LEARNING CROSS-LINGUAL INFORMATION WITH MULTILINGUAL BLSTM FOR SPEECH SYNTHESIS OF LOW-RESOURCE LANGUAGES

Quanjie Yu[1,2], *Peng Liu*[1,2], *Zhiyong Wu*[1,2,3], *Shiyin Kang*[3], *Helen Meng*[1,3], *Lianhong Cai*[1,2]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

{yuqj13, liup12}@mails.tsinghua.edu.cn,
{zywu, sykang, hmmeng}@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

Bidirectional long short-term memory (BLSTM) based speech synthesis has shown great potential in improving the quality of the synthetic speech. However, for low-resource languages, it is difficult to obtain a high quality BLSTM model. BLSTM based speech synthesis can be viewed as a transformation between the input features and the output features. We assume that the input and output layers of BLSTM are language-dependent while the hidden layers can be language-independent if trained properly. We investigate whether sufficient training data of another language (auxiliary) can benefit the BLSTM training of a new language (target) that has only limited training data. In this paper, we propose 1) a multilingual BLSTM that shares hidden layers across different languages and 2) a specific training approach that can best utilize the training data from both the auxiliary and target languages. Experimental results demonstrate the effectiveness of the proposed approach. The multilingual BLSTM can learn the cross-lingual information, and can predict more accurate acoustic features for speech synthesis of the target language than the monolingual BLSTM that is trained with only the data from the target language. Subjective test also indicates that multilingual BLSTM outperforms the monolingual BLSTM in generating higher quality synthetic speech.

Index Terms— Speech synthesis, bidirectional long short-term memory (BLSTM), low-resource, multilingual, cross-lingual

1. INTRODUCTION

In recent years, statistical parametric speech synthesis (SPSS) has become popular because it has many advantages. Hidden Markov model (HMM)-based speech synthesis, deep neural networks (DNN)-based speech synthesis and bidirectional long short term memory (BLSTM) based speech synthesis are instances of SPSS. HMM-based speech synthesis is effective to model the evolution of speech signals as a stochastic sequence of acoustic feature vectors [1] and can obtain a high quality acoustic model using even a relatively small size corpus. On the other hand, DNN or BLSTM has sophisticated network architecture and needs moderate or large corpus (phonetically and prosodically rich) to train a good model [2, 3, 4]. Given that recording training data is often at great expense, the available training data is always very limited, especially for low-resource languages of a particular specific speaker.

To deal with the lacking training data problem, in HMM-based synthesis, speaker adaptive training technique is proposed to train an average voice model using different speakers' training data and then adapt the average voice model to a specific speaker [5]. [6] investigated the similar technique in DNN-based speech synthesis utilizing multi-task learning [7] and transfer learning technologies.

In this paper, we propose a multilingual BLSTM, in which the hidden layers are shared across different languages while the input and output layers are language-dependent. The BLSTM recurrent neural network can be considered as a model that learns a complicated feature transformation through hidden layers and output layer. BLSTM based speech synthesis can transform the linguistic features to acoustic features. In BLSTM based synthesis, BLSTM can be decomposed into hidden layers for linguistic features transformation and output layers for acoustic feature regression. The shared hidden layers and separate regression layer of each language in multilingual BLSTM are jointly trained with two language corpora. Multilingual BLSTM can then learn knowledge across multiple languages and transfer the knowledge from one language to another.

2. BIDIRECTIONAL LSTM RECURRENT NEURAL NETWORK (BLSTM)

A recurrent neural network (RNN) is able to deal with the correlations between data points embodied in (time) sequential data. The input vectors are fed into the hidden layer of RNN one at a time. The hidden layer output activations of the last time step are also fed into the hidden layer. In this way, the structure can exploit all the available input information up to the current time step. The feedforward process of RNN is:

$$\mathbf{h}_t = \mathcal{H}(\mathbf{W}_{\mathrm{xh}}\mathbf{x}_t + \mathbf{W}_{\mathrm{hh}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{h}}) \quad (1)$$

$$\mathbf{y}_t = \mathbf{W}_{\mathrm{hy}}\mathbf{h}_t + \mathbf{b}_{\mathrm{y}} \quad (2)$$

$\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_T)$ is the input vector sequence, $\mathbf{h} = (\mathbf{h}_1, ..., \mathbf{h}_T)$ is the hidden state vector sequence computed from input vector sequence, and $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$ is the output vector sequence. $\mathcal{H}$ is the activation function for hidden state. $\mathbf{W}_{\mathrm{xh}}$, $\mathbf{W}_{\mathrm{hh}}$ and $\mathbf{W}_{\mathrm{hy}}$ represent the input-hidden, hidden-hidden and hidden-output weight matrices repectively. $\mathbf{b}_{\mathrm{h}}$ and $\mathbf{b}_{\mathrm{y}}$ denote the bias vectors for hidden state vectors and output vectors.

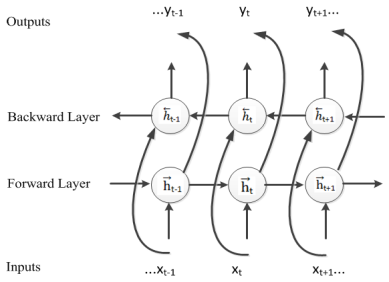

Fig. 1. Bidirectional recurrent neural network (BRNN).

The disadvantage of RNN is that it can only access the previous inputs. Bidirectional RNN (BRNN) can access both previous and future inputs utilizing the bidirectional architecture [8], as shown in Fig. 1. The feedforward process of BRNN include forward $\overrightarrow{\mathbf{h}}$, and backward hidden sequence $\overleftarrow{\mathbf{h}}$.

In text-to-speech (TTS) synthesis task, the long time span contexts in an utterance need to be modeled. However, the RNN and BRNN structure is only able to retain short term memory beacause of the vanishing gradient problem. Long short term memory (LSTM) [9] recurrent neural network is designed to tackle with long time lags. An LSTM layer consists of memory blocks which are a set of connected blocks. A single memory block is shown in Fig. 2. Each block contains four types of units: one or more recurrently connected memory cells, input gate, output gate and forget gate. These three gates are multiplicative units which simulate read, write and reset operations for memory cells. The feedforward process of LSTM is:

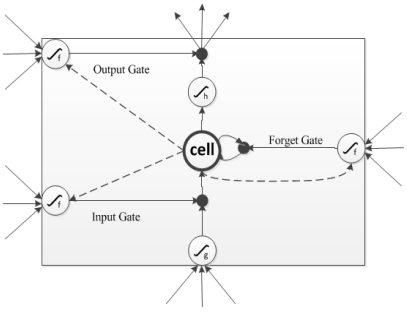

Fig. 2. Long short term memory (LSTM) block.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{\mathrm{xi}}\mathbf{x}_t + \mathbf{W}_{\mathrm{hi}}\mathbf{h}_{t-1} + \mathbf{W}_{\mathrm{ci}}\mathbf{c}_{t-1} + \mathbf{b}_{\mathrm{i}}) \quad (3)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{\mathrm{xf}}\mathbf{x}_t + \mathbf{W}_{\mathrm{hf}}\mathbf{h}_{t-1} + \mathbf{W}_{\mathrm{cf}}\mathbf{c}_{t-1} + \mathbf{b}_{\mathrm{f}}) \quad (4)$$

$$\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tan\mathbf{h}(\mathbf{W}_{\mathrm{xc}}\mathbf{x}_t + \mathbf{W}_{\mathrm{hc}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{c}}) \quad (5)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{\mathrm{xo}}\mathbf{x}_t + \mathbf{W}_{\mathrm{ho}}\mathbf{h}_{t-1} + \mathbf{W}_{\mathrm{co}}\mathbf{c}_{t-1} + \mathbf{b}_{\mathrm{o}}) \quad (6)$$

$$\mathbf{h}_t = \mathbf{o}_t \tan h(\mathbf{c}_t) \quad (7)$$

where σ is a logistic function; i, f, o and c respectively represent input gate, forget gate, output gate and cell memory [10].

We replace the hidden units in forward layer and backward layer of BRNN by LSTM blocks to derive the bidirectional long short term memory (BLSTM) recurrent neural network. This structure can exploit long term memories in both directions of a sequence.

3. APPROACHES

3.1. Monolingual BLSTM

A BLSTM based TTS synthesis approach was proposed by [4], where BLSTM uses the converted linguistic features as input and acoustic features as output. The BLSTM model can be considered as a sophisticated transformation model that learns the relationships between input linguistic features and output acoustic features. Fig. 3 shows two BLSTM models of language 1 and language 2 that are trained separately. We call each model the monolingual BLSTM, where the input linguistic features and the corresponding output acoustic features of a BLSTM come from one single language. It should be noted that different languages may have different linguistic contextual information as input features. Hence, the input vectors of monolingual BLSTM for different languages may be different.

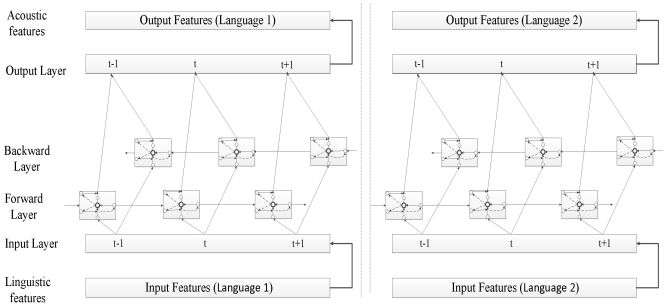

Fig. 3. Architecture of monolingual BLSTM.

3.2. Multilingual BLSTM

In the above monolingual BLSTM for TTS synthesis, the input and output layers are language dependent. We assume that the hidden layers (i.e. the forward layer and the backward layer) of a BLSTM can be language independent, which transform the input linguistic features to an internal language independent representation. And such internal representation can be shared across different languages. With this assumption, we propose a multilingual BLSTM that shares hidden layers across different languages and a specific training approach to train such multilingual BLSTM.

Fig. 4 shows the architecture of the proposed multilingual BLSTM whose input layer and hidden layers are shared across different languages, while the output layer is not shared. Different languages have their own output layers and related weight matrices from hidden layers. Because each language may have its unique linguistic features, different languages may correspond to different dimensional input features. To solve such problem, the input feature vectors of different languages are combined together to form a single uniform representation of input features, as shown in the lower part of Fig. 4. The dimension of the uniform input features equals to the

sum of the input feature dimensions of language 1 and language 2. When the current input features are from language 1, the uniform input features are constructed by concatenating language 1's input features with appending all zeros. The uniform representations of language 2's input features are similarly constructed by prepending zeros to language 2's real input features. Then multilingual BLSTM accepts uniform input features to the input layer. The hidden layers of multilingual BLSTM are perceived as feature transformations and can be shared across languages. The shared hidden layers of different languages can transform the uniform input features to internal representation that can provide benefits to both languages. The output layers then use the commonly internal representations to predict the acoustic features of different languages.

As we can see, multilingual BLSTM is trained with speech data from both low-resource (or resource-limited) language and resource-rich language (with sufficient training data). This strategy is a kind of multi-task learning: the tasks of resource-limited and resource-rich languages are trained simultaneously. The cross-lingual information captured by the hidden layers of multilingual BLSTM leads to better performance in TTS than the monolingual BLSTM.

4. EXPERIMENTS

4.1. Experimental setup

The CMU_ARCTIC_SLT corpus is used as English (ENG) recordings. To verify that our proposed approach can learn cross-lingual transfer information of different languages pairs, Mandarin (MAN) and Cantonese (CAN) corpora that were recorded by different female speakers in broadcast news reading style are adopted in our experiments. Each corpus is phonetically and prosodically rich. Table 1 shows statistics of the utterance numbers for each corpus. The English corpus consists of 550 speech utterances, with each utterance having around 15 words. The Mandarin and Cantonese corpora both contain 2,000 sentences of speech utterances, with average 20 Chinese syllables for each utterance. For the 550 English speech utterances, 500 utterances are used as the training set for TTS synthesizers, while the remaining 50 utterances serve as the test set. We have two language pairs in our experiments, Mandarin-English and Cantonese-English for multilingual BLSTM training and evaluation.

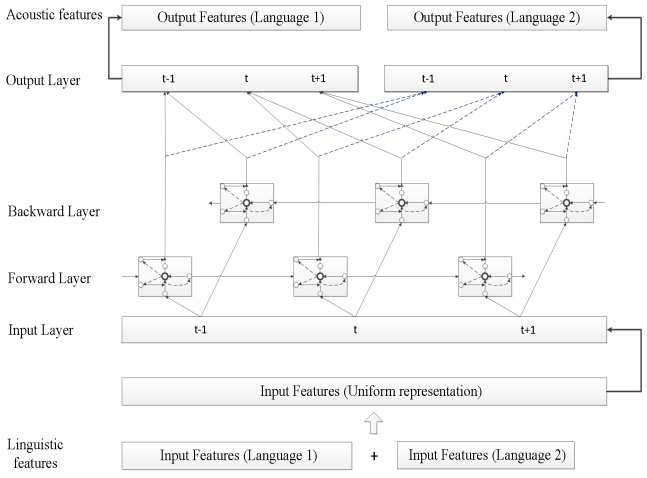

Fig. 4. Architecture of multilingual BLSTM.

Table 1. *Utterance number in each corpus of different languages.*

Language	Numbers
English	550
Mandarin	2000
Cantonese	2000

All speech recordings are stored as Microsoft wave file with the sampling rate of 16 kHz. The acoustic features are then extracted with a frame window of 25 ms length, and frame shift of 5 ms, 35 order Mel Generalization Cepstrum (MGC), voiced/unvoiced (V/UV) flag, log-F0 together with their delta and delta-delta deviations are extracted, which serve as the acoustic parameters for TTS synthesizers. As for the linguistic features, the phonetic and prosodic contexts of English include quin-phone, stress of syllable, the position of a phone, syllable and word in phrase and sentence, the length of word and phrase, TOBI and POS of word, with 503 dimensions in total; and the phonetic and prosodic cotntexts of Mandarin and Cantonese include tri-syllable, the tone of the syllable, the position of a syllable in prosodic word, prosodic phrase, intonational phrase and sentence, the length of prosodic word, prosodic phrase, intonational phrase and sentence, totalling 806 and 797 dimensions respectively.

To conduct the experiments, 3 kinds of TTS synthesizers are to be compared, including HMM-based, monolingual BLSTM-based and multilingual BLSTM-based synthesizers.

In HMM-based TTS, each HMM phone model is five-state, left-to-right topology with single Gaussian, diagonal covariance distributions. 35 order MGC , V/UV flag and log-F0 with their delta and delta-delta are simultaneously modeled using the context-dependent HMMs. A decision tree-based context clustering algorithm is used to cluster HMM states using the minimum description length (MDL) [11] criterion. The parameters of HMM models are first trained using the maximum likelihood (ML) sense and then retrained by the minimum generation error (MGE) [12]. 500 utterances in the training set of English corpus were used to train the HMM models. For speech synthesis, a parameter generation module [13] is used to generate smooth feature parameters with dynamic feature, and then feed to the STRAIGHT [14] vocoder to generate synthetic speech.

In monolingual BLSTM based TTS, the training data is still 500 English utterances in the training set. The input feature vector includes 503 dimensions of phonetic and prosodic contextual information. The output feature vector contains a voiced/unvoiced (V/UV) flag, log-F0, MGC, totally 37dimesions. At synthesis stage, the output features predicted by the BLSTM directly serve as the acoustic parameters of the vocoder input.

For multilingual BLSTM based TTS, the 2,000 utterances of Mandarin (Cantonese) are used as the training data of language 1 (auxiliary language), while the 500 English utterances in the training set are used as the training data of language 2 (target language). In multilingual BLSTM, we use two bidirectional LSTM layers with 100 units for each layer to capture the long time span contextual effect of the training data. For BLSTM training, we use Rmsprop [15] to minimize the mean square error between the output features and groundtruth. Rmsprop is a form of stochastic gradient descent in which the gradient is normalized by the magnitude of recent gradients. It is robust by utilizing pseudo curvature information. It is also applicable of using mini batch learning for it can nicely handle stochastic objectives.

We use RNNLIB [16] as the implementation of our monolingual or multilingual BLSTM neural network, and HTS [17] for HMM-based speech synthesis.

4.2. Experimental results and analysis

To evaluate the performance of the proposed multilingual BLSTM, we conducted a set of objective and subjective evaluations on the 50 English utterances from the test set.

To verify that multilingual BLSTM trained with resource-rich and resource-limited training data can increase the prediction accuracy of the acoustic features of resource-limited language over the monolingual BLSTM trained using only the resource-limited speech data, following objective experiments were conducted. As the most popular SPSS method, the HMM-based method is also compared.

Table 2. *Objective evaluation results on HMM, monolingual, multilingual MAN-ENG and multilingual CAN-ENG BLSTM.*

Measures / Model	LSD (dB)	V/U Err rate (%)	F0 RMSE (Hz)
HMM	5.16	7.9	21.3
Monolingual	7.23	11.3	29.3
MAN-ENG	5.94	8.9	22.5
CAN-ENG	5.84	9.1	23.1

For objective evaluation, we calculate the distortions between the acoustic parameters of the original test utterances and the predicted parameters by different models (HMM, monolingual BLSTM or multilingual BLSTM). The distortions are calculated frame by frame and then averaged on all frames of the test set. For F0, root mean-square error (RMSE) is calculated. For voiced/unvoiced (V/UV), swapping error rate is counted. Normalized distance in log spectral distance (LSD) is computed for spectrum distortion.

The results of objective evaluations are shown in Table 2. For both language pairs, multilingual BLSTM outperforms monolingual BLSTM in predicting all acoustic parameters and such improvements are significant. This indicates that the proposed multilingual BLSTM can really learn the cross-lingual transfer information. Furthermore, different resource-rich auxiliary languages (Mandarin or Cantonese) with same amount of data can provide similar performance boost. However, multilingual BLSTM-based method still shows slightly worse performance than the traditional HMM-based method, which might suggest that more parameters in BLSTM should be learnt.

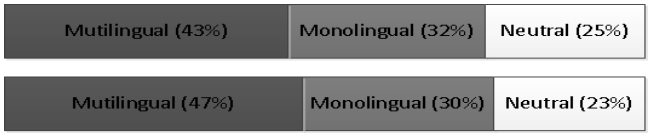

Fig. 5. Preference on Mandarin-English (above) and Cantonese-English (below) multilingual BLSTMs and monolingual BLSTM.

For subjective evaluation, we performed AB preference test between monolingual BLSTM English TTS and multilingual BLSTM Mandarin-English (or Cantonese-English) TTS. For each test sentence, we generated paired synthetic speeches from English monolingual BLSTM, Mandarin-English (or Cantonese-English) multilingual BLSTM synthesizers. Each pair of the synthetic speeches were randomly played in a sound-proof studio. Ten listeners were invited to listen to the synthetic speeches, and judge the quality of which speech is better. The listener can also select no preference if it is difficult to distinguish which one is better.

From the preference scores in Fig. 5, we can see that the quality of the synthetic speech generated by the multilingual BLSTM outperforms the quality of the synthetic speech by monolingual BLSTM, which validates the effectiveness of the proposed method

Table 3. *Objective evaluation results on different number of the training utterances.*

Measures / Utterances	LSD (dB)	V/U Err rate (%)	F0 RMSE (Hz)
250	6.89	9.7	25.9
200	7.29	11.2	28.5
150	8.17	12.1	30.2
120	9.23	12.9	33.5
100	9.81	13.5	35.2

For speech synthesis of low-resource languages, it would be valuable to find the least amount data for generating synthetic speech with satisfied speech quality. Based on the above experimental setup, we tried to reduce the amount of the training data of English in multilingual BLSTM (for Mandarin-English pair). We tried different settings of the number of the training utterances, from 500, 250, 200, 150, 120, 100. The objective measures of different settings are shown in Table 3. As can been see, when the number of training utterances is about 150 utterances (amount to 7 minutes), LSD is 8.17, voiced/unvoiced swapping errors is 12.1 and RMSE is 30.2. These measurements are closes to the performance of the monolingual BLSTM TTS engine trained with 500 utterances.

5. CONCLUSIONS

In this paper, we proposed a multilingual BLSTM that shares hidden layers across different languages. With this architecture, the cross-lingual information can be learned and benefit the training of a TTS synthesizer for low-resource or resource-limited languages. The experiments with two language pairs (Mandarin-English, Cantonese-English) validates the effectiveness of the proposed method. Both objective and subjective evaluations indicate that multilingual BLSTM can predict more accurate acoustic features for speech synthesis than the monolingual BLSTM. Future work will be focused on whether such multilingual BLSTM network can be used for cross-lingual speech synthesis.

6. ACKNOWLEDGEMENTS

This work is supported by National Basic Research Program of China (2012CB316401), National High Technology Research and Development Program of China (2015AA016305), National Natural Science Foundation of China (NSFC) (61375027, 61433018, 61370023, 61171116), joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15), Major Program for National Social Science Foundation of China (13&ZD189) and Hong Kong SAR Government General Research Fund (14205814).

7. 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. Zen, A. Senior, and Schuster M., "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[3] 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. ICASSP*, 2014, pp. 3829–3833.

[4] Y. Fan, Y. Qian, F. Xie, and F.K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. INTERSPEECH*, 2014, pp. 1964–1968.

[5] J. Yamagishi, T. Nose, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[6] Y. Fan, Y. Qian, F.K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[7] R. Caruana, *Multitask learning*, Springer, 1998.

[8] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, pp. 2673–2681, 1997.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] A. Graves, *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer, 1997.

[11] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *The Journal of the Acoustical Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[12] Y.J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, 2006, pp. 89–92.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[14] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[15] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[16] A. Graves, "RNNLIB: A recurrent neural network library for sequence learning problems," [OL] [2015-07-10], http://sourceforge.net/projects/rnnl/.

[17] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.