TDNN-BASED MULTILINGUAL MIX-SYNTHESIS WITH LANGUAGE DISCRIMINATIVE TRAINING

Gulijiamali Maimaitiaili* Zhiyong Zhang† Aisikaer Rouzi†
*School of Mathematical Science of Xinjiang Normal University, Urumqi, China
E-mail: guljamal123@sina.com Tel: +86-18999876139
† Tsinghua University, Beijing, China
E-mail: zhangzhiyong.115@163.com Tel: +86-18910665504
E-mail: askhar@xju.edu.cn Tel: +86-13391688228

***Abstract*—We propose to build a time delay neural network based, Mandarin and Uyghur, bilingual TTS system. To facilitate the phone sharing across the two languages, we design multilingual question set, which includes language specific, language independent and IPA sharing questions. Neural network based language discriminative approach is also used to get better network output during mix-training. Language discriminative information is augmented as auxiliary feature to linguistic features at input level and control the output of a feed-forward deep neural network at the output layer. Preliminary experimental results show that the multilingual mix-synthesis models can be constructed using the proposed language discriminative training architecture. Monolingual and multilingual system performance are evaluated and compared for both languages, language discriminative codes also show to be efficient to distinguish the contextual linguistic features from different languages and can help to control the output features.**

***Index Terms*—** Speech synthesis, multilingual, time delay neural network (TDNN), phone sharing, data augmentation

I. INTRODUCTION

Most Text to Speech (TTS) systems today assume that the input is in a single language written script. However, due to the growing influence between different languages, we now see code-mixing in text becoming more common in bilingual and multilingual communities. Therefore, a multilingual TTS system, in which one engine can synthesize multiple languages or even mixed-languages, is in a great demand with the rapid improvements of TTS technologies.

Hidden Markov model (HMM) based speech synthesis has dominated SPSS in the past decades, due to high effectiveness to model the evolution of speech signals as a stochastic sequence of acoustic feature vectors and can obtain a high quality acoustic model using even a relatively small size corpus [1]. HMM-based bilingual TTS system has been proposed for English-Mandarin code switched TTS [2]. This approach uses speech databases in both languages from the same speaker and a single TTS system that shares phonetic space is built. With shared phones, the system has a smaller footprint and synthesis quality much better for mixed-language synthesis.

However, the naturalness of synthetic speech rendered through HMM-based synthesis system is not as good as that of the best samples from unit-selection speech synthesizers. This is mainly caused by three factors: quality of vocoder, accuracy of acoustic model, and effect of over-smoothing. To get the high accuracy of training model, the use of deep neural networks (DNNs) has been proposed. Several independent studies have demonstrated that DNNs can produce more natural synthesized speech than the conventional HMM-based speech synthesis in various training conditions [3-10]. One reason for the success of DNNs compared to HMMs is that they can provide a better and more efficient representation of complex dependencies between linguistic and acoustic features. To model the sequential nature of speech, the DNNs are extended to recurrent neural networks, especially long short-term memory networks (LSTMs), which capture the correlations among consecutive frames [10-12]. Study in [13] have proposed a multilingual BLSTM based speech synthesis system that shares hidden layers across different languages.

The DNN-based speaker adaptation also outperformed the HMM-based systems in terms of naturalness and speaker similarity[14]. Several studies have also explored DNNs for speaker adaptation in TTS [14-19]. In DNNs, the adaptation techniques have been applied at three different levels of input [14, 15, 17, 18, 19, 20], model [14, 16, 20], and output [14, 15, 20]. DNN-based speaking style adaptation of Lombard speech has been proposed and also confirmed that the DNNs are able to adapt better to the Lombard style than HMMs [20]. However, there are no previous studies on language discrimination controlling in DNN-based multilingual or mix-synthesis.

In this work, we investigate a TDNN-based Mandarin and Uyghur bilingual TTS system. To the best of our knowledge, it is the first attempt for neural network based speech synthesis and mix-synthesis for Uyghur language. We design language specific, language independent and IPA sharing questions, to facilitate phone sharing across the two languages in mix-TTS. To get the better network output in mix-training, we also consider neural network based adaptation and controlling methods. We augment the language discriminative information as auxiliary features to linguistic features at input level and output level. The augmented information enables the network to distinguish the contextual linguistic features from different languages at the input layer, and constrain the output features of different languages at the output layer. As these controlling techniques are performed at different levels, they may be usefully combined. We have performed experimental analysis on the performance of each individual discriminative training and that of their combinations.

II. FEATURE DESIGNING

For building a mix-synthesis system, the most important steps of data preparation are deciding a phone set to cover all

speech sounds in different languages, designing multilingual question set and creating input training data. Additionally, we also hope such a phone and question set can be compact enough to facilitate phone sharing across languages and make a reasonable sized training model.

The phone set we used is the union of all phones in Mandarin and Uyghur, contains 48 and 42 phones respectively. The questions set that we designed for multilingual synthesis include:

a) Language specific question: e.g. Does the tone of current phone is 0?, since there is no tone in Uyghur, that kind of question only specific for Mandarin. For Uyghur e.g. Does the current phone belongs to vowel, which contain /a/, /æ/, /e/, /i/, /o/, /u/, /ø/, /y/, 8 vowels in Uyghur.

b) Language independent question: mainly includes numeric values (e.g. the number of words in the phrase, the relative position of the current frame in the current phoneme) and silence pattern (e.g. is current silence pau?)

c) Phone sharing IPA question: to facilitate the phone sharing between two different languages, we explore the sharing IPA between them and designing a set of IPA sharing questions.

IPA (International Phonetic Alphabet) is an international standard to transcribe speech sounds in any spoken language. It classifies phonemes according to their phonetic-articulatory features. Phonemes of different languages labeled by the same IPA symbol should be considered as the same phoneme by ignoring the language-dependent aspects of speech perception.

We found twelve consonants /p/, /m/, /f/, /t/, /n/, /l/, /k/, /ŋ/, /x/, /s/, /j/, /w/, and six vowels (ignoring the tone information) /a/, /e/, /o/, /i/, /u/, /y/, can be shared between the two languages according to their IPA symbols. The sizes of different type of questions are listed in table 1.

Table.1 size of multilingual mix question set

Question type	language	size
Language specific	Mandarin	319
	Uyghur	233
Language independent	common	22
IPA sharing question	common	110
total		684

III. MULTILINGUAL MIX-SYNTHESIS ARCHITECTURE

In DNN-based monolingual speech synthesis, the input linguistic features and the corresponding output acoustic features from one single language, different languages have their different linguistic contextual information. Thus, the input and output layers are language dependent.

In multilingual speech synthesis, though the input and output layer of DNN are language dependent, but the hidden layers can be considered as language independent, which transforms the input of linguistic features to an internal language independent representation, and the internal representation can be shared across different languages. Because each language has its unique linguistic features, different languages may correspond to different dimensional input features, which include binary answers to questions about linguistic contexts and numeric values, etc. The single uniform representation of input features from different languages has obtained by using multilingual question set designed in section 2. Dimension of the uniform input features equals to the sum of the input feature dimensions of different language, when the current input features are from language 1, the uniform input features are constructed by concatenating the input features from language 1 with appending all zeros (representations of language specific questions from language 2) and vice versa, as shown in lower part of fig 1. Then, mix model accepts uniform input features, the hidden layers of model are perceived as feature transformations and can be shared across different languages. The output layers then use the commonly internal representations to predict the acoustic features of different languages.

Joint training with multiple languages may increase the perplexity of the model. In order to help the network learning more about language variation during training and to get more language discriminative output features from network, we augment language specific information as auxiliary features to the input and output features during the model training.

IV. LANGUAGE DISCRIMINATIVE TDNN TRAINING

In the process of training, we use TDNN (time delay neural network) architecture for both duration and acoustic model training. TDNN has been shown to be effective in modeling long range temporal dependencies [21], and uses a modular and incremental design to create larger networks from sub-components [22]. TDNN architecture which models long term temporal dependencies with training times comparable to standard feed-forward DNNs and uses sub-sampling to reduce computation during training also shows the effectiveness in learning wider temporal dependencies in both small and large data scenarios on LVCSR tasks [23] and get better results than recurrent neural network. Splicing increasingly wide asymmetric context as the layer rise architecture is used in our system according to different time splicing experiment and experience from [23]. We splice frames of offsets [-2, 2], {0}, {-1, 2}, {-3, 4}, {-7, 2}, {0} at six hidden layers of duration model and offsets of [-2, 2], {-1, 2}, {-3, 4}, {0} at four hidden layers of acoustic model training. Fig 1 shows the TDNN acoustic model training architecture used in this paper.

The use of augmented or auxiliary features is an widely used approach in speaker adaptive neural network architecture in which the linguistic features are augmented with additional speaker-specific features computed for each speaker at both training and test stages. Studies in [14,15,17] have successfully used the auxiliary information such as gender, speaker identity, age or i-vector for speaker adaptation in DNN-based speech recognition and synthesis. In this work, we augment the language specific information as auxiliary features to linguistic features at input level. We use one-hot vector codes for auxiliary features. If there are *N* languages in

the training set, the standard one-hot vector language discrimination code for the i th language can be defined as:

$$S_i = (s_1, s_2, \mathrm{L}\ , s_N) = \begin{cases} s_n = 1, & n = i \\ s_n = 0, & n \neq i \end{cases} \quad (1)$$

The augmented values enable distinguishing the contextual linguistic features from different languages. Besides, we also experiment to constrain the output features with language discrimination code and combination both of input and output layer.

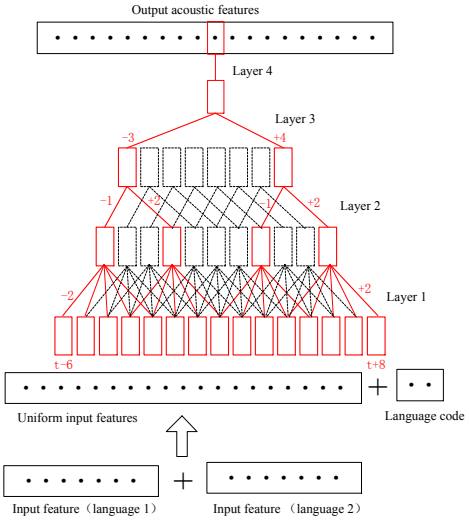

Fig.1 TDNN training architecture in multilingual mix-synthesis

V. EXPERIMENTS

A. Experimental setup

In training the multilingual TTS, the 1898 sentence speech utterances were used as training set, including 948 Mandarin and 950 Uyghur. 10% of training data was used as development set and 50 utterances were used as evaluation set. All speech data are recorded as newspaper reading style by different native female speaker for each language with sampling rate of 16 kHz.

The full contextual labels were generated from the text files, which were available along with speech, using the text analysis process of different languages. As for the linguistic features, the contextual labels include quin-phone following with position and length related features of phone, syllable, word, phrase, sub-sentence and sentence. Besides, contextual labels of Mandarin include tone of the syllable, prosody and part-of-speech features. The part-of-speech and prosody related features for Uyghur were not ready in time for this paper, owing to there are not enough labeling data for them now.

The overall multilingual synthesis system was done with the Kaldi toolkit [24]. We trained two TDNN model for duration and acoustic features separately. For duration modeling, the input comprises binary features derived from a subset of the multilingual questions set designed in section 2, with 684 dimension in total. Frame-aligned data for TDNN training was created by forced alignment using the HMM system. The output is durations for every phone. 6 hidden layers of Relu activation function with 1024 nodes were used. L2 regularization was applied to the weights with penalty factor of 1e-3, exponential decay learning rate was applied with initial value 0.02, the mini-batch size was 256, and momentum was 0.2.

For acoustic modeling, the input uses the same features as duration prediction, to which 4 numerical features are appended to provide information about the position and durations of frame within the phoneme. In total, the input feature vector was 688 in dimension. The output of acoustic features comprise 60-D MGCs (Mel Generalization Cepstrum), 1-D BAPs (Band Aperiodicity), 1-D F0 and their corresponding delta and delta-delta features. The F0 was linearly interpolated and an extra V/UV feature was added to acquire the voice/unvoiced information at runtime synthesis. Thus, in total, the output feature was 187 dimensional. The acoustic model consists of 4 hidden layers with 1024 hidden units using Relu as activation function in each layer. L2 regularization was applied to the weights with penalty factor of 1e-2, exponential decay learning rate was applied with initial value 0.0015, the mini-batch size was 256, and momentum was 0.2.

In acoustic and duration model, the input features were normalized to the range of [0.1, 0.99] by using the min-max normalization and the output features were normalized to zero mean and unit variance. The development and evaluation set were normalized by the values derived from the training data.

At synthesis time, duration is predicted first, and is used as an input to the acoustic model to predict the speech parameters. To generate smooth parameter trajectories, the maximum likelihood parameter generation (MLPG) algorithm was applied on predicted acoustic parameters using the global variances of training data, and spectral enhancement postfiltering is applied to the resulting MGC trajectories. Finally, the WORLD [25] vocoder is used to synthesize the waveform.

B. Objective evaluation

To evaluate the performance of our system, we conducted a set of objective evaluations on the 50 utterances from the different languages test set. While the objective metrics do not map directly to perceptual quality, they are often useful for system tuning. The mel-cepstral distortion (MCD), band aperiodicity distortion (BAP), voiced/unvoiced prediction error (VUV), root mean squared error (RMSE) and Pearson correlation were computed between predicted and original acoustic parameters of the entire evaluation set.

First, we investigated the performance of monolingual TTS for Uyghur language, since this is the first attempt for neural network based Uyghur TTS. For monolingual Uyghur speech synthesis, we conducted two different kinds of model training architecture experiments, including TDNN and BLSTM. All the model parameters for both approaches are the same as described in section A. In BLSTM, we used Merlin

toolkit[12], with configuration of five feedforward hidden layers of 1024 hyperbolic tangent units each, followed by a single BLSTM layer with 512 units for duration model. For acoustic model, we configure three feedforward hidden layers of 1024 hyperbolic tangent units each, followed by a single BLSTM layer with 512 units. Results are presented in Table 2. TDNN based training architecture get slight better performance in our experiment under our experimental condition.

Table.2 Experimental results of monolingual Uyghur TTS

Method	MCD(dB)	BAP(dB)	F0-RMSE(Hz)	CORR	VUV(%)
TDNN	7.504	0.285	17.651	0.686	10.887
BLSTM	7.709	0.302	18.660	0.649	11.220

Second, we investigated the performance of our proposed multilingual mix-synthesis for different languages and language discriminative training approaches at the different layers. Our baseline system is the mix-training of two languages without any discrimination approach. We added the language discriminative code to output and input layer separately, and also added to both of them at the same time. Performance of multilingual mix-synthesis for Uyghur test utterances presented in table 3 and for Mandarin in table 4. It can be seen from the table that all the discriminative training in different layers get the better performance than the baseline system, and adding language discriminative code to both input and output layer get the best result in both multilingual experiment. It also can be seen that language discrimination in input layer is more efficient than output controlling. As for Mandarin test in table 4, we also give the TDNN based monolingual Mandarin synthesis results for comparison.

Table.3 Results of multilingual TTS for the Uyghur test utterances

Method	MCD(dB)	BAP(dB)	F0-RMSE(Hz)	CORR	VUV(%)
baseline	8.070	0.314	19.629	0.606	11.979
Output	8.057	0.314	19.426	0.610	11.779
Input	7.634	0.291	18.356	0.653	11.182
Combination	7.635	0.292	18.336	0.653	11.090

Table.4 Results of multilingual TTS for the Mandarin test utterances

Method	MCD(dB)	BAP(dB)	F0-RMSE(Hz)	CORR	VUV(%)
monolingual	5.077	0.259	35.224	0.822	5.930
baseline	5.936	0.277	37.189	0.802	6.984
Output	5.927	0.276	37.243	0.804	6.931
Input	5.829	0.265	35.356	0.819	6.209
Combination	5.825	0.263	35.485	0.819	6.251

It has to be noticed that, in all monolingual and multilingual synthesis results in table 4, the F0-RMSE of Mandarin are significantly high. It also indicates that the differences between the tone and no-tone language. The f0 fluctuation of Mandarin utterances is significantly higher than Uyghur, this may be one of the main reason that affects the multilingual mix-synthesis performance. The f0 trajectory differences between two languages are also clearly shown in Fig 3 and Fig 4, which include f0 trajectory of original and utterances synthesized by the monolingual and multilingual combination approach. Other evaluation metrics including MCD, BAP and VUV are lower than Uyghur test, due to Mandarin training data have rich contextual features such as prosody and part-of-speech than Uyghur.

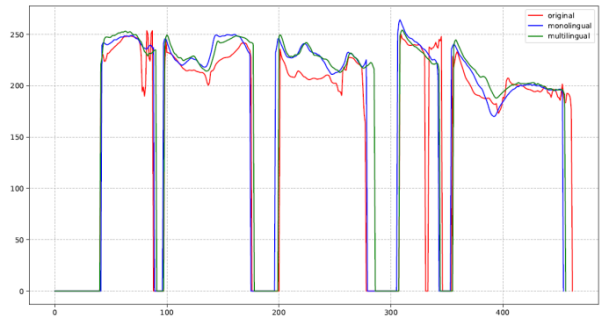

Fig.3 F0 trajectory for held-out Uyghur utterances: “he came here for a trip”

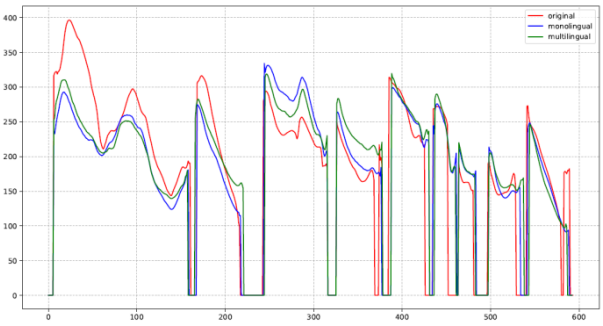

Fig.4 F0 trajectory for held-out Mandarin utterances: “It's like being in the arms of the earth”

C. *Subjective evaluation*

We conducted listening tests to assess the naturalness of the synthesized speech obtained from different training architecture. Two MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) tests [26] were conducted to assess the naturalness of Mandarin and Uyghur synthesis utterances. 20 native listeners participated in each test. Each listener rated 15 sets which were randomly selected from the testing utterances. For Mandarin listening test, each set consisted of 5 stimuli of the same sentence generated by each of the four multilingual systems and one monolingual system plus the copy-synthesis speech used as the hidden reference. For Uyghur utterance listening test, there is one more stimuli for BLSTM. The listeners were asked to rate each stimulus from 0 (extremely bad for naturalness) to 100 (same naturalness as the reference speech).

The MUSHRA scores for all the Mandarin and Uyghur synthesis utterances are presented in Fig. 5 and Fig.6. In both two listening test, all the results of different approach are almost the same trend with objective test. The difference between input layer discrimination and combination approach is not significant and gets the better performance than other systems within multilingual synthesis. It also can be clearly shown from the figure that, mix-synthesis quality gradually increases with the discrimination approaches of different layer, and combination approach is very close to non-mixed. From

that point of view we can see that more efficient language discrimination code is very essential in our task, actually we also investigate the language vector learning approaches, but not ready in time for this paper. Lack of data, language and speaker differences are the greatest difficult for further enhancing the performance of our mix-synthesis.

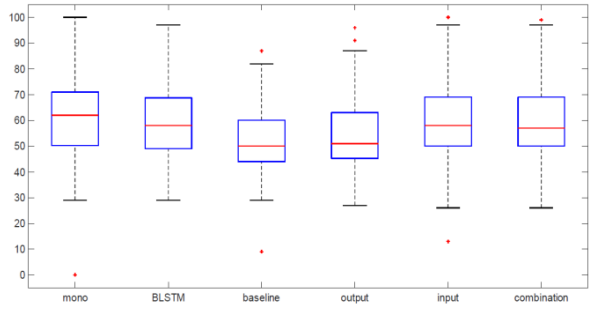

Fig.5 Box plot of MUSHRA results for Uyghur synthesis utterances

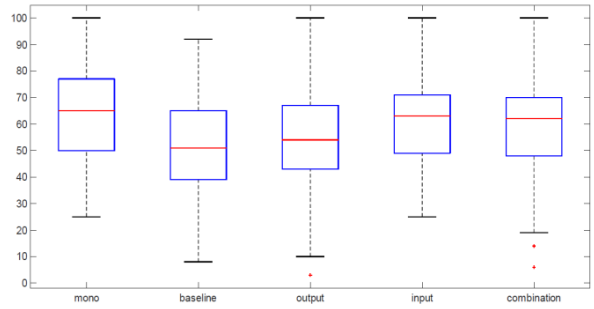

Fig.6 Box plot of MUSHRA results for Mandarin synthesis utterances

VI. CONCLUSIONS

In this paper, a systematic experimental analysis was conducted on multilingual mix-synthesis for Mandarin and Uyghur language. To facilitate the phone sharing between two languages, we design multilingual mix-question set and investigate the IPA sharing between them. A TDNN based training architecture is used for duration and acoustic model training. In order to help the network to learn more about language variation and to get more language discriminative output features, a language discriminative training approach is used and the experimental results of adding language constraint codes to different layer are compared. The preliminary experimental results show that multilingual mix-synthesis quality is very close to non-mixed. In the future, we would like to investigate more complicated models, such as multitask training architecture, and also like to investigate more efficient language sharing and language discriminative code learning architecture for further enhancing the performance of the system.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61462087 and NO. 61751316.

REFERENCES

[1] H. Zen, K. Tokuda, and A.W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, no. 11, pp.1039–1064, 2009.
[2] H. Liang, Y. Qian, and F. K. Soong, “An HMM-based bilingual (Mandarin-English) TTS,” Proceedings of SSW6, 2007.
[3] Heiga Zen, Andrew Senior, and Mike Schuster, “Statistical parametric speech synthesis using deep neural networks,” in ICASSP, 2013, pp. 7962–7966.
[4] Z.-H. Ling, L. Deng, and D. Yu, “Modeling spectral envelopes using Restricted Boltzmann Machines and Deep Belief Networks for statistical parametric speech synthesis,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp.2129–2139, 2013.
[5] S. Kang, X. Qian, and H. Meng, “Multi-distribution deep belief network for speech synthesis,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2013.
[6] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, “On the training aspects of deep neural network (DNN) for parametric TTS synthesis,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2014.
[7] H. Zen and A. Senior, “Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis,” in Proc.IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2014.
[8] 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 Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2015.
[9] B. Ur´ıa, I. Murray, S. Renals, and C. Valentini-Botinhao, “Modelling acoustic feature dependencies with artificial neural networks: Trajectory-RNADE,” in Proc IEEE ICASSP, 2015.
[10] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong, “TTS synthesis with bidirectional lstm based recurrent neural networks,” in Interspeech, 2014, pp. 1964–1968.
[11] Heiga Zen and Has¸im Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in ICASSP, 2015, pp. 4470–4474.
[12] Wu, Z, Watts, O, King, S, “Merlin: An Open Source Neural Network Speech Synthesis System,” in 9th ISCA Speech Synthesis Workshop, 2016.
[13] Quanjie YU, Peng LIU, Zhiyong WU, Shiyin KANG, Helen MENG, Lianhong CAI, “Learning Cross-lingual Information with Multilingual BLSTM for Speech Synthesis of Low-resource Languages,” [in] Proc. ICASSP. Shanghai, China, 20-25 March, 2016.
[14] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King, “A study of speaker adaptation for dnn-based speech synthesis,” in Interspeech, 2015.
[15] Blaise Potard, Petr Motlicek, and David Imseng, “Preliminary work on speaker adaptation for dnn-based speech synthesis,” Tech. Rep., Idiap, 2015.
[16] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, “Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” in ICASSP, 2013, pp. 7893–7897.

[17] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” in ASRU, 2013, pp. 55–59.
[18] Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, Junichi Yamagishi, “Adapting and Controlling DNN-based Speech Synthesis Using Input Codes,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),SP-L4.3, 4905-4909, Mar 2017.
[19] Moquan Wan, Gilles Degottex, Mark J. F. Gales, “Integrated speaker-adaptive speech synthesis,” ASRU 2017: 705-711
[20] Bajibabu Bollepalli, Manu Airaksinen, Paavo Alku, “LOMBARD SPEECH SYNTHESIS USING LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORKS,” ICASSP 2017: 5505-5509
[21] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, “Phoneme recognition using time-delay neural networks,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328–339, Mar. 1989.
[22] A. Waibel, “Modular construction of time-delay neural networks for speech recognition,” Neural computation, vol. 1, no. 1, pp.39–46, 1989.
[23] Peddinti Vijayaditya, Povey Daniel, Khudanpur Sanjeev, “A time delay neural network architecture for efficient modeling of long temporal contexts,” In INTERSPEECH-2015, 3214-3218.
[24] Povey D, Ghoshal A, “The Kaldi speech recognition toolkit,” Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2011: 1-4
[25] M. MORISE, F. YOKOMORI, and K. OZAWA, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, 2016.
[26] Sebastian Kraft and Udo Z¨olzer, “Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality,” in Linux Audio Conference, Karlsruhe, DE, 2014.