INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multi-Task Learning for Mispronunciation Detection on Singapore Children's Mandarin Speech

Rong Tong, Nancy F. Chen and Bin Ma

Institute for Infocomm Research, Singapore
{tongrong,nfychen,mabin}@i2r.a-star.edu.sg

Abstract

Speech technology for children is more challenging than for adults, because there is a lack of children's speech corpora. Moreover, there is higher heterogeneity in children's speech due to variability in anatomy across age and gender, larger variance in speaking rate and vocal effort, and immature command of word usage, grammar, and linguistic structure. Speech productions from Singapore children possess even more variability due to the multilingual environment in the city-state, causing inter-influences from Chinese languages (e.g., Hokkien and Mandarin), English dialects (e.g., American and British), and Indian languages (e.g., Hindi and Tamil). In this paper, we show that acoustic modeling of children's speech can leverage on a larger set of adult data. We compare two data augmentation approaches for children's acoustic modeling. The first approach disregards the child and adult categories and consolidates the two datasets together as one entire set. The second approach is multi-task learning: during training the acoustic characteristics of adults and children are jointly learned through shared hidden layers of the deep neural network, yet they still retain their respective targets using two distinct softmax layers. We empirically show that the multi-task learning approach outperforms the baseline in both speech recognition and computer-assisted pronunciation training.

Index Terms: automatic speech recognition (ASR), multi-task learning (MTL), human-computer interaction (HCI), computer-assisted pronunciation training (CAPT), computer-assisted language learning (CALL)

1. Introduction

It is conventional wisdom that in general learning a language at a young age is one of the primary factors for one's proficiency in language acquisition. Language learning is an important subject for primary to college level students in multi-lingual counties like Singapore. A computer-assisted language learning (CALL) system is specially useful for students in such multilingual environment to provide self-paced pronunciation training.

Compared to adults' speech, it is more challenging to conduct speech processing research on children's speech. One primary reason is the lack of children's linguistic corpora. Data collection on children's speech is much more difficult than that of adults, as children usually have limited attention span and vocabulary, especially those at young ages. Another reason is the high acoustic and linguistic variabilities in children's speech: acoustic characteristics varies across age and gender [1, 2], there are larger variances in speaking rate and vocal effort [3] in children's speech, and children have immature command of word usage, grammar, and linguistic structure [4].

Various acoustic, prosodic and linguistic features are explored for children's speech processing studies. An automatic scoring system [5] for speech of middle school students was developed by using pronunciation, prosody, lexical and content features. Features obtained from multiple aspects are utilized in [6] on an automatic system to assess the proficiency of non-native children's speech from age 8 and above. Children's speech are collected in wide-band to capture high frequency signal [7]. Feature warping functions [8, 9] are explored to improve the performance of children's speech recognition.

Acoustic modeling is one of the most important components in speech technology. Acoustic analyses on children's speech [2] are conducted to study the characteristics of children's speech production for children's speech recognition. Deep neural network training is widely adopted in acoustic modelling. A convolutional, fully connected LSTM model structure is shown to outperform a standard LSTM model on children's speech recognition [4]. Phoneme specific discriminative classifiers and DNN derived bottle-neck features are utilized in pronunciation assessment [10]. Multi-distribution DNN is used for mispronunciation detection and diagnosis in English [11]. Transfer learning based logistic regression classifiers are adopted in [12] for mispronunciation detection.

To take advantage of the existing adults' speech data for children's acoustic modeling, vocal tract length normalization (VTLN) is applied on adults' data to compensate for the vocal tract differences between adult and children [13, 14, 15]. A stochastic feature mapping method is proposed to transform out-of-domain adults data for children's speech recognition [16].

In this paper, we compare two data augmentation approaches for non-native children's Mandarin mispronunciation detection. The first approach is a straightforward data combination method, it disregards the child and adult categories and consolidates the two datasets together as one entire set for children's acoustic modeling.

The second approach is multi-task learning (MTL) [17] using deep neural networks (DNN). Multi-task learning has been successfully adopted in various speech applications, like multilingual speech recognition [18, 19, 20], information retrieval [21], speech synthesis [22] and spoken language understanding [23]. The key concept of MTL is to train a shared model that performs well on different tasks. Optimizing for separate objective functions at the same time can be viewed as an effective form of regularization, preventing the jointly trained model from overfitting, thus increasing the generalization power of the model. Specificity in this work, the acoustic characteristics of adults and children are jointly learned through shared hidden layers of a deep neural network, yet they still retain their respective targets using two distinct softmax layers.

http://dx.doi.org/10.21437/Interspeech.2017-520

2. Adults and children's Mandarin speech

2.1. Mandarin Chinese Background

Mandarin Chinese is a monosyllabic language, where each Chinese character is a single syllable. Pinyin is one of the most widely adopted romanization format of Chinese characters. For example, a Pinyin *SHAN1* consists of an initial (*SH*), a final (*AN*) and a tone (*1*). There are five tones in Mandarin: Tone 1 to Tone 5, in which Tone 5 is neutral and has no specific pitch contour, it is analogous to an unstressed syllable in English. Since Tone 5 is not well-defined, in this work, we focus on the mispronunciation detection of Tone 1 to Tone 4.

Non-native Mandarin speakers may make different types of mistakes, ranging from phonetic [24], to lexical tone [25], and to fluency aspects [26]. In this paper, we focus on the detection of the phonetic and tonal mistakes in children's Mandarin.

2.2. Mandarin production variations across age and language background

Differences between children's and adults' speech can be characterized from two aspects: acoustic and linguistic. The acoustic differences are attributed to children's physical development (shorter vocal tract, smaller size in vocal folds and tongue), thus children have higher pitch and formant frequencies than those of adults. Linguistically, since children are less linguistically developed than adults, they often have a more limited vocabulary and are more likely to mispronounce certain phonemes even in their native first language.

Speech productions of Singapore children [1] possess even more variability due to the multilingual environment in the city-state, causing inter-influences from Chinese languages (e.g., Hokkien and Cantonese), English dialects (e.g., American and British), Malay languages (e.g., Bahasa Melayu and Bahasa Indonesia) and Indian languages (e.g., Hindi and Tamil). For example, retroflex fricatives and affricates (e.g., /SH/ in Pinyin) in Mandarin might be influenced from Hokkien, where the retroflex place of articulation is not phonemic.

To better visualize the pronunciation variability across different subjects depending on age and language background (monolingual or multilingual), Figure 1 compares the posterior probability histogram of /SH/ on *Baseline-adult* model (Section 5.2) pronounced by speakers from the following three test sets : native adults (*adult-cn-tst*), native children (*kids-cn-tst*) and Singapore children (*kids-sg-tst*) (Section 4). Compared to the posterior of /SH/ produced by native adults, higher variations are observed in native children's speech. The posterior variations are even larger in Singapore children's than native children's. This corresponds with the documentation of Singapore children making the most mistakes in pronouncing the retroflex fricative /SH/ in Mandarin [1].

3. Multi-task learning for children's language learning

The concept of multi-task learning (MTL) is learning multiple tasks jointly to take advantage of the common information shared among different tasks [17]. The common assumption is that the tasks are related but different, hence there are underlying common characteristics that can be shared. Various application of multi-task learning can be categorized into two groups: for the first group, each task has its unique output [18, 19]; for the second group, one task can have multiple outputs [17]. The proposed multi-task learning for children falls into the first group. The sharing of adult and children's training data helps the joint model characterize speaker-independent acoustic properties better, thus improve both acoustic modeling of adults and children. This is especially beneficial to modeling children's speech, since linguistic resources for such are limited.

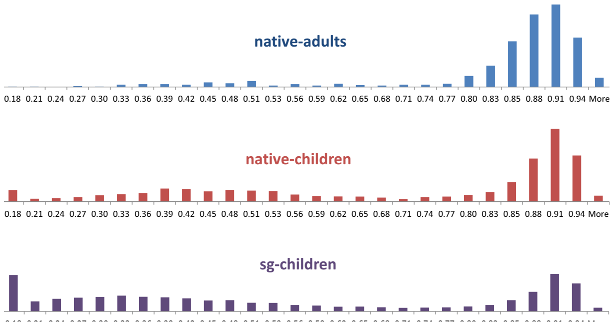

Figure 1: *Posterior probability histogram of /SH/ pronounced by speakers in 3 test sets: mainland China adults (native-adults), mainland China children (native-children) and Singapore children (sg-children)*

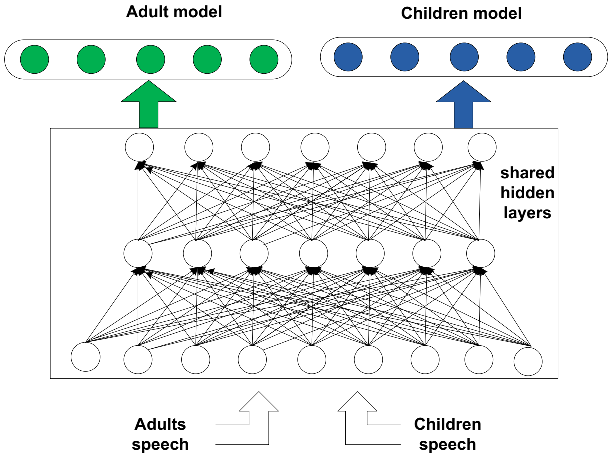

Figure 2: *Multi-task learning DNN for adults and children*

3.1. Multi-task deep neural network for acoustic modeling

Figure 2 illustrates the proposed multi-task learning deep neural network structure for adults and children's acoustic modelling. The adults' and children's speech are first consolidated together to train several lower hidden layers, then two different output layers are trained for adults and children respectively.

Given two training sets from adults and children : $D_a = \{x^a, y^a\}$ and $D_c = \{x^c, y^c\}$, where x^a and x^c are input features for adults and children respectively, y^a and y^c are the corresponding targets for adults and children respectively. Following the standard deep neural network training process, a probabilistic model with L hidden layers can be learnt jointly from the two feature sets x^c and x^a. For each hidden layer $l \in \{1..L\}$:

$$h_l(x^c, x^a) = \sigma(W_l h_{l-1}(x^c, x^a) + b_l) \qquad (1)$$

where σ is an activation function, W_l is the weight matrix for l-th hidden layer, h_{l-1} is the probabilistic model of the previous layer, and b_l is a bias vector for layer l. Note that the hidden layers are trained with all training features from both adults and children.

To capture the differences of the adults and children in acoustic modelling, a task dependent output layer is trained for each task:

$$\hat{y^a} = \text{softmax}(W_a\,\sigma(W_L + b_L))$$
$$\hat{y^c} = \text{softmax}(W_c\,\sigma(W_L + b_L)) \qquad (2)$$

where $\hat{y^a}$ and $\hat{y^c}$ are the outputs of adult and children tasks respectively, W_a and W_c are the task dependent weight matrix of adult and children respectively, W_L and b_L are the weight matrix and the bias vector of the shared hidden layers.

During DNN training, the goal of training is to minimize the global cost ϵ, which is a linear combination of the costs of each task:

$$\epsilon = \lambda_a \epsilon_a + \lambda_c \epsilon_c \qquad (3)$$

where ϵ_a and ϵ_c are the cost functions, λ_a and λ_c are the weights for the adults and children respectively.

3.2. Multi-task learning for mispronunciation detection

For mispronunciation detection on non-native children's Mandarin speech, we adopt the context aware mispronunciation detection framework [24, 27]. Under this framework, mispronunciation detection is first performed at the phonetic level, followed by the tone level, and the output of the two levels are combined in the last level to provide syllable level feedback. In both phonetic level and tone level, the context information is incorporated in the detection process.

4. Speech Corpora

4.1. Native Mandarin Corpus - adult (adult-cn)

The native Mandarin training set *adult-cn-trn* consists of data from three different corpora. A large portion of native adult data comes from King-ASR-118 mobile speech corpus.[1] The training set consists of utterances from 975 speakers recorded with various type of mobile phones.

To capture the characteristics of microphone channel and reading-style speech, two read speech Mandarin corpus are incorporated in *adult-cn-trn*. One is the HKU96 Putonghua Corpus [28], which consists of a total of 20 native Putonghua speakers, each speaking hundreds of utterances. The whole corpus is included in *adult-cn-trn*.

Another portion is the training part of THCHS-30 corpus [29]. It is a Mandarin corpus, where the speech data are recorded with microphone in clean environment. The training set consists of speech from 30 speakers and the test set has speech of 10 speakers. The test set of the THCHS30 corpus is used as the ASR test data for native adult, noted as *adult-cn-tst*.

4.2. Native Mandarin Corpus - children (kids-cn)

The native Mandarin children's corpus is recorded in mainland China. It is recorded from 256 gender balanced speakers, each reading 300 utterances. The speakers are primary school students of ages 7-12, all the speakers are from the northern China. The corpus is separated into training and test sets: the training set consists of speech from 228 speakers, it is included in the children's acoustic model training set *kids-cn-trn*. The test set *kids-cn-tst* consists of speech from 28 speakers, used to evaluate the ASR performance.

4.3. Non-native Mandarin Corpus (kids-sg)

The non-native corpus is a corpus recorded from primary school students in Singapore [1]. The corpus consists of speech recorded from 255 students aged in 7-12, each reads 330 utterances. The corpus is separated into three portions, the first set *kids-sg-trn* is for acoustic model training, it consists of speech of 203 speakers; the second set *kids-sg-dev* consists of speech from 36 speakers, used as a development set for mispronunciation detection; and the last set *kids-sg-tst* consists of speech of 12 speakers, it is used as mispronunciation detection test set. Both development and test set are combined as ASR test set for non-native children. More details about the corpus and the transcription can be found in [1].

Table 1: *Acoustic model training data*

Data	no. spks	no. utts	hrs w/sil	hrs wo/sil
adult-cn-trn	1015	387503	300	212
kids-cn-trn	228	65142	74	29
kids-sg-trn	203	61705	99	37

Table 2: *ASR test set, mispronunciation detection development (dev) and test (test) sets of non-native speakers*

Data	no. spks	no. utts	hrs w/sil	hrs wo/sil
adult-cn-tst	10	2495	6.2	4.9
kids-cn-tst	28	8107	9.2	3.5
kids-sg-dev	36	10587	23	8
kids-sg-tst	12	4077	8.8	2.9

Table 1 summarizes the acoustic model training data. The number of speakers (no. spks), number of utterances (no. utts) and audio length in hours including (hrs w/sil) and excluding (hrs wo/sil) silences are reported. Note that the children's audio length are significantly reduced when the silences are excluded, as the children's sentences are shorter than those of adults.

Table 2 shows the statistics of the ASR test set, the non-native mispronunciation detection development and test sets. In Singapore children's development and test sets, half of the speakers are aged 8-9 years and the other half are aged 10-12.

5. Experiments

5.1. Evaluation metric

The proposed two data augmentation methods are evaluated on automatic speech recognition (ASR) and non-native mispronunciation detection. The ASR performance is measured by Pinyin error rate (PER), which is the percentage of wrongly recognized Pinyin; The mispronunciation detection performance is measured by False acceptance Rate (FAR) and False Rejection Rate (FRR). FAR is the percentage of mispronounced tests that system failed to detect and FRR is the percentage of correctly pronounced tests that system erroneously detected as mispronunciation. In practical CALL applications, minimizing FRR is more important than FAR, as we do not want to discourage the learner when their non-native pronunciation is correct.

5.2. Automatic speech recognition (ASR)

5.2.1. Acoustic modeling

Two baseline acoustic models are deep neural network based acoustic models [27]. Both DNN models have 4 hidden layers and the DNNs are trained on top of a GMM-HMM model with 175 tone dependent phones and 8502 tied states. ***Baseline-adult*** is trained from the native adult Mandarin training sets *adult-cn-trn*. ***Baseline-kids*** is trained from the composite set of native and non-native children's data *kids-cn-trn* and *kids-sg-trn*.

Another two DNN models are trained by combining the adults and children data *adult-cn-trn*, *kids-cn-trn* and *kids-sg-trn* using different approaches. Both DNN models are trained from GMM-HMM models with the same number of phone and

[1] Chinese Mandarin Mobile Speech Recognition Database, http://www.speechocean.com/en-News/783.html

Table 3: *ASR results for native and non-native Mandarin adult and children speech (Pinyin unigram)*

AM /PER(%)	adult-cn-tst	kids-sg (dev+tst)	kids-cn-tst
Baseline-kids	52.22	45.91	46.17
Baseline-adult	27.57	47.38	36.25
Mixed	31.06	33.58	22.56
Multi-task	24.36	28.51	21.71

Table 4: *Singapore children's Mandarin phone and syllable mispronunciation detection, all results do not consider tone*

kids-sg	phone(%)		syllable(%)		
	FAR	FRR	FAR	FRR	Average
Baseline-kids	22.4	9.5	15.9	18.7	16.6
Mixed	21	8.2	15	15	14.8
Multi-task	20.3	7.7	13.9	14.4	14.08

states as the baseline models. To take advantage of the augmented training data, each DNN has 6 hidden layers. ***Mixed*** is trained by consolidating the adult and children sets together as one entire training set. ***Multi-task*** is trained with the proposed multi-task learning DNN, as illustrated in Fig 2.

5.2.2. *ASR results*

Table 3 reports the ASR results of the three test sets on the four aforementioned DNN models. A Pinyin unigram is used in the decoding process. The performance are reported as Pinyin error rate (PER). Comparing the two baseline acoustic models, the ***Baseline-kids*** has slightly better performance on matched test set (*kids-sg*), while the performances are much worse on the native Mandarin sets. This is reasonable as the training data for ***Baseline-kids*** is far less than that of ***Baseline-adult***.

When we consolidate all the adult and children's data in ***Mixed*** training, the ASR accuracy of both children tests are improved. This reveals that the adults speech characteristics compensate the children's. However the ASR error on *adult-cn-tst* set is increased, this might attribute to the acoustic mismatch between children's speech and adults' speech.

With the same training data but a different training approach, the ASR performance on all test sets are improved on ***Multi-task***. This validates the effectiveness of the multi-task learning, tasks compensate each other by learning from the shared information, while task specific optimization ensures the task specific characteristics are still preserved.

5.3. Mispronunciation detection

We validate the acoustic modeling performance on non-native children's Mandarin mispronunciation detection. The mispronunciation detection is conducted following the context aware multilayer framework [24].

5.3.1. *Phone and syllable mispronunciation detection*

Table 4 reports the mispronunciation detection performance of three children's acoustic models on *kid-sg-tst*. Both phone and syllable level detection results are reported. The average error rates are reported in the last column.

The mispronunciation detection errors at the phone and syllable levels are consistently improved by augmenting children's training data with adults speech. Multi-task learning methods gain more performance improvements than the straightforward data consolidation approach.

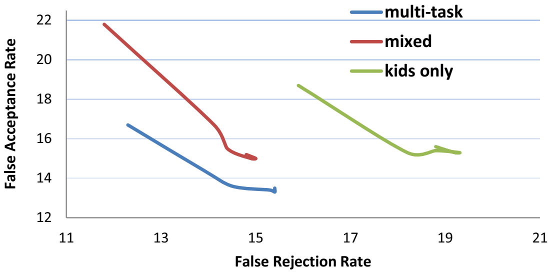

Figure 3: *Singapore children's Mandarin mispronunciation syllable error detection with different acoustic weights*

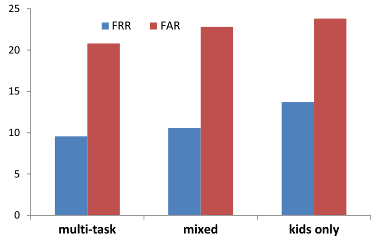

Figure 4: *Singapore children's mispronunciation detection on /SH/*

Table 5: *Tone recognition results of Singapore children's Mandarin test kids-sg-tst*

AM/ accuracy (%)	tone 1	tone 2	tone 3	tone 4
Baseline-kids	68.5	55.6	51.5	66.2
Mixed	83.1	73.6	70.0	80.3
Multi-task	83.3	75.3	72.0	81.4

Figure 3 illustrates the detection error trade-off (DET) plot of Singapore children's Mandarin syllable error detection results with different acoustic model weights during decoding process. The x-axis is the False Rejection rate and y-axis shows the False Acceptance Rate. Similar trends can be observed in Table 4. The DET curve of the multi-task approach is the closest to the bottom left corner, indicating higher performance of the Multi-task learning on mispronunciation than the straightforward data consolidation approach.

Figure 4 shows the mispronunciation detection of /SH/ on the three children's models. The proposed multi-task training method achieved the best performance for mispronunciation detection of /SH/, indicating that the multi-task learning approach works better on modelling the inter-speaker variability.

5.3.2. *Tone recognition*

Table 5 shows tone recognition results on non-native children test set *kids-sg-tst*. Tone recognition performance is consistent with phone and syllable error detection performance: the Mixed model outperforms baseline model, and the Multi-task learning further improves the performance of the Mixed model.

6. Conclusion

In this work, we proposed to use multi-task learning to improve children's Mandarin acoustic modeling by leveraging on a larger set of adult data. We showed that multi-task learning outperforms the baseline data augmentation approach of consolidating children and adult speech corpora into one single dataset on tasks such as automatic speech recognition and mispronunciation detection of phonetic, lexical tones and syllable errors.

7. References

[1] Nancy F. Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li, "Singakids-mandarin: Speech corpus of singaporean children speaking mandarin chinese," in *Interspeech*, 2016.

[2] Matteo Gerosa, Diego Giuliani, and Fabio Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10, pp. 847–860, 2007.

[3] Alexandros Potamianos, Shrikanth Narayanan, and Sungbok Lee, "Automatic speech recognition for children," in *Eurospeech, vol. 97, pp. 2371 - 2374*, 1997.

[4] Hank Liao, Golan Pundak, Olivier Siohan, Melissa Carroll, Noah Coccaro, Qi-Ming Jiang, Tara N. Sainath, Andrew Senior, Francoise Beaufays, and Michiel Bacchiani, "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015.

[5] Keelan Evanini and Xinhao Wang, "Automated speech scoring for non-native middle school students with multiple task types," in *Interspeech*, 2013.

[6] Khairun nisa Hassanali, Su-Youn Yoon, and Lei Chen, "Automated scoring of non-native children's spoken language proficiency," in *Slate*, 2015.

[7] M. Russell, S. Darcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," in *Signal Processing Letters, IEEE, vol. 14, no. 12, pp 1044-1046*, 2007.

[8] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Workshop on Child, Computer and Interaction (WOCCI)*, 2014.

[9] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," in *IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp 603 - 616*, 2003.

[10] Mauro Nicolao, Amy V.Beeston, and Thomas Hain, "Automatic assessment of English learner pronunciation using discriminative classifiers," in *ICASSP*, 2015.

[11] Kun Li, Xiao-Jun Qian, and Helen M. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing PP(99):1-1*, 2016.

[12] Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wanga, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," in *Speech Communciation, Volume 67, March 2015, Pages 154 - 166* , 2015.

[13] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for us english: Child interaction with living-room-electronic-devices," in *Workshop on Child, Computer and Interaction (WOCCI)*, 2014.

[14] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *IEEE Workshop on Spoken Language Technology (SLT)*, 2014.

[15] Romain Serizel and Diego Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, pp. 1–26, 2016.

[16] Joachim Fainberg, Peter Bell, Mike Lincoln, and Steve Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Interspeech*, 2016.

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

[18] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP pp. 7304-7308*, 2013.

[19] Michael L. Seltzer and Jasha Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *ICASSP*, 2013.

[20] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.

[21] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *NAACL*, 2015.

[22] Qiong Hu, Zhizheng Wu, Korin Richmond, Junichi Yamagishi, Yannis Stylianou, and Ranniery Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," in *Proc. Interspeech*, Dresden, Germany, September 2015.

[23] Gokhan Tur, "Multitask learning for spoken language understanding," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1.

[24] Rong Tong, Nancy F. Chen, Bin Ma, and Haizhou Li, "Context aware mispronunciation detection for mandarin pronunciation training," in *INTERSPEECH*, 2016.

[25] Rong Tong, Nancy F. Chen, Boon Pang Lim, Bin Ma, and Haizhou Li, "Tokenizing fundamental frequency variation for mandarin tone error detection," in *ICASSP 2015, April 19-24, 2015*. 2015, pp. 5361–5365, IEEE.

[26] Rong Tong, Boon Pang Lim, Nancy F Chen, Bin Ma, and Haizhou Li, "Subspace gaussian mixture model for computer-assisted language learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5347–5351.

[27] Rong Tong, Nancy F. Chen, Bin Ma, and Haizhou Li, "Goodness of Tone (GOT) for non-native mandarin tone recognition," in *INTERSPEECH 2015, Dresden Germany, September 6-10, 2015*, 2015, pp. 801–805.

[28] Y.-Q. Zu, W.-X. Li, M.-C. Ho, and C. Chan, "HKU96 a Putonghua corpus (CD-ROM version) ," 1996.

[29] Dong Wang, Xuewei Zhang, and Zhiyong Zhang, "THCHS-30 : A Free Chinese Speech Corpus," 2015, http://arxiv.org/abs/1512.01882.