human standard for judgment. The validation is carried out by experimenting with different mispronunciation detection architectures using acoustic models trained by different criteria. It is found that it is preferable to use frame-level cross-entropy to train the hidden layer parameters. Classifier based mispronunciation detection is further experimented with using features computed by transfer learning based GOP and it is shown that it also helps to achieve better results.

© 2017 Acoustical Society of America. <https://doi.org/10.1121/1.5011159>

[MAH]

Pages: 3165–3177

I. INTRODUCTION

With the accelerating process of globalization and development of computing technology, computer assisted language learning (CALL) that makes use of speech and language technologies to facilitate second-language learning has gained a growing interest over the last two decades. Computer assisted pronunciation training, which provides opportunities for learners practicing their pronunciation and gives automatic feedback, is one of the most popularly deployed applications. Generally, there are two kinds of pronunciation feedbacks. One is to give learners pronunciation scores.^{1–6} The other is to help a language learner by pinpointing erroneous pronunciations, which is often referred to as automatic mispronunciation detection. Research on automatic mispronunciation detection has been comprehensively carried out at the phone-level,^{7–14} word-level,^{15,16} and phrase-level.¹⁷ In addition, some studies extend the detection task to mispronunciation diagnosis, which identifies the type of incorrect phone(s) produced in place of the canonical phone to provide more useful feedback for the learner.^{18,19}

For automatic mispronunciation detection, there has been a great deal of research work and lots of methods have been proposed. They range from confidence based methods,^{6–9} classifier based methods,^{10–12} and continuous speech recognition based methods.^{13,14,18} Among them, the goodness of pronunciation (GOP)⁸ based method, which can be regarded as a confidence measure based approach, is the

most widely used. The classifier based mispronunciation detection approach uses manually designed mispronunciation detection features as front-end and statistical classifiers such as support vector machine (SVM) or neural network logistic regression (NNLR) classifier as the back-end to make the final decision.^{10–12} The continuous speech recognition based method uses a phone recognizer to decode the input waveforms to find out the possible errors.^{13,14,18,19}

In early research on GOP based mispronunciation detection, Gaussian mixture model–hidden Markov models (GMM-HMMs) were used as the acoustic models to compute the GOP scores. Traditionally, the GMM-HMMs are trained by maximum likelihood (ML) criterion. The ML criterion is an objective function in the automatic speech recognition (ASR) task that seeks to maximize the joint likelihood of data as well as its label. In ASR, the performance of a system is often evaluated in terms of word error rate (WER). Discriminative training (DT) of the acoustic models were used to reduce WER. The most commonly used discriminative training methods include minimum classification error,²⁰ maximum mutual information (MMI),^{21,22} and minimum phone error (MPE).^{23,24}

By contrast to the task of speech recognition, mispronunciation detection task concerns about evaluation metrics different from WER. The commonly used metrics include false rejection (FR) (correct pronunciations detected as incorrect), false acceptances (FA) (errors detected as correct), true acceptance (TA) (correct pronunciations detected as correct), and true rejection (TR) (errors detected as incorrect). Some works use precision and recall (or precision and

^{a)}Electronic mail: hwanghao@gmail.com

recall curve) to evaluate system performance. These metrics can be effective, however, there has often to be an empirical trade off among the multiple objectives when they are used as the training targets. To solve the problem, we introduced maximum F1-score criterion (MFC) that explicitly maximizes F1-score and presented GMM-HMM training algorithm according to MFC in Refs. 25 and 26. It was shown that the proposed method had significantly improved detection results over the ML trained GMM-HMM based acoustic models.

Recently, progress in deep learning for ASR (Ref. 27) triggered a lot of work on applying a deep neural network (DNN) to CALL system.^{4,5,11,12,18,28-30} DNN acoustic models has shown better detection results than ML trained GMM-HMMs in GOP based mispronunciation detection.^{4,12} Normally, the DNN parameters are initialized using generative pretraining and then fine-tuned under the cross-entropy (CE) criterion. However, the CE minimization objective is also used for reducing WER and is not directly linked to the optimization of mispronunciation detection performance. As claimed in Ref. 26, the problem on how the golden standard could be incorporated in supervised training of the DNN still remains.

Most recently, both our work and the results by Hsu³¹ showed that applying the MFC objective function to supervised training of DNN based acoustic models yields better mispronunciation detection performance. In this paper, we propose transfer learning based GOP, in which the hidden layers are borrowed for DNN trained by CE objective function on native speech data and only the top softmax layer of the network is trained by MFC. Results show detection performance can be significantly improved and MFC training of the top softmax layer is more robust to over-training than MFC training of the whole network. In light of this, we bare the thought that GOP based mispronunciation detection can be fundamentally separated into a two-stage based transfer learning framework: The first stage uses DNN (or GMM) based hidden layers borrowed from acoustic models trained for native speech recognition to discriminate the standard (native) pronunciations in the acoustic space. The second stage uses a trainable softmax layer to approximate the golden standard of a human expert.

For the first stage, a question is, what is an appropriate criterion for training the hidden-layer acoustic models? Inconsistent with our intuition, results presented in our previous work²⁶ showed that for GMM-HMM based acoustic models, MPE and MMI training aiming at reducing WER for speech recognition did not clearly lead to better detection performance. For DNN based acoustic models, it was demonstrated in Ref. 32 that DNN can be trained with sequence classification criteria named the state-level minimum Bayesian risk (sMBR) which uses exactly the same lattice-based methods that have been developed for GMM-HMMs (MPE) to improve speech recognition performance. However, it was shown in Ref. 31 that while sMBR can considerably improve the ASR performance, it did not provide additional gain in mispronunciation detection that employs the CE estimated DNN acoustic models. By comparing the mispronunciation detection results among different models trained by a variety

of training criteria, we found those DNN models which have lower frame-level CE objective function achieve better detection performance. To further verify this idea, we present detection results using minimum CE training of the GMM-HMMs, i.e., increasing the log posterior probability of the reference senone label. We found GMM-HMM based acoustic models with better CE criterion (or generally speaking, high frame-level senone classification accuracy) obtain better mispronunciation detection results.

For the second-stage, we use a trainable softmax layer instead of simple phone-dependent thresholds in conventional GOP to learn complex human standard of judgment. The softmax layer takes scores output by DNN as inputs, and generates senone posterior probabilities to compute GOP scores. The softmax layer in this paper is trained by the MFC objective function,^{25,26} which is more suitable for the case when only imbalanced training data are available.

The transfer learning based GOP is further experimented with by using subtly designed experiments: We extend the conventional GOP framework using GMM-HMM acoustic models to transfer learning based GOP, which utilizes GMMs as the hidden layer and a trainable softmax layer as the output layer. Experimental results show the transfer learning based GOP also outperforms the conventional GOP using GMM-HMMs. Transferring from CE trained GMM-HMMs obtains better results than from conventional ASR training criteria such as ML/MPE/MMI in transfer learning based GOP.

In the final part of the experiments, we evaluate mispronunciation detection results using DNN-HMM acoustic models with/without MFC training for NNLR classifier based mispronunciation detection. It is shown MFC training of the softmax layer of the DNN also helps to achieve better results for classifier based mispronunciation detection. We hope this work also serves as a comparison study of mispronunciation detection task with different acoustic models under various training criteria. Our major contributions include the following.

- We propose to use transfer learning based GOP, in which the hidden layers of the DNN are borrowed from native speech recognition and while the softmax layer is trained by task-dependent objective function (MFC). We demonstrate that the hidden layers originally built for speech recognizer are transferable to mispronunciation detection task and transfer learning based GOP can improve the F1-score and shows to be more robust to over-training than MFC training of the whole network parameters.
- The conventional GOP based mispronunciation detection is explored as a two-stage based transfer learning framework: The first stage uses hidden layer(s) to discriminating pronunciations in the acoustical space and the second stage uses a trainable softmax layer to learn the golden standard of a human expert to make the final decision. For the hidden layer that extracting features for phonetic discrimination, we found it is preferable to use cross-entropy minimization as the objective function to train both the DNN-HMM and GMM-HMM acoustic models rather than sequential classification criteria such as MPE and MMI.
- We design and carry out a series of thorough experiments to evaluate various acoustic models (GMM-HMM, DNN-

HMM) trained by a variety of training criteria (MPE, MMI, and CE minimization) under the conventional GOP and the transfer learning based GOP frameworks. Detection results using NNLR classifier based mispronunciation detection are presented as well. These are hopeful to provide as a full investigation of the existing mispronunciation detection models, frameworks and training criteria.

The remainder of this paper is organized as follows. Section II briefly reviews the GOP based mispronunciation detection method using GMM-HMMs and DNN-HMMs. Section III describes the traditional training criteria in acoustic modeling and introduces the MFC objective function. The corresponding derivatives for network parameter update are presented as well. In Sec. IV the experimental evaluations and detailed analysis on the results are described. Section V draws the conclusion of this study.

II. GOP BASED MISPRONUNCIATION DETECTION

A. GOP based framework using GMM-HMMs

The task of mispronunciation detection is to verify whether a pronunciation is correct or not. For phone-level mispronunciation detection, GOP (Ref. 8) is a commonly used quantitative score to measure the quality of a pronunciation. Given a set of speech utterances, let $\mathcal{O}_n$ be the acoustic observations of the n th phonetic segment and q_n the canonical phone label of the segment, the GOP score of segment n is calculated as log phone posterior of the canonical phone q_n .⁸

$$\text{GOP}(\mathcal{O}_n, q_n) = \frac{1}{T_n} \log \frac{p(\mathcal{O}_n | q_n) P(q_n)}{\sum_{q \in Q(n)} p(\mathcal{O}_n | q) P(q)}, \quad (1)$$

where T_n is the duration (in frames) of the segment and $Q(n)$ represents all the possible pronunciations. q_n is the canonical pronunciation. $p(\mathcal{O} | q)$ is the observation probability of model q that can be computed by using a forward-backward pass. $P(q)$ is the prior of phone q . After the GOP score has been calculated, a threshold (or bias) b is applied to make the final decision,

$$\text{GOP}(\mathcal{O}_n, q_n) > b \begin{cases} \text{yes,} & \text{correct pronunciation,} \\ \text{no,} & \text{mispronunciation.} \end{cases} \quad (2)$$

The GMM-HMM based acoustic models are often trained with maximum likelihood (ML) criterion. In Refs. 25 and 26, it was shown that MFC training of the GMM-HMM

based acoustic models achieved significant mispronunciation detection improvement.

B. GOP based framework using DNN-HMMs

Figure 1 shows the schematic diagram of GOP based framework using DNN-HMM based acoustic models. A deep neural network is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs.²⁷ Given the input observation $\mathbf{o}_t$ of frame t , the network is a nonlinear function of $\mathbf{o}_t$ using a sequence of L layers (hidden plus output) and the top softmax layer of the DNN (denoted as the L th layer) directly outputs the posterior probability of a HMM state (senone) s by using the softmax non-linearity

$$p(s | \mathbf{o}_t) = \frac{\exp(z_{t,s}^L)}{\sum_{s'=1}^S \exp(z_{t,s'}^L)}, \quad (3)$$

where $z_{t,s}^L$ is the s th output by the affine transform in the softmax layer (the L th layer) given $\mathbf{o}_t$ as the neural network input, and S the total number of HMM states (senones). Let s_t denote the senone label of frame t generated by forced alignment, the GOP score of q_n in (1) is the time normalized log phone posterior (LPP) of q_n and can be approximated by¹²

$$\begin{aligned} G(n) &= \text{LPP}(q_n | \mathcal{O}_n) = \frac{1}{T_n} \log P(q_n | \mathcal{O}_n) \\ &\approx \frac{1}{T_n} \sum_{t=t_s(n)}^{t_e(n)} \log p(s_t | \mathbf{o}_t), \end{aligned} \quad (4)$$

where $t_s(n)$ and $t_e(n)$ stand for the initial frame and the last frame of segment n , $\mathbf{o}_t$ represents the input observations of the frame t , and s_t is the senone label of the frame t generated by force alignment with the given canonical phone q_n . An advantage of this GOP computation is that it avoids forward-backward computations needed by Eq. (1). If we use Eq. (4), a discriminant function for error detection can be defined as

$$d(n) = -G(n) + b. \quad (5)$$

It can be seen that $d(n) > 0$ is interpreted as erroneous and $d(n) \leq 0$ as correct. Hence the posterior probability of being a mispronunciation is

$$y_n = p(\text{error} | \mathcal{O}_n) = \sigma(d(n)), \quad (6)$$

FIG. 1. (Color online) GOP based mispronunciation detection using DNN.

where $\sigma(\cdot)$ is the sigmoid function

$$\sigma(u) = \frac{1}{1 + e^{-u}}. \quad (7)$$

The decision rule becomes

$$y_n = \sigma(d(n)) > 0.5 \begin{cases} \text{yes,} & \text{mispronunciation,} \\ \text{no,} & \text{correct pronunciation.} \end{cases} \quad (8)$$

If we define a fixed dimensional mispronunciation detection feature vector $\tilde{\mathbf{x}}_n = [-G(n) \ 1]^\top$ for segment n , we get $d(n) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n$ where $\tilde{\mathbf{w}} = [1 \ b]^\top$. The decision rule in Eq. (8) can be written as $\sigma(\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n) > 0.5$ as mispronunciation and otherwise correct pronunciation. As claimed in Ref. 26, the GOP based mispronunciation detection can be considered as a classifier based approach as well, which uses the GOP score as the feature input and use a pre-set threshold as the back-end classifier. More specifically, we can see the GOP based method uses an logistic regression (LR) classifier with a two-dimensional feature input. Training the acoustic models according to some criterion can be regarded as an automatic feature learning process.

III. TRAINING CRITERIA OF THE ACOUSTIC MODELS

A. CE minimization

In GOP based mispronunciation detection that makes use of DNN as the acoustic models, the DNN parameters are normally trained by minimizing the CE objective function that has been widely used in speech recognition:

$$C = - \sum_s d_s \log p(s|\mathbf{o}_t), \quad (9)$$

where d_s is the target probability of state s , typically taking values of one or zero. Minimization of this objective function can be regarded as to maximize the log posterior probability of the reference senone label, in a sense, to maximize senone classification accuracy. Purely optimizing the CE objective for reducing WER on the L1 (native) data might not achieve the best mispronunciation detection result in conventional GOP based method because the human-annotated results on the L2 (non-native) data are not fully exploited in supervised training of the DNN.

B. Accuracy maximization

Automatic mispronunciation detection can be viewed as a binary classification task. A phone segment can be classified as either a correct or erroneous pronunciation by the system. Accuracy, the number of correctly classified phone segments was used as an evaluation metric in Refs. 12 and 14 and can be computed as

$$\text{Accuracy} = \frac{N_{TA} + N_{TR}}{N} \times 100\%, \quad (10)$$

where N_{TA} is the number of TA and N_{TR} is the number of TR. A possible function used for maximizing this training

target can be achieved by maximizing the likelihood as in Ref. 12,

$$\mathcal{L} = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n}, \quad (11)$$

where t_n is the human-annotated result of segment n . $t_n = 1$ is for mispronunciation and 0 otherwise. The optimization can be achieved by minimizing the negative logarithm of the likelihood function¹²

$$-\ln \mathcal{L} = - \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}. \quad (12)$$

Optimization of Eq. (12) can be viewed as to improve the binary classification accuracy. However, using classification accuracy as the training target is only good for symmetric (balanced) data sets.³³ Mispronunciation detection often faces the standard problems that are caused by an imbalanced distribution of classes in the corpus.³³ The amount of phones labeled as errors by human annotators is often extremely small in relation to the overall number of phones. In this case, when optimizing Eq. (12), the gradient based optimization tends to simply tune the threshold (bias in LR) and mark all the phone realizations as correct, which leads to the problem of *accuracy paradox*.³⁴ Therefore, Eq. (12) is not an appropriate training criterion for imbalanced data.

C. Maximum F1-score criterion

For binary classification task on imbalanced data, a variety of performance evaluation metrics have been proposed such as F1-score, Cohen's κ , ROC (receiver operating characteristic), AUC (area under curve), etc. For information retrieval, F1-score, the harmonic mean of precision and recall, is a widely used performance metric, particularly for tasks with imbalanced data sets. F-measure maximizing has been widely discussed in machine learning research community. According to Ref. 36, a variety of methods have been proposed for optimizing F-measures, which generally fall into two paradigms, the decision theoretic analysis (DTA) and empirical utility maximization (EUM).

The DTA approach³⁷ aims at learning a probabilistic model and then predicts labels that maximizes the expected value of the chosen F-measure. The EUM approach aims at training a classifier that maximizes the F-measure on a labeled training set. Due to the high computational complexity in DTA based approaches, the EUM based methods are more popular. In EUM based methods, optimizing the F-measure directly is often difficult as the F-measure is non-convex. Thus, approximation methods are often used instead. Reference 38 gave an efficient algorithm for maximizing a convex lower bound of F-measures for SVMs, and showed it worked well on text classification task. Reference 39 gave an efficient algorithm to maximize a non-convex approximation to F-measures using logistic regression models, and showed it works well on a text summarization problem. More discussions about the two approaches to F-measure maximization can be found in Ref. 36.

In Refs. 25 and 26, we proposed to use the MFC objective function for GMM-HMM based acoustic models, aiming at maximizing F1-score, hence simultaneously improving precision, recall and accuracy. The objective function was similar to that originally proposed in Ref. 39 for logistic regression based classifier on text classification task, while in Refs. 25 and 26 we optimize the GMM-HMM parameters using extended Baum-Welch algorithms for mispronunciation detection task. Our method can be regarded as an EUM based approach. Though simple, the method has been found to be effective and is convenient to implement. The following section describes the MFC objective function and derives the training algorithm of the DNN parameters.

1. The MFC objective function

After the GOP scores of all the phonemes in the utterances are calculated and the mispronunciations are detected according to Eqs. (1) and (2), F1-score can be computed to evaluate the performance of the system based on the detection results of the machine and the annotations of a human evaluator. F1-score is the harmonic mean of precision and recall computed from the number of mispronunciations detected by both the computer and human evaluator, defined as

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

and

$$\text{Precision} = \frac{N_{WW}}{N_D} \times 100\%, \quad (14)$$

$$\text{Recall} = \frac{N_{WW}}{N_W} \times 100\%. \quad (15)$$

N_{WW} is the number of phones marked as mispronunciations by both the computer and the human evaluator. N_D is the total number of mispronunciations detected by the machine, N_W is the number of mispronunciations judged by the human evaluator. By replacing precision and recall in Eq. (13) with Eqs. (14) and (15), F1-score can be rewritten as

$$F1 = \frac{2N_{WW}}{N_D + N_W}. \quad (16)$$

Let $I(\cdot)$ be the step indicator function. In terms of the GOP based error decision rule in Eq. (2), N_{WW} and N_D can be expressed as

$$N_{WW} = \sum_{n=1}^N I(d(n))t_n, \quad (17)$$

$$N_D = \sum_{n=1}^N I(d(n)), \quad (18)$$

$$N_W = \sum_{n=1}^N t_n, \quad (19)$$

where t_n is the human-annotated result of segment n . $t_n = 1$ when segment n is marked as mispronunciation and 0

otherwise. Because of the step indicator function in Eqs. (17) and (18), F1-score is not differentiable, which makes it difficult to optimize F1-score by using gradient based methods. To surpass this problem, we can use the sigmoid function to transform the step indicator function $I(\cdot)$ into a continuous function. By replacing the indicator function $I(\cdot)$ with $\sigma(\cdot)$ in Eqs. (17) and (18), the smoothed number of errors marked by both the machine and human annotator can be

$$N_{WW}^S = \sum_{n=1}^N \sigma(d(n))t_n. \quad (20)$$

Similarly, the smoothed number of errors detected by the machine can be

$$N_D^S = \sum_{n=1}^N \sigma(d(n)). \quad (21)$$

By replacing N_{WW} and N_D with N_{WW}^S and N_D^S in F1-score computation (16), we obtain a smooth form of F1-score, denoted as the maximum F1-score criterion (MFC), which can be written as follows:

$$\mathcal{F}_{\text{MFC}} = \frac{2N_{WW}^S}{N_D^S + N_W} \quad (22)$$

$$= \frac{2 \sum_{n=1}^N \sigma(d(n))t_n}{\sum_{n=1}^N \sigma(d(n)) + N_W}. \quad (23)$$

From the equation we can see, as N_W remains fixed given the human-annotation results, the MFC objective function can be maximized by simultaneously increasing N_{WW}^S and decreasing N_D^S .

2. MFC training of the deep neural network

When gradient based method is used for DNN optimization, the derivatives of $\mathcal{F}_{\text{MFC}}$ with respect to the neural network weights are needed. We only provide the derivative of $\mathcal{F}_{\text{MFC}}$ with respect to the affine transform $\mathbf{W}^L$ in the top softmax layer (the L th layer of the network). Given the augmented input $\mathbf{o}_t$ at time t , let $\mathbf{z}_t^L$ be the output of affine transform in the top softmax layer, the derivatives of the objective function $\mathcal{F}_{\text{MFC}}$ with respect to the s -th element of $\mathbf{z}_t^L$ is

$$\frac{\partial \mathcal{F}_{\text{MFC}}}{\partial z_{t,s}^L} = \frac{2}{N_D^S + N_W} \frac{\partial N_{WW}^S}{\partial z_{t,s}^L} - \frac{2N_{WW}^S}{(N_D^S + N_W)^2} \frac{\partial N_D^S}{\partial z_{t,s}^L}, \quad (24)$$

where

$$\frac{\partial N_{WW}^S}{\partial z_{t,s}^L} = \sigma'(d(n)) \frac{\partial d(n)}{\partial z_{t,s}^L} t_n, \quad (25)$$

$$\frac{\partial N_D^S}{\partial z_{t,s}^L} = \sigma'(d(n)) \frac{\partial d(n)}{\partial z_{t,s}^L}. \quad (26)$$

FIG. 2. (Color online) (a) Transfer learning based GOP using DNN: The hidden layers and the softmax layer for speech recognition are trained using CE minimization on the L1 training set; The softmax layer for mispronunciation detection is trained using MFC on the L2 training set. (b) NNLR classifier based mispronunciation detection (Ref. 12).

In Eqs. (25) and (26), $\sigma'(u) = \sigma(u)(1 - \sigma(u))$ and

$$\frac{\partial d(n)}{\partial z_{t,s}} = \frac{1}{T_n} (p(s|\mathbf{o}_t) - A(s, s_t)). \quad (27)$$

$A(s, s_t)$ is the “senone accuracy” of state s and is equal to 1 when s and the aligned label s_t are the same, and 0 otherwise. $p(s|\mathbf{o}_t)$ is the senone posterior in Eq. (3). The derivative of $\mathcal{F}_{\text{MFC}}$ with respect to $\mathbf{W}^L$ in segment n is

$$\frac{\partial \mathcal{F}_{\text{MFC}}}{\partial \mathbf{W}^L} = \sum_{t=t_s(n)}^{t_e(n)} \frac{\partial \mathcal{F}_{\text{MFC}}}{\partial \mathbf{z}_t^L} (\mathbf{a}_t^{L-1})^\top, \quad (28)$$

where $\mathbf{a}_t^{L-1}$ is the output of the layer prior to the softmax layer at time t . The derivative of $\mathcal{F}_{\text{MFC}}$ with respect to the deeper layer weights can be computed according to the chain rule.

3. Transfer learning based GOP Using DNN

In MFC training of the DNN, rather than train the whole DNN parameters, we can train only the softmax layer of the network and keep the hidden layers fixed. This can be viewed as a transfer learning based GOP, in which the hidden layers are borrowed from speech recognizer while the softmax layers are trained by MFC for mispronunciation detection. The GOP based mispronunciation detection framework using transfer learning based GOP is depicted in Fig. 2(a).

IV. EXPERIMENTS AND RESULTS

A. Databases

The proposed method is evaluated on a mispronunciation detection task for Uighur college students who had been learning Mandarin Chinese in Xinjiang University. Two databases are used.

- (1) The L1 database is the “863 project” Mandarin speech database uttered by native Mandarin speakers, which is a corpus for continuous speech recognition collected for the Chinese National “863 Project.” It has a total of about 110-h recordings spoken by 160 speakers (80 females and 80 males). The database contains 92 243 utterances. In our experiments, 86 271 utterances are selected as the L1 training set, the rest are used for evaluating speech recognition performance of the acoustic models.
- (2) L2 database is a non-native speech corpus, containing mandarin speech uttered by 100 Uighur speakers (50 male and 50 female). Each speaker was asked to read

three sets of prompted texts. Each set contains 50 single Chinese characters, 25 words and 20 short sentences. The database has been annotated by two expert phoneticians (annotator A and annotator B). The annotation efforts involved pinpointing phone-level mispronunciations and tonal errors at the syllable level. At the sentence level, the two annotators also provided three mean opinion scores (MOS) indicating the correctness of pronunciations, speaking fluency and prosody of the utterance. After cleaning, the database contains about 14.1 h of speech, 25 673 utterances. 18 643 utterances of the database are used as the L2 training set and 7030 are used for mispronunciation detection evaluation (L2 test set). Table I shows the four subsets of the data.

Table II shows the annotation results on the L2 corpus by the two phoneticians. It can be seen that the Uighur speakers are more likely to produce tonal errors. In this paper, we focus only on the phone-level mispronunciations without consideration of tonal error detection, which normally belongs to another separate research topic.^{40,41} To compare the machine-human results with the human-human performance, the inter-annotator agreement results for the pronunciation error annotations are also listed. By assuming annotator A as the golden standard, the detection results of annotator B against annotator A are presented in Table III. In Table III, the inter-rater agreement is also measured by Cohen’s kappa coefficient (κ).³⁵

B. Configurations and results

Mispronunciation detection results with various acoustic models (GMM-HMMs, DNN-HMMs) trained by different objective functions in different mispronunciation detection frameworks (conventional GOP, transfer learning based GOP and classifier based method) are shown in this section. Precision, Recall and F1-score are used as performance evaluation metrics. The detections results are also presented in terms of FR, FA, TA, TR, classification accuracy, and Cohen’s κ .

TABLE I. Division of the native (L1) and nonnative (L2) corpora.

Data set	Duration	# utterances	# phones	# errors
L1 Training	100.6 h	86 271	2095K	NA
L1 Testing	6.9 h	5972	142 K	NA
L2 Training	10.0 h	18 643	98 174	3077
L2 Testing	4.1 h	7030	39 668	1022

TABLE II. Annotation results on non-native (L2) corpus.

	A	B
# Phone errors	4099	4039
# Tonal errors	21 611	22 561
Correctness	3.84	3.78
Fluency	3.93	3.95
Prosody	2.79	2.71

1. Conventional GOP using DNN-HMMs

a. CE training on the L1 dataset. For DNN based acoustic models, we use a network with 5 hidden layers and 1024 units per layer and sigmoid activation. The spectral front-end uses a 39-dimensional vector, consisting of 13 MFCCs and their Δ and $\Delta\Delta$ with mean normalization. The HMM set has 67 phones (28 initials and 37 finals plus silence and short pause). The input feature of the DNN has a dimension of 429 by concatenating a context of 5 frames left and right of the current frame (11 frames in total). The output of the softmax layer has a dimension of 205, which represents the posterior probabilities of context-independent HMM states (senones).

First the DNN is trained using a standard speech recognition recipe: The weights and biases of hidden layers are initialized with stacked restricted Boltzman machines (RBMs)²⁷ using the Kaldi toolkit⁴² on the L1 training set. Then a softmax layer is added on the top of the network and the DNN parameters are fine-tuned according to CE minimization on the L1 training set. Table IV shows the MFC objective function $\mathcal{F}_{\text{MFC}}$ and mispronunciation detection results on both the L2 training set and testing set by training different layer(s) of the DNN using different criteria. In addition to mispronunciation detection, we carried out syllable output speech recognition experiments to evaluate recognition performance of the acoustic models. Language model is removed from decoding process to obtain a good evaluation of the acoustic resolution. Speech recognition performance in terms of syllable error rate (SER) using the corresponding models are presented in Table V.

Before detecting mispronunciations using the DNN based acoustic models, we need to obtain the optimal phone thresholds. According to the discussion in Sec. II B, the thresholds can be optimized using gradient based method along with the network parameters. In this paper, we first compute the GOP scores on all the L2 training set using Eq. (4) and then tune the phone-dependent thresholds according to MFC using grid search adopted in Ref. 25. Finally the mispronunciations are detected on the L2 test set. Using the

TABLE III. Inter-annotator results on the L2 corpus.

	P	R	F1	A	κ
Phone errors	78.7	77.5	78.1	98.7	0.783
Tonal errors	88.0	91.9	89.9	93.0	0.837
	TA	FA	TR	FR	—
Phone errors	96.5	0.7	2.3	0.6	—
Tonal errors	61.8	2.7	31.2	4.2	—

TABLE IV. Detection results in $\mathcal{F}_{\text{MFC}}$ and F1-score (%).

Criterion		Training		Testing	
Hidden	Softmax	$\mathcal{F}_{\text{MFC}}$	F1	$\mathcal{F}_{\text{MFC}}$	F1
CE _{L1}	CE _{L1}	0.441	48.5	0.413	45.4
DBN _{L1}	CE _{L1}	0.382	40.7	0.367	39.9
sMBR _{L1}	sMBR _{L1}	0.447	47.4	0.420	44.9
CE _{L2}	CE _{L2}	0.306	35.9	0.298	38.3
CE _{L2} –	CE _{L2} –	0.413	45.3	0.341	40.9
MFC _{L2}	MFC _{L2}	0.864	88.3	0.493	49.9
CE _{L1}	MFC _{L2}	0.598	63.5	0.508	53.6
DBN _{L1}	MFC _{L2}	0.551	59.2	0.445	47.3
sMBR _{L1}	MFC _{L2}	0.596	62.4	0.510	52.5
CE _{L2}	MFC _{L2}	0.550	58.3	0.473	51.0

DNN of which the hidden layers and the softmax layer are all trained by CE minimization on the L1 training set (denoted as CE_{L1} for the hidden and CE_{L1} for the softmax in Table IV), the F1-score on the L2 test set is 45.4%.

b. DBN based hidden-layer. Table IV then presents mispronunciation detection results using a DNN of which the softmax layer trained by CE minimization on the L1 training set (denoted as CE_{L1}) while the hidden layer are deep belief network (DBN) composed by a stack of RBMs trained in an unsupervised manner on the L1 training set (DBN_{L1}). For speech recognition, it is shown in Table V that DNN with DBN_{L1} hidden layers and CE_{L1} softmax layer yields inferior SER than DNN with CE_{L1} trained hidden layers and softmax layer. For mispronunciation detection, it is also observed that DNN with DBN_{L1} hidden layers and CE_{L1} softmax layer obtains lower F1-score than DNN with CE_{L1} trained hidden layers and softmax layer. These results indicate the features extracted by supervised training on native (L1) speech data are more discriminating and benefit both L1 speech recognition and L2 mispronunciation detection. The output by L1 phonetically-aware hidden layers shows better in mispronunciation detection task than unsupervised derived features output by DBN.

c. sMBR training. Based on the fact that GMM-HMM based speech recognition systems are often discriminatively trained using sequence-based criteria, such as MPE (Ref. 23) or MMI (Ref. 22) that are more directly related to recognition

TABLE V. Speech recognition results (%) in SER.

Criterion		L1		L2	
Hidden	Softmax	Training	Testing	Training	Testing
CE _{L1}	CE _{L1}	6.6	14.8	55.7	55.2
DBN _{L1}	CE _{L1}	24.2	26.3	64.9	65.0
sMBR _{L1}	sMBR _{L1}	5.5	13.5	63.2	60.4
CE _{L2}	CE _{L2}	49.2	48.4	16.9	33.9
CE _{L2} –	CE _{L2} –	48.4	47.7	18.5	34.0
MFC _{L2}	MFC _{L2}	63.8	65.8	97.8	95.8
CE _{L1}	MFC _{L2}	15.7	23.9	57.8	57.4
DBN _{L1}	MFC _{L2}	61.3	63.1	86.9	85.8
sMBR _{L1}	MFC _{L2}	12.1	20.6	61.4	59.5
CE _{L2}	MFC _{L2}	56.7	57.7	43.2	54.5

accuracy, DNN based acoustic models can be trained with sequence classification criteria using exactly the same lattice-based methods and shows considerably better performance.³² It is seen in Table V that sMBR training of the DNN has yielded better speech recognition result. The sMBR trained DNN is then used for detecting mispronunciations using conventional GOP method. Mispronunciation detection results using sMBR trained DNN are reported in Table IV. By comparing the detection results of sMBR_{L1} + sMBR_{L1} with CE_{L1} + CE_{L1}, we also observed that sMBR training of the DNN, which has shown better speech recognition performance, does not clearly helps to improve mispronunciation detection.

d. CE training on the L2 dataset. In mispronunciation detection, the acoustic models are normally initialized using the native (L1) speech data. We report the detection results using DNN trained on the L2 training set using CE minimization to find out whether better recognition performance on the L2 dataset leads to better detection performance. It is seen in Table V when the whole network is trained by CE minimization on the L2 training set (CE_{L2} for the hidden layers and CE_{L2} for the softmax layer), SER on the L2 test set is 33.9%, better than that obtained by DNN trained with CE minimization on the L1 training set (CE_{L1} for the hidden and CE_{L1} for the softmax, 55.2%). This suggests that the acoustic models do have, to some extent, adapted to the non-standard L2 accent and obtain recognition improvement on non-native (L2) speech. However, mispronunciation detection F1-score does not show better than that of CE trained DNN on the L1 dataset.

We further experiment with CE training on a cleaned L2 dataset. We remove all the mispronounced phone segments in the L2 data and train the DNN using CE on the cleaned data (denoted as CE_{L2-}). For speech recognition, it can be seen in Table V that the SERs on the L1 dataset decrease while the SERs on the L2 dataset increase slightly. We think this is because training the acoustic models on the cleaned L2 data will be a little more adaptive to native pronunciations and reduce the tolerance to mispronunciations. As a further result, for mispronunciation detection, it can be observed the DNN trained by CE_{L2-} shows better detection F1-scores than DNN trained by CE_{L2}. This is because the CE_{L2-} model improves discriminations between the correct and incorrect pronunciations.

2. Transfer learning based GOP using DNN-HMMs

a. MFC training of all layers. From the DNN with all layers trained by CE on L1 training set, we conduct MFC training of all the layers. The MFC training procedure is an iterative training process using gradient ascent according to derivative in Eqs. (24)–(28), which can be summarized as follows.

- (1) Initialize the DNN based acoustic models;
- (2) Iterative MFC training:
 - (a) Compute the GOP scores of all the phone segments in the L2 training set;

- (b) Search for the phone-dependent thresholds $\mathbf{b}$ that maximize $\mathcal{F}_{\text{MFC}}$ on the L2 training set;
- (c) Compute N_{WW}^S and N_D^S using the GOP scores and thresholds obtained in 2(a) and 2(b);
- (d) Compute the derivatives of $\mathcal{F}_{\text{MFC}}$ with respect to the corresponding layer (softmax layer, hidden layers) parameters of the DNN according to Eqs. (24)–(28);
- (e) Update the network parameters;
- (f) Goto step 2(a) unless convergence or maximum number of iterations is reached.

In our implementation, the errors are back-propagated after each training phone-segment. Strictly, N_{WW}^S and N_D^S should be re-calculated whenever the network parameters are changed. So, it is time consuming when the batch-size is small. In this work, we re-calculated N_{WW}^S and N_D^S only after a training epoch on the entire data set. Based on the fact that N_{WW}^S and N_D^S do not change too much in each training epoch, this implementation shows to work well in practice.

As for the results of MFC training, it is found MFC training of all the layers (both the hidden layers and the softmax layer are trained by MFC on the L2 training set, denoted as MFC_{L2} + MFC_{L2}) achieves an F1-score of 49.9% on the L2 test set, better than the DNN of which both the hidden layers and softmax layer are trained by CE on the L1 training set (CE_{L1} + CE_{L1}, 45.4%). However, it can be observed that the MFC objective function ($\mathcal{F}_{\text{MFC}}$) improves dramatically while only a moderate F1 improvement on the test set is obtained, indicating that MFC training of the whole network is liable to over-training.

b. Transfer learning from CE_{L1} hidden layers. Instead of training all the layers by MFC, we tune only the affine transform in the softmax layer according to MFC (CE_{L1} for the hidden layers and MFC_{L2} for the softmax). An F1-score of 53.6% is achieved, significantly ($p < 0.001$) better than conventional GOP based method using CE_{L1} trained DNN (F1 = 45.4%) and the DNN with all layers trained by MFC (F1 = 49.9%), with only a moderate improvement of $\mathcal{F}_{\text{MFC}}$. Since the hidden layers are remain fixed, this approach can be viewed as a transfer learning approach with hidden layers borrowed from L1 speech recognition task. Mispronunciation detection task utilizes feature representations output by the hidden layers trained for L1 ASR while the softmax layers are discriminatively fine-tuned by task-dependent criteria. The results indicate that the feature transformation represented by the hidden layers in the L1 data trained DNN can be effectively transferred to detect mispronunciations in L2 speech.

c. Transfer learning from other differently trained hidden layers. In addition to transfer from CE_{L1} trained hidden layers, we further experiment with hidden layers borrowed from three differently trained DNNs. All the three setups train the top softmax layer according to MFC on the L2 training set (MFC_{L2}). The hidden layers of the DNNs are, respectively, DBN trained on the L1 training data (DBN_{L1}), sMBR trained on the L1 training data (sMBR_{L1}), and CE

trained on the L2 training set (CE_{L2}). We can see MFC training of the softmax layer combining with these hidden layer setups also leads to better detection results, but is inferior to CE_{L1} trained hidden layer + MFC_{L2} trained softmax layer. By comparing the detection result of CE_{L1} + MFC_{L2} with that of CE_{L2} + MFC_{L2} , we can conclude that better capability of discriminating pronunciations in standard native speech (L1 speech) rather than none-native (L2) speech helps to obtain better detection result.

Table VI demonstrates the detection results in terms of precision and recall obtained by the corresponding DNNs. The numbers of TA, FA, TR, and FR normalized by the total number of phone realizations in the data set, detection accuracy and Cohen's κ are displayed in Table VII. Results suggest that the maximization of F1-score yields improvements of both precision and recall over the CE_{L1} + CE_{L1} baseline. The six metrics in Table VII can also be improved by MFC training of the DNN.

Based on the above results and comparisons, we have seen GOP based mispronunciation detection can be casted into a two-stage based transfer learning framework: Discriminating the pronunciations in the acoustic space (more specifically, the native pronunciation space) using features extracted by the hidden layers (trained by CE_{L1}), and approximating the golden standard of a human annotator using a trainable softmax layer (trained by MFC_{L1}). A further verification of this hypothesis will be carried out in Sec. IV B 4.

3. Conventional GOP using GMM-HMMs

This section demonstrates conventional GOP based mispronunciation detection results using GMM-HMMs for comparisons with those using DNN-HMMs. Some results were originally presented in Ref. 26 and are listed here for the completeness of this work and convenience for discussions. Similar to the DNN-HMMs, context-independent (CI) GMM-HMMs are used due to the limited non-native (L2) training data, especially the limited amount of mispronounced training data. Another reason for using CI models is that context information is not easy to implement considering possible mispronunciations around the target phone.¹⁰ The spectral front-end uses 39 dimensional vector, consisting of 13 MFCCs and their Δ and $\Delta\Delta$ with mean normalization. The HMM set has 67 phones (28 initials and 37 finals plus silence and short

TABLE VI. Detection results in precision and recall (%).

Criterion		Training		Testing	
Hidden	Softmax	P	R	P	R
CE_{L1}	CE_{L1}	53.1	44.6	47.1	43.9
DBN_{L1}	CE_{L1}	45.5	36.9	40.7	39.1
$sMBR_{L1}$	$sMBR_{L1}$	49.5	45.4	44.3	45.5
CE_{L2}	CE_{L2}	43.8	30.5	36.2	40.7
CE_{L2-}	CE_{L2-}	51.0	40.9	38.9	43.2
MFC_{L2}	MFC_{L2}	97.3	80.8	56.8	44.4
CE_{L1}	MFC_{L2}	70.9	57.5	60.0	48.5
DBN_{L1}	MFC_{L2}	66.3	53.6	51.3	43.9
$sMBR_{L1}$	MFC_{L2}	69.0	56.9	58.4	47.7
CE_{L2}	MFC_{L2}	66.9	51.7	51.4	50.7

TABLE VII. Detection results in TA, FA, TR, FR, accuracy (A)(%), and Cohen's κ .

Data	Hidden/top	TA	FA	TR	FR	A	κ
L2 Train	CE_{L1}/CE_{L1}	95.6	1.7	1.4	1.2	97.0	0.477
	DBN_{L1}/CE_{L1}	95.5	2.0	1.1	1.4	96.6	0.376
	$sMBR_{L1}/sMBR_{L1}$	95.4	1.7	1.4	1.3	96.9	0.469
	CE_{L2}/CE_{L2}	95.6	2.2	1.0	1.2	96.6	0.354
	CE_{L2-}/CE_{L2-}	95.6	1.9	1.3	1.2	96.9	0.438
	MFC_{L2}/MFC_{L2}	96.8	0.6	2.5	0.1	99.3	0.874
	CE_{L1}/MFC_{L2}	96.1	1.3	1.8	0.7	97.9	0.628
	DBN_{L1}/MFC_{L2}	96.0	1.5	1.7	0.9	97.7	0.577
	$sMBR_{L1}/MFC_{L2}$	96.1	1.4	1.8	0.8	97.9	0.614
	CE_{L2}/MFC_{L2}	96.1	1.5	1.6	0.8	97.7	0.570
L2 Test	CE_{L1}/CE_{L1}	96.2	1.4	1.1	1.3	97.3	0.435
	DBN_{L1}/CE_{L1}	96.0	1.6	1.0	1.5	97.0	0.372
	$sMBR_{L1}/sMBR_{L1}$	95.9	1.4	1.2	1.5	97.1	0.438
	CE_{L2}/CE_{L2}	95.5	1.5	1.1	1.8	96.6	0.387
	CE_{L2-}/CE_{L2-}	95.7	1.5	1.1	1.7	96.8	0.393
	MFC_{L2}/MFC_{L2}	96.6	1.4	1.1	0.8	97.7	0.489
	CE_{L1}/MFC_{L2}	96.6	1.3	1.3	0.8	97.9	0.543
	DBN_{L1}/MFC_{L2}	96.3	1.4	1.1	1.1	97.4	0.457
	$sMBR_{L1}/MFC_{L2}$	96.5	1.3	1.2	0.9	97.8	0.510
	CE_{L2}/MFC_{L2}	96.2	1.2	1.3	1.2	97.5	0.507

pause), each HMM state is a mixture of 8 Gaussians. Detection results using GMM-HMMs are listed in Table VIII. The GMM-HMMs are trained by different criterion (ML/MPE/MMI/MFC). The results in precision and recall are listed in Table IX. Speech recognition results using the corresponding models are listed in Table X.

The baseline GMM-HMMs are ML trained on either the L1 or the L2 training set (ML_{L1} or ML_{L2}). The corresponding detection F1-score on the L2 test set are, respectively, 0.382 and 0.381. Upon these models, the GMM-HMMs are trained by some discriminative training criterion that aims at reducing WER popular in speech recognition (MMI/MPE) on either the L1 or L2 training set. It is shown whilst the acoustic models have been well trained by MMI and MPE before they are used for mispronunciation detection, only MPE training on the L1 data (MPE_{L1}) shows some F1-score improvement (2.7%). This indicates that discriminative training aiming at reducing WER does not explicitly lead to F1-score improvement in automatic mispronunciation detection.

In the final part of Table VIII, we report the detection results of MFC training of the GMM-HMMs. The

TABLE VIII. Mispronunciation detection results using GMM-HMMs.

Criterion	Training		Testing	
	$\mathcal{F}_{MFC}$	F1	$\mathcal{F}_{MFC}$	F1
ML_{L1}	0.368	38.6	0.367	38.2
ML_{L2}	0.367	38.8	0.363	38.1
MMI_{L1}	0.365	38.6	0.360	38.7
MMI_{L2}	0.321	33.9	0.320	34.6
MPE_{L1}	0.396	41.4	0.392	40.9
MPE_{L2}	0.364	38.4	0.350	37.5
MFC_{L2}	0.616	66.5	0.454	48.8

TABLE IX. Detection results in precision and recall (%) using GMM-HMMs.

Criterion _{train data}	Training		Testing	
	P	R	P	R
ML _{L1}	41.8	35.9	37.6	38.9
ML _{L2}	43.2	35.3	38.2	38.1
MMI _{L1}	42.0	35.7	38.2	39.3
MMI _{L2}	41.4	28.8	37.9	31.9
MPE _{L1}	42.9	40.0	39.5	42.3
MPE _{L2}	40.9	36.3	33.4	42.7
MFC _{L2}	79.5	57.2	57.1	42.7

GMM-HMMs are initialized by ML_{L1} and then optimized according to MFC using the extended Baum-Welch algorithms proposed in Refs. 25 and 26. Refer to Refs. 25 and 26 for more details about MFC training of the GMM-HMMs. It is seen the F1-score on the L2 test set improves to 48.8%, significantly better than that from ML trained GMM-HMMs. So far, we have seen MFC training of the acoustic models benefits both GMM-HMM and DNN-HMM based mispronunciation detection. From the SERs reported in Table X, we can see when the MFC trained GMM-HMMs are used for speech recognition, SER rises dramatically, showing that detection F1-score and speech recognition accuracy are two different objectives. This is similar to DNN based acoustic models.

4. Transfer learning based GOP using GMM-HMMs

In conventional GOP based mispronunciation detection, the golden standard is approximated by using tunable thresholds, which might be too simple to precisely learn complex judgment criterion. As observed in GOP based mispronunciation detection based on DNN-HMMs, using a softmax layer trained by MFC has shown better mispronunciation detection results. Inspired by this, it is reasonable that we can extend the conventional GOP based framework using GMM-HMMs to a new detection architecture, which uses GMMs as the hidden layer and a softmax output layer (transfer learning GOP using GMM-HMMs). Given the input MFCC frames, the GMM hidden layer outputs acoustic scores (log likelihood) and feeds them into the softmax layer which outputs the senone posterior probabilities to compute the GOP scores. The affine transform in the softmax layer is first initialized randomly and then fine-tuned by MFC. The system architecture is depicted in Fig. 3.

TABLE X. Recognition results (%) in SER using GMM-HMMs.

Criterion	L1		L2	
	Training	Testing	Training	Testing
ML _{L1}	32.5	34.1	68.4	67.5
MMI _{L1}	25.2	27.9	69.0	67.3
MPE _{L1}	22.6	25.4	68.0	67.0
ML _{L2}	33.7	35.0	55.2	57.4
MMI _{L2}	37.6	38.6	34.2	43.9
MPE _{L2}	41.3	41.7	32.2	41.9
MFC _{L2}	58.7	60.3	77.2	76.3

FIG. 3. (Color online) Transfer learning based GOP framework using GMMs as the hidden layer.

For the GMM based hidden layer, it is possible that the Gaussians are borrowed from GMM-HMMs trained by different training criteria. By comparing speech recognition and mispronunciation detection results from GMM-HMMs and DNN-HMMs, it is natural to argue that better native (L1) speech recognition performance leads to better mispronunciation detection result. However, what has been puzzling in our previous work²⁶ and the results presented in Table VIII, is that the GMM-HMMs trained by MMI or MPE which aims at reducing WER showed opposite mispronunciation detection results: while MPE trained GMM-HMMs on the L1 dataset show better F1-score than ML trained GMM-HMMs, MMI training even degrades the detection performance. On the L2 training set, MPE and MMI training of the GMM-HMMs though greatly improve speech recognition on the L2 test set, also degrade mispronunciation detection performance. As for the DNN-HMMs, it has also been shown that sMBR training improves speech recognition, nevertheless, the detection F1-score degrades.

According to these results, it is necessary to find out what objective function is suitable for training the hidden layers (DNN or GMM) for transfer learning in GOP based mispronunciation detection. In Table XI we show the average frame log posterior probabilities of the canonical senones using different training approaches. The corresponding mispronunciation detection results using these models within the conventional GOP and the transfer learning based GOP frameworks are summarized in Table XII. We can see compared with the ML and MMI trained GMM-HMMs, the MPE trained GMM-HMMs show higher frame-level senone posterior probabilities (in a sense, better senone classification accuracy) and accordingly, achieve better detection results, in both the conventional GOP and the softmax GOP frameworks. The same thing happens to DNN based acoustic models: the CE trained DNN with lower CE objective than the sMBR trained DNN yields better detection F1-score.

TABLE XI. Frame-level cross entropy and senone accuracy.

Model	Criterion	$\mathcal{F}_{CE}$	Accuracy (%)
GMM	MLE _{L1}	3.16	57.8
GMM	MMI _{L1}	3.20	58.2
GMM	MPE _{L1}	2.81	58.4
GMM	CE _{L1}	1.74	59.2
DNN	CE _{L1}	0.28	89.7
DNN	sMBR _{L1}	0.50	85.4

TABLE XII. Mispronunciation detection results using GOP and transfer learning GOP.

	Hidden	Criterion	F1	P	R	A
GOP	GMM	MLE _{L1}	38.2	37.6	38.9	96.8
	GMM	MMI _{L1}	38.7	38.2	39.3	96.8
	GMM	MPE _{L1}	40.9	39.6	42.3	96.8
	GMM	CE _{L1}	41.8	39.9	43.9	96.9
	DNN	CE _{L1}	45.4	47.1	43.9	97.3
	DNN	sMBR _{L1}	44.9	44.3	45.5	97.1
Transfer Learning GOP	GMM	MLE _{L1}	42.0	40.3	43.8	96.9
	GMM	MMI _{L1}	41.6	40.9	42.3	96.9
	GMM	MPE _{L1}	43.6	43.3	44.0	97.1
	GMM	CE _{L1}	44.4	44.0	44.9	97.1
	DNN	CE _{L1}	53.6	60.0	48.5	97.9
	DNN	sMBR _{L1}	52.5	58.4	47.7	97.8

To further validate the idea that CE minimization leads to better F1-score in transfer learning based GOP, we train the GMM-HMMs according to CE minimization on the L1 training set (CE_{L1}), which is achieved by optimizing the following objective function:

$$\mathcal{F}_{CE} = - \sum_t \log \frac{p^k(\mathbf{o}_t | s_r)}{\sum_{s=1}^S p^k(\mathbf{o}_t | s)}, \quad (29)$$

where T is all the training frames in the L1 training set and $0 < k < 1$ is a commonly applied exponential scaling factor in discriminative training to reduce dynamic range of the probabilities.^{23,24} s_r is the referenced senone label using Viterbi alignment against the phone transcriptions. S is the total number of states (senones) in the HMM set. The extended Baum-Welch (EBW) algorithms²⁴ are applied to update the Gaussian means and variances,

$$\mu_{sm} = \frac{\chi_{sm}^n - \chi_{sm}^d + D_{sm} \hat{\mu}_{sm}}{\beta_{sm}^n - \beta_{sm}^d + D_{sm}}, \quad (30)$$

$$\sigma_{sm}^2 = \frac{Y_{sm}^n - q Y_{sm}^d + D_{sm} (\hat{\sigma}_{sm}^2 + \hat{\mu}_{sm}^2)}{\beta_{sm}^n - \beta_{sm}^d + D_{sm}} - \mu_{sm}^2, \quad (31)$$

where $\hat{\mu}_{sm}$ and $\hat{\sigma}_{sm}$ are, respectively, the current mean and variance for Gaussian m in state s (consider Gaussians with a single dimension for simplicity). μ_{sm} and σ_{sm} are the mean and variance to be updated. The smoothing factor D_{sm} is a Gaussian-dependent factor which is empirically determined for each Gaussian component and typically set to $D_{sm} = E \beta_{sm}^d$. The constant $E = 3.0$ was chosen in the experiments. In Eqs. (30) and (31), the occupation-data, sum-of-data, and sum-of-square-data of the “numerator” are

$$\beta_{sm}^n = k \sum_{t=1}^T \psi_{sm}(t), \quad (32)$$

$$\chi_{sm}^n = k \sum_{t=1}^T \psi_{sm}(t) o(t), \quad (33)$$

$$Y_{sm}^n = k \sum_{t=1}^T \psi_{sm}(t) o^2(t), \quad (34)$$

where $\psi_{sm}(t)$ is the posterior probability of being in Gaussian mixture m within state s at time t . The statistics of the “denominator” can be calculated as follows:

$$\beta_{sm}^d = k \sum_{t=1}^T \psi_s(t) \psi_{sm}(t), \quad (35)$$

$$\chi_{sm}^d = k \sum_{t=1}^T \psi_s(t) \psi_{sm}(t) o(t), \quad (36)$$

$$Y_{sm}^d = k \sum_{t=1}^T \psi_s(t) \psi_{sm}(t) o^2(t), \quad (37)$$

where $\psi_s(t)$ is the posterior probability of being in state s against all the possible states at time t .

The CE minimization training of the GMMs is an iterative process starting from the ML trained GMM-HMMs. It is seen in Table XI the CE objective decreases and hence the frame-level senone classification accuracy improvement is accordingly obtained in general. As for mispronunciation detection, it is shown in Table XII the CE trained GMM-HMMs achieve better detection results than the MPE/MPE/MMI trained models. By further comparing the results of conventional GOP and transfer learning based GOP that has the same hidden layer(s), mispronunciation detection F1-score improves substantially, indicating the superiority of approximating the human standard using a MFC trained softmax layer to simple thresholds in conventional GOP based framework.

5. NNLR classifier based mispronunciation detection

Instead of training the acoustic models in GOP based mispronunciation, an alternative to incorporating the golden standard of human annotator in a mispronunciation detection system is to follow the classifier based approach, i.e., train an additional back-end classifier supervised by the annotated results.^{10,12} For example, in Ref. 12, mispronunciation detection was casted into a binary classification task. The outputs of DNN based acoustic models were used to compute segmental mispronunciation detection features as the input of an NNLR based classifier and human-annotated results (golden standard) were used to train back-end classifier. In classifier based approach, the acoustic models are used as feature extractor which remain fixed and are not optimized.

The NNLR classifier based mispronunciation detection scheme¹² using DNN acoustic models is demonstrated in Fig. 2(b). The NNLR classifier is a three-layer neural network with a hidden layer and a output layer that consists of multiple logistic regression (LR) classifiers, each for classifying one phone. DNN acoustic models are exploited to compute mispronunciation detection features and NNLR classifier is adopted as the back-end to detect mispronunciations. Different from the GOP based approach, NNLR classifier exploits log phone posterior (LPP) of all the possible phones as the mispronunciation features, not merely LPP of the

canonical phone. The LPP feature of segment n is expressed as

$$f_{LPP}(n) = [LPP(q_1|\mathcal{O}_n), LPP(q_2|\mathcal{O}_n), \dots, LPP(q_M|\mathcal{O}_n)]^\top, \quad (38)$$

where M is the number of all the possible phones. Furthermore, log posterior ratio (LPR) feature between the competitive phone against the canonical phone were used to detect salient pronunciation replacement¹²

$$f_{LPR}(n) = [LPR(q_1), LPR(q_2), \dots, LPR(q_M)]^\top, \quad (39)$$

where

$$LPR(q_i) = \frac{1}{T_n} (\log P(q_i|\mathcal{O}_n) - \log P(q_n|\mathcal{O}_n)). \quad (40)$$

q_i represents a competitive phone, more details about implementing LPP and LPR features for classifier based mispronunciation detection can be found in Refs. 10 and 12.

The training of the NNLR classifier follows the recipe in Ref. 12: The weights and biases of hidden layers are initialized with pretrained stacked RBMs, then phone-specific LR classifiers are added above the hidden layers with randomly initialized weights. Finally, all the NNLR parameters are discriminatively optimized according to MFC with back-propagation using gradient ascent. LPR feature was not adopted in this work because no obvious improvement was shown in our experiments.

The NNLR classifier based mispronunciation detection results are shown in Table XIII. First we use LPPs computed by the CE trained DNN as features for the NNLR based classifier, which achieves an F1-score of 50.4%, 5.0% better than that by GOP framework using CE trained DNN (45.4%). This mainly benefits from the supervision of the golden standard in the training of the back-end classifier. Then we use LPPs from the MFC trained DNN as inputs to the NNLR classifier, an F1-score of 54.5% is yielded, 4.1% better than that using LPPs from CE trained DNN (50.4%). This emphasizes the importance of feature learning when using classifier based mispronunciation detection. Finally, we compare the GOP and the classifier based approaches both using the DNN acoustic models with MFC trained softmax layer. It is seen the GOP based approach (53.6%) is comparable to the classifier based approach (54.5%), indicating training the softmax layer of the front-end DNN has an almost equivalent effect of overall training of the DNN and an additional back-end classifier. Furthermore, the softmax layer is trained by MFC, indicating that it not necessary to build an explicit mapping to L1 senone label for native speech recognition in both GOP and classifier based mispronunciation detection.

TABLE XIII. Results (%) with NNLR classifier in (F1)-score, (P)recision, (R)ecall, and (A)ccuracy.

DNN	Classifier	F1	P	R	A
CE _{L1} + CE _{L1}	NNLR	50.4	51.4	49.5	97.5
CE _{L1} + MFC _{L2}	NNLR	54.5	60.2	49.8	97.9

V. CONCLUSION AND FUTURE WORK

We have shown the effectiveness of a transfer learning approach to GOP based mispronunciation detection when applying MFC training to DNN-HMM based acoustic models. The MFC training is achieved by optimizing only the softmax layer rather than the whole DNN parameters, i.e., the hidden layers for mispronunciation detection are borrowed from speech recognition DNN the final softmax layer is task-dependent. Thus, GOP based mispronunciation detection can be cased into a two-stage based transfer learning framework: The first stage uses hidden layer(s) to extract senone discriminating features and the second approximates human standard using a trainable softmax layer. Results have shown the DNN composed by hidden layers borrowed from native speech recognizer and a MFC trained softmax layer yields F1-score improvement over the DNN with all the layers trained by the CE objective function, and obtain better detection results and robustness to over-training than DNN of which all the layers are trained by MFC.

We carried out thorough investigations into the use of transfer learning GOP based mispronunciation detection. The investigation was conducted by experimenting with different mispronunciation detection architectures proposed in literature that make use of different acoustic models (GMM-HMM, DNN-HMM) trained by different criteria. For the hidden layer(s) in the first stage, we found it is preferable to use frame-level CE minimization as the objective function to train the model parameters. Better frame-level senone classification accuracy on native speech data leads to better mispronunciation detection results. For the softmax layer, it is preferable to use MFC training to approximate human standard on a imbalanced training dataset. We further investigated into NNLR classifier based mispronunciation detection using DNN based acoustic models. It is shown mispronunciation detection features computed from the DNN with MFC trained softmax layer benefits both the GOP and the NNLR classifier based mispronunciation detection. These results have shown MFC training of the DNN filled the performance gap between the GOP and the classifier based mispronunciation detection frameworks.

Compared with the inter-rater annotation results on the L2 database and mispronunciation detection performance achieved by other DNN based systems reported in Refs. 11, 12, 30, and 31, the best detection result from transfer learning based GOP method presented in this work (53.6% in F1-score) is fairly low and needs to be further improved in the future work. We have also seen the Uighur learners of Mandarin are more likely to produce tonal errors. How to build a tonal error detector to facilitate the learning of Mandarin is a subject for future work.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61365005, 61663044, and 61761041 and the Funds for Creative Research Groups of Higher Education in Xinjiang.

- ¹C. Cucchiaroni, H. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, Santa Barbara, CA (IEEE, New York, 1997), pp. 622–629.
- ²K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Commun.* **51**(10), 883–895 (2009).
- ³D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Comput. Speech Lang.* **25**(2), 282–306 (2011).
- ⁴W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Proceeding of Interspeech*, Lyon, France [International Speech Communication Association (ISCA), Baixas, France, 2013], pp. 1886–1890.
- ⁵A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proceedings of Interspeech*, Singapore (ISCA, Baixas, France, 2014), pp. 1468–1472.
- ⁶J. Zhang, C. Huang, M. Chu, F. K. Soong, and W. Ye, "Generalized segment posterior probability for automatic Mandarin pronunciation evaluation," in *Proceedings of ICASSP*, Hawaii (IEEE, New York, 2007), pp. 201–204.
- ⁷H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proceedings of Eurospeech*, Budapest, Hungary (ISCA, Baixas, France, 1999), pp. 851–854.
- ⁸S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.* **30**(2-3), 95–108 (2000).
- ⁹F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *Proceedings of ICASSP*, Las Vegas (IEEE, New York, 2008), pp. 2077–2080.
- ¹⁰S. Wei, G. Hu, Y. Hu, and R. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Commun.* **51**, 896–905 (2009).
- ¹¹W. Hu, Y. Qian, and F. K. Soong, "A new neural network based logistic regression classifier for improving mispronunciation detection of L2 language learners," in *The 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Singapore (IEEE, New York, 2014), pp. 245–249.
- ¹²W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based regression classifiers," *Speech Commun.* **67**, 154–166 (2015).
- ¹³T. Kawahara, M. Dantsuji, and Y. Tsubota, "Practical use of English pronunciation system for Japanese students in the CALL classroom," in *Proceedings of Interspeech*, Jeju, Korea (ISCA, Baixas, France, 2004), pp. 1689–1692.
- ¹⁴D. Luo, X. Yang, and L. Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus," in *Proceedings of Interspeech*, Florence, Italy (ISCA, Baixas, France, 2011), pp. 1593–1596.
- ¹⁵T. Cincarek, R. Gruhn, C. Gruhn, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Comput. Speech Lang.* **23**(1), 65–88 (2009).
- ¹⁶S. Robertson, C. Munteanu, and G. Penn, "Pronunciation error detection for new language learners," in *Proceedings of Interspeech*, San Francisco (ISCA, Baixas, France, 2016), pp. 2691–2695.
- ¹⁷S. Ronanki, B. Li, and J. Salsman, "Automatic pronunciation evaluation and mispronunciation detection using CMU Sphinx," in *Proceedings of the Workshop on Speech and Language Processing Tools in Education, 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India [Association for Computational Linguistics (ACL), Stroudsburg, VA, 2012], pp. 61–68.
- ¹⁸X. 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 *Proceedings of Interspeech*, Portland, OR (ISCA, Baixas, France, 2012), pp. 775–778.
- ¹⁹X. Qian, H. Meng, and F. K. Soong, "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training," *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(6), 1020–1028 (2016).
- ²⁰B. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Sign. Process.* **40**(1), 3043–3054 (1992).
- ²¹L. Bahl, P. Brown, P. Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings of ICASSP*, Tokyo, Japan (IEEE, New York, 1986), Vol. 11, pp. 49–52.
- ²²P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Comput. Speech Lang.* **16**(1), 25–47 (2002).
- ²³D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of ICASSP*, Orlando (IEEE, New York, 2002), pp. 105–108.
- ²⁴D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. thesis, Cambridge University, Cambridge, UK (2004).
- ²⁵H. Huang, J. Wang, and H. Abudureyimu, "Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning," in *Proceedings of Interspeech*, Portland, OR (ISCA, Baixas, France, 2012), pp. 815–818.
- ²⁶H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**, 787–797 (2015).
- ²⁷G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Sign. Process. Mag.* **29**, 82–97 (2012).
- ²⁸W. Hu, Y. Qian, and F. K. Soong, "A DNN-based acoustic modeling of tonal language and its application to mandarin pronunciation training," in *Proceedings of ICASSP*, Florence, Italy (ISCA, Baixas, France, 2014), pp. 3206–3210.
- ²⁹J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Commun.* **73**, 14–27 (2015).
- ³⁰A. Lee, Y. Zhang, and J. R. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *Proceedings of ICASSP*, Vancouver, Canada (IEEE, New York, 2013), pp. 8227–8231.
- ³¹Y. C. Hsu, M. H. Yang, H. T. Hung, and B. Chen, "Mispronunciation detection leveraging maximum performance criterion training of acoustic models and decision functions," in *Proceedings of Interspeech*, San Francisco (ISCA, Baixas, France, 2016), pp. 1886–1890.
- ³²B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proceedings of ICASSP*, Taipei, Taiwan (IEEE, New York, 2009), pp. 3761–3764.
- ³³X. Yang, A. Loukina, and K. Evanini, "Machine learning approaches to improving pronunciation error detection on an imbalanced corpus," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, South Lake Tahoe (IEEE, New York, 2014), pp. 300–305.
- ³⁴X. Zhu and I. Davidson, *Knowledge discovery and data mining: Challenges and realities* (Information Science Reference, Hershey, PA, 2007), p. 118.
- ³⁵J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.* **20**(1), 37–46 (1960).
- ³⁶I. Pillai, G. Fumera, and F. Roli, "Designing multi-label classifiers that maximize F-measures: State of the art," *Pattern Recog.* **61**, 394–404 (2017).
- ³⁷D. D. Lewis, "Evaluating and optimizing autonomous text classification systems," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, USA [Association for Computing Machinery (ACM), New York, 1995], pp. 246–254.
- ³⁸T. Joachims, "A support vector method for multivariate performance measures," in *Proceedings of ICML*, Bonn, Germany (ACM, New York, 2005), pp. 377–384.
- ³⁹M. Jansche, "Maximum expected F-measure training of logistic regression models," in *Proceedings of HLT/EMNLP*, Vancouver, Canada (ACL, Stroudsburg, VA, 2005), pp. 692–699.
- ⁴⁰S. Wei, H. Wang, Q. Liu, and R. Wang, "CDF-matching for automatic tone error detection in Mandarin CALL system," in *Proceedings of ICASSP*, Hawaii (IEEE, New York, 2007), pp. 205–208.
- ⁴¹J. Cheng, "Automatic tone assessment of non-native Mandarin speakers," in *Proceedings of the Interspeech*, Lyon, France (ISCA, Baixas, France, 2013), pp. 1299–1302.
- ⁴²D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU* (IEEE, New York, 2011), pp. 7304–7308.