Phonological Feature Based Mispronunciation Detection and Diagnosis using Multi-Task DNNs and Active Learning

Vipul Arora[1]*, Aditi Lahiri*[1]*, Henning Reetz*[2]

[1]Faculty of Linguistics, Philology and Phonetics, University of Oxford, UK
[2]Goethe University, Frankfurt am Main, Germany

{vipul.arora, aditi.lahiri}@ling-phil.ox.ac.uk, reetz@em.uni-frankfurt.de

Abstract

This paper presents a phonological feature based computer aided pronunciation training system for the learners of a new language (L2). Phonological features allow analysing the learners' mispronunciations systematically and rendering the feedback more effectively. The proposed acoustic model consists of a multi-task deep neural network, which uses a shared representation for estimating the phonological features and HMM state probabilities. Moreover, an active learning based scheme is proposed to efficiently deal with the cost of annotation, which is done by expert teachers, by selecting the most informative samples for annotation. Experimental evaluations are carried out for German and Italian native-speakers speaking English. For mispronunciation detection, the proposed feature-based system outperforms conventional GOP measure and classifier based methods, while providing more detailed diagnosis. Evaluations also demonstrate the advantage of active learning based sampling over random sampling.

Index Terms: computer-aided pronunciation training, phonological features, multi-task DNNs, active learning

1. Introduction

Pronunciation training is an important component in language learning. Computer-aided pronunciation training (CAPT) systems aim to detect mispronunciations in the speech of the learners of a new language (L2). A learner is given a sentence (from L2) to read; wherever his/her pronunciation deviates from the canonical pronunciation (as perceived by native speakers of L2), it is marked as a mispronunciation. Further, CAPT systems provide a feedback to help the learner in correcting those mispronunciations.

Various methods have been invented to detect mispronunciations. CAPT systems are generally built on the principles of automatic speech recognition systems, where the speech is modelled using dynamic models like hidden Markov models (HMMs) and finite state transducers (FSTs). Since the words read are known beforehand, mispronunciations can be detected using the likelihood of actual pronunciation. A measure commonly used to detect mispronunciations is the goodness-of-pronunciation (GOP) measure [1, 2]. Another method is based on mispronunciation classifiers trained for each phoneme [3, 4]. Yet another approach, known as extended recognition networks, is to modify the FSTs to detect mispronounced phonemes during decoding itself [5, 6].

For mispronunciation diagnosis, various kinds of feedback schemes have been proposed by different researchers [7]. The most common way is to specify the phoneme uttered in place of the canonical one [8, 9]. But as the underlying computational models are prone to errors, it is much more useful to construct explanatory models that allow a deeper analysis of the pronunciation in addition to giving a concise message to the learner. Another problem here is that many times the uttered phoneme might not correspond to any phoneme used in the decoder, thereby limiting the diagnosis. Moreover, often phoneme based feedback does not help a learner, as (s)he might not even perceive the L2 phoneme if it is not there in his/her L1 (native language). Towards this end, Li *et al.* [10] use subphonetic attributes, derived from acoustic and phonetic properties of speech, for rendering the feedback. They use GOP measure to detect mispronunciations at the attribute level, and then merge the subsequent results using neural networks to detect phone level mispronunciations.

In this paper, we propose the use of phonological features for CAPT. The utility of phonological features for speech analysis has been long known [11]. They are also useful for analysing error patterns in mispronunciations [12]. For CAPT, the use of broad phonetic features [13] and phonemic distinctions [14] has been considered by some works. Feedback in terms of phonological features is easily comprehensible to the learners and is effective in learning. It provides information on how to move the articulators in order to produce the target sounds. This paper presents two major contributions - firstly, the deep neural network (DNN) based acoustic models are improved with the help of multi-task learning, and secondly, the cost of transcription in terms of human supervision is reduced with the help of active learning.

The primary task of an acoustic model is to estimate the phoneme state probabilities for HMM based decoders. The secondary task in our proposed model is to estimate the phonological features from speech. Earlier models [10, 15] perform the two tasks serially. However, multi-task DNN learning [16] uses a shared representation, in terms of shared hidden layers, for both the tasks. It is found to improve the system performance for many applications like phoneme recognition [17], low-resource speech recognition [18], cross-language adaptation of the acoustic models [19], speech synthesis [20], object detection in images [21], etc.

Another challenge in training the CAPT systems is that they need phoneme-level transcribed data. On the other hand, present ASR systems mostly use word-level transcribed datasets. The latter datasets are now available much more abundantly than the former. Moreover, the labour needed from human annotators (expert teachers) is quite intense. In this paper, we also propose an active learning algorithm for training a CAPT system. With a large number of unannotated speech files at hand, active learning [22] involves selecting the most informative set of files to be labelled by human. Thus, a large improvement is obtained from annotating a smaller number of files. Active learning has been helpful for many machine learn-

Table 1: *Phonological features along with the corresponding phonemes*

VOC	AA AE AH AO AW AX AY EH ER EY IH IY OH OW OY UH UW
CONS	B CH D DH F G HH JH K P S SH T TH V Z ZH L M N NG R
CONT	DH F HH L S SH TH V Z ZH
OBSTR	B CH D DH F G JH K P HH S SH T TH V Z ZH
STR	CH S SH TH Z ZH
VOICE	B D DH G JH V Z ZH
SON	AA AE AH AO AW AX AY EH ER EY IH IY L M N NG OH OW OY R UH UW W Y
STOP	B CH D G JH K P T
LOW	AA AE AW AY
HIGH	CH IH IY JH SH UH UW W Y ZH
LAB	AO B F M OH OW OY P UH UW V W
COR	AE CH D DH EH EY IH IY JH L N R S SH T TH Y Z ZH
DOR	AA AO AW AY G K NG OH OW OY UH UW W
RTR	AH AX EH ER IH UH W
NAS	M N NG
LAT	L
RHO	ER R
RAD	HH

ing problems including ASR [23] and spoken language understanding [24].

2. Proposed system

This section describes the proposed CAPT system with English as L2.

2.1. Phonological Features

The present system uses 18 phonological features that characterise all the phonemes of English. These features are VOC (vowel), CONS (consonant), CONT (continuant fricative consonant), OBSTR (obstruant consonant), STR (strident), VOICE (voiced consonant), SON (sonorant), STOP (stop consonant), LOW, HIGH, LAB (labial), COR (coronal), DOR (dorsal), RTR (retracted tongue root), NAS (nasal), LAT (lateral), RHO (rhotic), RAD (radical). In addition, a silence marker (SIL) is used. Table 1 lists all the phonological features along with the corresponding phonemes (from the UK phoneme set used by ISLE dataset) that they characterise.

2.2. Dataset

For implementing the proposed system we use Interactive Spoken Language Education (ISLE) dataset [25]. It consists of 23 German and 23 Italian native speakers speaking English as L2. A speaker is given a sentence to read and his/her utterance is recorded. Apart from the word-level transcription, each utterance has two kinds of transcriptions - i) the canonical phonemes that the speaker should utter, and ii) the actual phonemes that the speaker actually utters.

2.3. Acoustic Model

The proposed system is based on DNN-HMM framework [26]. Speech is modelled as a sequence of phonemes. Each phoneme is modelled with 3 states of an HMM (monophone). The DNN acoustic model estimates two kinds of outputs - i) the posterior probabilities of HMM states, and ii) the probabilities of phonological features.

The input to the DNN acoustic model are the acoustic parameters, i.e., 23 Mel-scaled filter bank log energies with a context of ± 5 frames. The input is normalised to have zero mean and unit variance. There are 3 shared hidden layers, each with 512 neurons and ReLU (rectified linear unit) non-linearity. The first output layer consists of 150 neurons corresponding to the monophone states of HMMs (canonical transcriptions have 41 English phonemes and a silence phoneme; the actual transcriptions contain 8 extra phonemes that are not valid English phonemes). The layer has softmax non-linearity. The outputs are the posterior probabilities of states and are converted into likelihoods, by dividing with prior probabilities, so as to be used by the HMM. The second output layer of DNN consists of 19 neurons corresponding to the phonological features. Each neuron, having sigmoidal non-linearity, estimates the probability of the corresponding feature.

The speech files in training data are force-aligned with the corresponding actual transcriptions with the help of a 3-state GMM-HMM based phoneme recogniser, using 39 dimensional MFCC$+\Delta+\Delta\Delta$ acoustic parameters. The target values for the first output layer are constructed directly from the alignments, while those for the second output layer are constructed as in [15], by marking the absence or presence of a phonological feature at a particular time frame with 0 or 1, respectively.

The DNN is trained to minimise an objective function, which is a sum of the objective functions for the two output layers. The first output layer's objective function is the categorical cross entropy while that of the second output layer is squared error. The DNN weights are updated using AdaGrad [27] with Nesterov momentum over mini-batches of size increasing linearly from 256 to 1024 with each epoch. The output classes are balanced by applying weights to the objective function at each output neuron. A dropout rate of 10% is used for each hidden layer during training.

2.4. Mispronunciation Detection

When a phoneme in actual transcription does not match the one in canonical transcription, it is said to be mispronounced. The outputs of the acoustic model are used to detect mispronunciations, with the help of a neural network classifier.

During training, the speech files are force-aligned with the canonical transcriptions. The average values of phonological feature probabilities for a phoneme segment are used as the input to the mispronunciation classifier (phoneme-level). There is one hidden layer with 512 neurons and ReLU non-linearity. The output layer consists of 41 units corresponding to each canonical phoneme. Each neuron, having sigmoidal non-linearity, estimates the probability of correct pronunciation of the corresponding phoneme. The ground truth mispronunciations are obtained by aligning the canonical transcriptions with actual transcriptions, using Levenshtein distance. The target value of the output neuron corresponding to the canonical phoneme is set to 0 or 1 to mark if it is mispronounced or not, respectively; the target values of other neurons are left unspecified, i.e. they are not used for updating the network weights. In this work, we consider only substitution and deletion errors. Insertion errors are not considered for training and evaluation, but the FSTs are allowed to take care of them while decoding by introducing an optional phoneme at word boundaries. The training is carried out to minimise the squared error. The weight update scheme

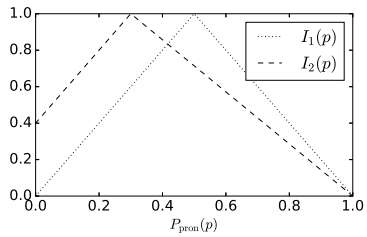

Figure 1: *Informativeness score $I(p)$ for phoneme p as a function of pronunciation score $P_{pron}(p)$*

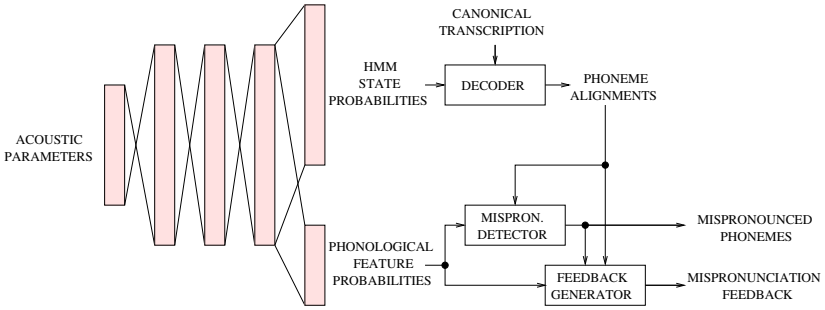

Figure 2: *Block schematic of the proposed CAPT system*

is same as used in Section 2.3, i.e., using AdaGrad, Nesterov momentum, data balancing and dropouts.

2.5. Mispronunciation Diagnosis

Mispronunciation is diagnosed in terms of phonological features. The average values of the phonological feature probabilities estimated by the acoustic model for a mispronounced canonical phoneme are compared with the canonical feature values for the same phoneme. The feature showing maximum deviation from its canonical value is used for constructing the feedback. The sign of deviation tells if the feature needs to be increased or decreased. For example, when *bag* is pronounced as *back*, the feature VOICE needs to be increased for the canonical phoneme /g/ to make it sound like the first sound in *goose*. Or, when *facing* is pronounced like *faking*, the system will ask the user to make the medial consonant with a hissing sound (STRIDENT feature) as the /s/ in *kissing*.

2.6. Informativeness Score for Active Learning

For the active learning algorithm, an informativeness score is computed based on the mispronunciation detection performance at the phoneme level. Phonemes that lie closer to the mispronunciation threshold are considered more informative than those away from the threshold. The informativeness score $I(p)$ for a canonical phoneme p is estimated from the probability of correct pronunciation $P_{\text{pron}}(p)$, which is the output of mispronunciation DNN, with the help of a triangular function. Two functions $I_1(p)$ and $I_2(p)$ have been proposed here as shown in Fig. 1. The utterance level informativeness score is the average of the informativeness scores of all its phonemes. The utterance files with the highest informative scores can be transcribed by human experts, and then be used for training.

A block schematic of the overall system is shown in Fig. 2.

3. Experiments

The dataset is divided into training and test sets, consisting of 19 and 4 speakers, respectively, from each language (German and Italian). The total audio duration of training set is 8 hours 24 minutes, while that of test set is 1 hour 34 minutes.

3.1. Evaluation of Multi-task DNN System

The proposed system is compared with two state-of-the-art methods [4] for detecting mispronunciations. One uses the goodness of pronunciation (GOP) measure, which is defined for

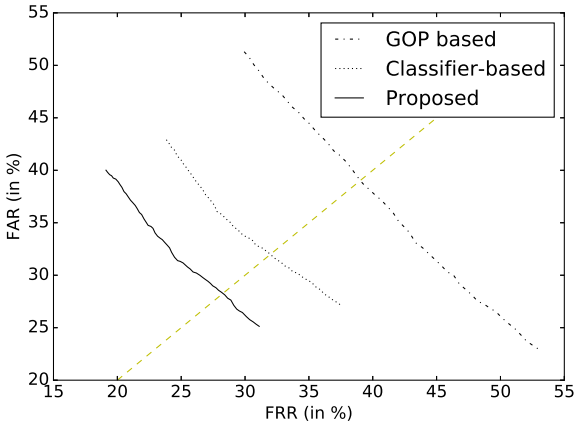

Figure 3: *Mispronunciation detection: False acceptance rate (FAR) vs false rejection rate (FRR) for different methods. The 45° dashed line shows where FAR=FRR.*

a canonical phoneme p as

$$\text{GOP}(p) = \log \frac{P(p)}{\max_{q \neq p} P(q)}$$

where, $\log P(q)$ is the average log posterior probability of phoneme q over the frames aligned with p. A DNN acoustic model has been used here to estimate the state probabilities for phoneme q; the maximum state probability is assigned as the phoneme probability. A low value of GOP indicates mispronunciation. The GOP scores for all the phonemes are scaled so as to have similar thresholds, close to 0.5.

The other baseline method uses a classifier based method without phonological features. The state log-probabilities, estimated by a DNN that maps input acoustic parameters to HMM phoneme states, are averaged for each phoneme q to get $\log P(q)$. The values $\log P(q), \forall q$, are fed into a neural network classifier to estimate $P_{\text{pron}}(p)$.

The performance of the system for mispronunciation detection is measured using false acceptance rate (FAR) and false rejection rate (FRR). FAR is the ratio of number of mispronounced phonemes falsely accepted by the system as correct to the total number of mispronounced phonemes. FRR is the ratio of number of correctly uttered phonemes falsely rejected as mispronounced to the total number of correctly pronounced phonemes. Since the two vary in opposite directions with the threshold, for easy comparison, we set the threshold to make

Table 2: *Mispronunciation detection: Equal error rate (EER) for different methods.*

	GOP-based	Classifier based	Proposed
EER (in %)	39.0	31.8	**28.3**

Table 3: *Feedback evaluation results for some of the most frequently mispronounced phonemes. Freq denotes the number of mispronunciations of the phoneme in the test set. Acc denotes feedback accuracy in %ge.*

Phoneme	Freq	Acc (%)	Phoneme	Freq	Acc (%)
R	171	98.8	AX	352	90.9
T	218	97.2	DH	185	87.0
EY	132	96.2	IH	250	79.2
AE	156	96.2	EH	107	60.7
IY	92	95.7	Z	105	60.0

the two equal, i.e., FAR=FRR:=EER (equal error rate). Fig. 3 shows the FAR vs FRR curves for the different mispronunciation detection methods. The EER attained by the different methods is shown in Table 2. The proposed system achieves a relative improvement of 27.4% over the GOP-based system and that of 11.0% over the classifier based system. This improvement could be attributed to the shared representation using multi-task DNNs, where both the tasks benefit from each other.

The phonological feature based feedback is evaluated using accuracy, which is defined as the number of mispronounced phonemes for which a correct feature (with a correct sign of deviation) is given as feedback divided by the total number of mispronounced phonemes. We evaluate feedback for all mispronunciations, independent of the detection performance. The proposed system attains a feedback accuracy of 81.1% averaged over all the phonemes (i.e., 2794 mispronunciations). Table 3 shows the feedback accuracy for some of the most frequently mispronounced phonemes in the test set. It is to be noted that a mispronunciation may entail mistakes in one or more features; catching any one of them is considered correct.

3.2. Evaluation of Active Learning

The entire training set is assumed to be unlabelled first. Let $\mathcal{S}_u$ denote the set of unlabelled utterances. Randomly, 36% of data is sampled, annotated and used to train the system. Let $\mathcal{S}_t$ denote the set of labelled utterances for training the system. The system so obtained is used to estimate informativeness scores for the utterances in $\mathcal{S}_u$. The most informative utterances from $\mathcal{S}_u$ are annotated, are transferred to $\mathcal{S}_t$ and the system is retrained. Thus, $\mathcal{S}_u$ is sampled incrementally in steps until it is empty.

The mispronunciation detection EER obtained by informativeness score based sampling is compared with that obtained by random sampling as baseline. The two schemes are employed to train the proposed system. The learning curves for the two schemes are shown in Fig. 4. We can see that the proposed $I(p)$ based sampling is efficacious in achieving lower error rates while reducing the amount of labelled data needed for training the system. The random and $I(p)$ based sampling curves meet at the same point when all the training data is used. We find that the performance with $I_2(p)$ deteriorates slightly when all the training set is used as compared to when 92% of it is used,

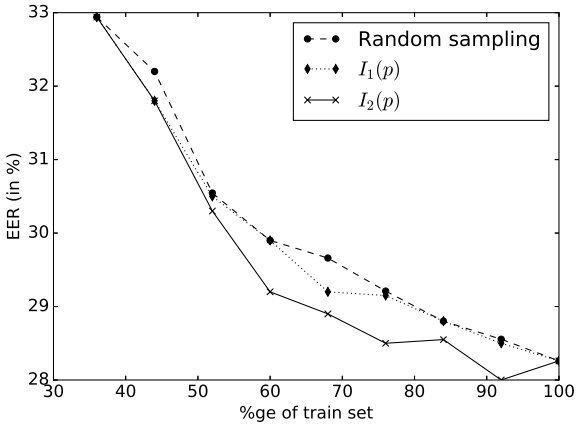

Figure 4: *Learning curve for mispronunciation detection equal error rate (EER), with respect to the percentage of training set used for training*

showing that $I_2(p)$ based sampling is able to reject the outlier data, which degrades the model.

The function $I_2(p)$ (refer to Fig. 1) gives more weight to mispronounced phonemes (i.e., with smaller P_{pron}) than the correctly pronounced ones, as compared to those given by $I_1(p)$. We find that $I_2(p)$ improves the performance more as compared to the symmetric $I_1(p)$. This might be due to the reason that the mispronounced phonemes occur quite scantily as compared to the correctly pronounced ones. Hence, giving more weight to mispronunciations incorporates more mispronounced phonemes into training, thereby, improving the performance.

4. Conclusion

This paper presents a computer aided pronunciation training (CAPT) system using phonological features. The main contributions are: i) a phonological feature based diagnosis system is presented to construct a corrective feedback for each mispronounced phoneme for effective learning; ii) a multi-task DNN based acoustic model is presented to estimate the HMM state probabilities and phonological feature probabilities using a shared representation; iii) an active learning based approach is presented to train the system with minimal annotated data. The proposed system outperforms the state of the art methods for detecting mispronunciations, attaining more than 11% relative improvement in equal error rate. Also, the proposed system can be used for any pair of L1 and L2. Future work includes incorporating native speaker data for training over canonical pronunciations. We hope that this will improve the robustness of the system to any L1. We would like to test the system for teaching other languages (L2) as well.

5. Acknowledgements

This research was supported by the ERC Proof of Concept FLEX-SR award no. 632226.

6. References

[1] S. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[2] Y. C. Hsu, M. H. Yang, H. T. Hung, and B. Chen, "Mispronuncia-

tion detection leveraging maximum performance criterion training of acoustic models and decision functions," in *INTERSPEECH*, 2016, pp. 2646–2650.

[3] J. van Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in dutch," *The Journal of the Acoustical Society of America (JASA)*, vol. 134, no. 2, pp. 1336–1347, 2013.

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

[5] W. K. Lo, S. Zhang, and H. M. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system." in *INTERSPEECH*, 2010, pp. 765–768.

[6] A. Lee and J. R. Glass, "Mispronunciation detection without non-native training data." in *INTERSPEECH*, 2015, pp. 643–647.

[7] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in Computer Assited Pronunciation Training: technology push or demand pull?" in *INTERSPEECH*, 2002.

[8] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[9] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 564–579, 2015.

[10] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6135–6139.

[11] R. Jakobson, C. G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The distinctive features and their correlates.* Acoustics Laboratory MIT, 1951.

[12] L. Ward, A. Stefani, D. Smith, A. Duenser, J. Freyne, B. Dodd, and A. Morgan, "Automated screening of speech development issues in children by identifying phonological error patterns," in *INTERSPEECH*, 2016, pp. 2661–2665.

[13] M. Kane, J. P. Cabral, A. Zahra, and J. Carson-Berndsen, "Introducing difficulty-levels in pronunciation learning." in *Workshop on Speech and Language Technology in Education (SLaTE)*. Citeseer, 2011, pp. 37–40.

[14] J. C. Koreman, P. Wik, O. Husby, and E. Albertsen, "Universal contrastive analysis as a learning principle in CAPT," in *Workshop on Speech and Language Technology in Education (SLaTE)*, 2013, pp. 172–177.

[15] V. Arora, A. Lahiri, and H. Reetz, "Attribute based shared hidden layers for cross-language knowledge transfer," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 617–623.

[16] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[17] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6965–6969.

[18] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5592–5596.

[19] P. Bell, J. Driesen, and S. Renals, "Cross-lingual adaptation with multi-task adaptive networks." in *INTERSPEECH*, 2014, pp. 21–25.

[20] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4460–4464.

[21] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[22] B. Settles, "Active learning literature survey," University of Wisconsin, Madison, Computer Sciences Technical Report 1648, 2010.

[23] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE transactions on speech and audio processing*, vol. 13, no. 4, pp. 504–511, 2005.

[24] H.-K. J. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding." in *INTERSPEECH*, 2005, pp. 437–440.

[25] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The isle corpus of non-native spoken english," in *Language Resources and Evaluation Conference (LREC)*. European Language Resources Association, 2000, pp. 957–964.

[26] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2014.

[27] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.