DBLSTM-based Multi-task Learning for Pitch Transformation in Voice Conversion

Runnan Li[1], *Zhiyong Wu*[1,2], *Helen Meng*[1,2], *Lianhong Cai*[1]

1 Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University

2 Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

lirn15@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

Abstract

While both spectral and prosody transformation are important for voice conversion (VC), traditional methods have focused on the conversion of spectral features with less emphasis on prosody transformation. This paper presents a novel pitch transformation method for VC. As the correlation of spectral features and fundamental frequency in pitch perceptions has been proved, well-converted spectrum should benefit to pitch transformation. Motivated by this, a multi-task learning (MTL) framework based on deep bidirectional long short-term memory (DBLSTM) recurrent neural network (RNN) has been proposed for pitch transformation in VC. DBLSTM is used to model the long short-term dependencies across speech frames for spectral conversion; the converted spectrum and the source pitch contour are further simultaneously modeled to generate the converted target pitch contour and voiced/unvoiced flag; the above tasks are incorporated with the MTL framework to enhance the performances of each other. Experimental results indicate the proposed method outperforms the conventional approaches in pitch transformation.

Index Terms: voice conversion, pitch transformation, deep bidirectional long short-term memory (DBLSTM), multi-task learning (MTL)

1. Introduction

Voice conversion (VC) is an important speech processing technique with the purpose to modify the characteristics of the speech uttered by a source speaker be perceived as if spoken by a specified target speaker while maintaining the speech content and naturalness [1]. From speech, the speaker identity can be characterized in two aspects: the voice quality that is determined by spectral content representing the physical attributes of the speaker; and the speaking style that is acoustically realized in prosodic features such as pitch contour, duration of words, rhythm, etc. Hence, a voice conversion system generally contains two parts: spectral transformation and prosody transformation [2].

Most of the existing voice conversion (VC) systems deal with the conversion of spectral features. Several statistical methods have been proposed to estimate a mapping function between the spectral features of the source and target speeches, including Gaussian mixture models (GMM), deep neural networks (DNN) and recurrent neural networks (RNN). GMM based method assumes the features have random distributions and the conversion can be achieved by maximizing conditional probability calculated from a joint probability of source and target features [3][4][5]. DNN derived method is based on learning processes using neuron structures and perform the conversion from source features to target ones directly in the non-linear manner, while RNN further extends DNN by taking into temporal correlations into account [6][7][8].

On the other hand, prosodic features such as pitch contours and speaking rhythm also contain important cues of speaker identity. It has been proved prosody information is beneficial to recognize speakers that are familiar to us. It should also be noted that there are correlations between the spectral features and the prosodic features. For example, the pitch range is divided into three 'registers' as proposed in [9], with each register associated with vocal fry, tense voice and falsetto respectively from the lowest range to the highest. While different kinds of spectral slopes are associated with different types of phonations, the co-variation between F0 and spectral cues should exist and has been proved in [10]. Nevertheless, in nowadays VC researches [6][7][8], such correlations have usually been ignored by performing spectral and prosody transformations separately. It is usually assumed relatively good results can be obtained with a simple statistical mean and variance scaling of F0 conversion methods [8][11], sometimes together with average speaking rate modification [12][13][14]. Although more advanced prosody conversion techniques have been proposed [15][16][17][18][19], how to utilize the correlations between the spectral features and the prosodic features to improve the prosody transformation performance still needs further investigation.

To address the above issue, this paper proposes a new pitch transformation method by virtue of deep bidirectional long short-term memory (DBLSTM) based multi-task learning (MTL) framework to model the correlations between spectral and prosodic features. To fully exploit the correlations, the method divides the pitch transformation task into three tasks: 1) the spectrum conversion task that converts the spectrum of the source speech to the target, 2) the pitch contour transformation task that accepts the converted spectrum and the source pitch contour to generate the transformed pitch contour, and 3) the voiced/unvoiced (V/UV) flag prediction task that predicts the V/UV flag of the converted speech from the source V/UV flag and converted pitch contour. The three tasks are incorporated in the MTL framework to enhance the performances of each other. DBLSTMs are used for each of the three tasks to model both preceding and succeeding long short-term temporal context dependencies [20][21] among speech frames.

2. Background Methodologies

The proposed method is based on deep bidirectional long short term memory (DBLSTM) recurrent neural network (RNN) and multi-task learning (MTL). This section give a brief review of the related background methodologies.

2.1. Deep Bidirectional Long Short-term Memory

For a given input sequence $\mathbf{x}=(x_1,x_2,\ldots,x_T)$, standard RNN computes the hidden vector sequence $\mathbf{h}=(h_1,h_2,\ldots,h_T)$ and the output vector sequence $\mathbf{y}=(y_1,y_2,\ldots,y_T)$ from $t=1$ to T according to the following equations:

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$y_t = W_{hy}h_t + b_y \quad (2)$$

where W_{xh}, W_{hh} and W_{hy} are the input-hidden, hidden-hidden and hidden-output weight matrices; b_h and b_y are the hidden and output bias vectors; and $\Phi(.)$ is the activation function in hidden layer. In long short-term memory (LSTM), $\Phi(.)$ can be implemented by LSTM block with the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t \tanh(c_t) \quad (7)$$

where σ is the logistic sigmoid function, while i, f, o and c are the *input gate, forget gate, output gate* and *memory cell* activation vectors respectively [21][22].

Since conventional RNNs process sequences in temporal order but ignoring the future context, bidirectional recurrent neural networks (BRNNs) is proposed to utilize contextual dependencies in both preceding and succeeding directions [23]. BRNNs process the data in both forward and backward directions with two separate hidden layers and then feed forward to the same output layer. Similar as the structure of deep neural network (DNN), multiple BRNNs hidden layers can be stacked on the top of each other to create deep BRNNs. Deep bidirectional long short-term memory (DBLSTMs) can be achieved by replacing the conventional activation function of hidden layers in deep BRNNs with LSTM cells [24].

2.2. Multi-task Learning

For several different but related tasks, multi-task learning (MTL) provides a way to train a universal model. The tasks involved in MTL generally consists of one main task and one or more secondary tasks. By combining related tasks, MTL can reduce overfitting of a specific task and make the learned representations universal across tasks. While the secondary task(s) are related to the main task and complementary at the same time, MTL is believed to be able to generalize better predictions than single-task learning model [25]. The use of MTL fashion in automatic speech recognition [26] and speech synthesis [27] have produced significant achievements.

2.3. Training Procedure

Backpropagation (BP) is the most widely used method in training artificial neural networks. However, conventional BP algorithm is not suitable for sequential model while it only accepts feed-forward inputs. Backpropagation through time (BPTT) [29][30], extended from conventional BP algorithm, provides the solution by unfolding RNNs into standard feed-forward networks through time steps. In conjunction with BPTT, mini-batch-based Adam [31] algorithm is used as the optimizer with Keras [32] deep learning framework to build and evaluate various model structures efficiently. As for the loss function, mean squared error (MSE) is adopted.

3. DBLSTM-based Multi-task Learning for Pitch Transformation

3.1. Pitch Transformation in Voice Conversion

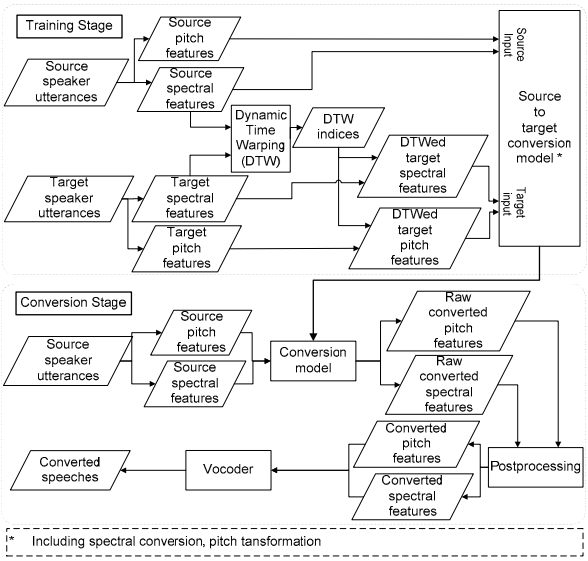

Figure 1: Overall architecture of voice conversion.

Fig. 1 depicts the overall architecture of voice conversion, in which pitch transformation is involved and is the main focus of this work. In the training stage, vocal tract related features (i.e. spectral) and vocal fold related features (i.e. pitch contour and V/UV) are extracted from the parallel source and target speaker utterances. Dynamic time warping (DTW) algorithm is then used to align the feature sequences of the source and target speakers. Each feature vector of source speaker is now mapped with a time aligned (DTWed) target feature vector. The source and target features are then normalized with z-score normalization and fed to train the conversion model that involves both spectral conversion and pitch transformation.

In the conversion stage, spectral and pitch features are extracted from source utterances and processed with the well trained conversion model to generate raw converted features. Maximum likelihood parameters generation (MLPG) [28] is employed as the postprocessing procedure to generate smooth speech parameter sequence (final converted features). The de-normalized spectral or pitch features are used as the mean of MLPG input probability density function (pdf) and the global variance of target features from all training data is used as the variance of the MLPG input pdf. The smoothed features are then sent to the vocoder to generate the converted speeches.

In conventional approaches, spectral conversion and pitch transformation are generally processed separately. While in this work, the two tasks are combined together in the MTL framework and the derived model is able to capture the correlations between spectral and pitch features.

3.2. Conventional Approaches

3.2.1. Linear and DNN based approaches

This work also implements linear transformation (LT) and DNN as the baseline approaches for comparison. As one of the most widely used pitch transformation method in VC, LT converts the source pitch contour (in log F0) by equalizing the mean and the standard deviation to target pitch contour:

$$y = \frac{V_t}{V_s}(x - M_s) + M_t \quad (8)$$

where **x**, **y** are the source pitch contour and converted pitch contour; *M, V* are the mean and standard deviation of source pitch contour and target pitch contour.

The DNN based model is developed as suggested by [20] which transfers the pitch contour frame by frame. The DNN based architecture consists of three stacked hidden layers and is trained by backpropagation algorithm using parallel vectors consisting of spectral features and pitch contour features.

3.2.2. DBLSTM based approach

The DBLSTM based model employed for both spectrum conversion and pitch transformation is illustrated in Fig. 2. As state-of-the-art in spectrum conversion, the DBLSTM based model accepts the combination of spectral features and pitch contours as the input vector **X** to predict the converted spectral features **Y**. Using the same structure, the DBLSTM based model is also developed to convert pitch features (i.e. pitch contour and V/UV).

It should be noted the two models for spectral and pitch conversion are generally trained separately in conventional voice conversion systems, which leads to two issues: 1) weak in exploiting the correlations between converted spectrum and pitch contour, 2) unable to make use of learned representations of spectral conversion during the learning process of pitch transformation even though these two tasks are related.

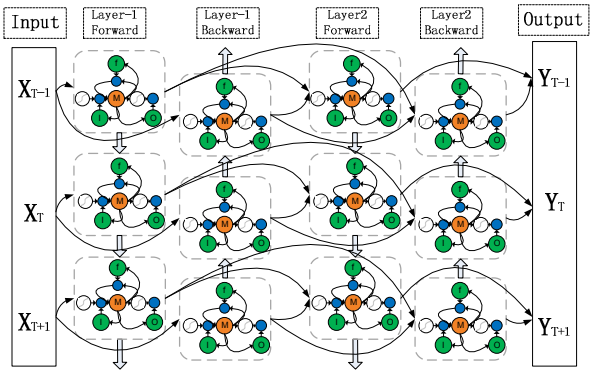

Figure 2: The basic DBLSTM based model.

3.3. DBLSTM-based Multi-Task Learning Model

To address the above issues, we propose a novel DBLSTM-based multi-task learning (MTL) model that can conduct pitch transformation and spectrum conversion simultaneously in the MTL framework. In the proposed model, we divide the transformation into three tasks: 1) the spectrum conversion task, 2) the pitch contour transformation task, and 3) the V/UV prediction task. Following the MTL fashion, we set the pitch contour transformation task as the main task and the other two tasks as secondary tasks. As shown in Fig. 3, three DBLSTM subnets are used: the first one (DBLSTM_1) accepts source spectral features and pitch contour as the input to perform spectrum conversion, the second one (DBLSTM_2) accepts the converted spectrum from DBLSTM_1 and source pitch contour as the input to conduct pitch transformation, while the last one (DBLSTM_3) accepts the converted pitch contour from DBLSTM_2 and source V/UV as the input to predict the voiced/unvoiced flag of converted speech. The three subnets are combined together to form a single entire MTL-DBLSTM model. As each task has its own loss function, during training stage, a single loss value is computed as weighted sum of the output of the three loss functions. In tradition, the loss weight of the main task is set higher than secondary tasks' to enhance the training performance of the main task.

By combining spectral and pitch conversion together, the proposed MTL-DBLSTM model can gain benefits by sharing the learned representations universal across tasks. In addition, the output of DBLSTM_1 (i.e. the converted spectrum) carries more inherent information about target speech. The correlation between the converted spectrum and target pitch contour can be fully exploited in DBLSTM_2. Similarly, V/UV prediction can also benefit from using converted pitch contour.

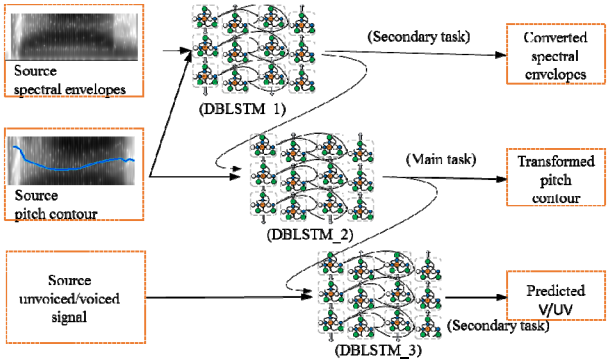

Figure 3: DBLSTM based multi-task learning model.

4. Experiments

4.1. Experimental Setup

The parallel speech database used for experiments is the CMU ARCTIC corpus [33]. While different speakers with different gender and accents have significant differences in pitch, we choose SLT (female from U.S) as the source speaker and AWB (male from Scotland) as the target speaker. The Mel-cepstral coefficients (MCEPs) and F0 (with V/UV flag) are extracted by STRAIGHT [34] and served as spectral and pitch features respectively. There are 1132 parallel utterances, from which the first 1000 parallel utterances are used as the training set, the next 100 utterances are used as the validation set and the rest 32 utterances are used as the test set.

The input of LSTM requests fixed length for the purpose to train the network with mini-batches, thus we extend feature sequence of each utterance to 1200 frames long with zero padding. Each frame has 27-dimentional features that consist of 25-dimensional MCEPs, one dimensional log F0 and one dimensional V/UV flag. Four different approaches for pitch transformation are compared:

- **LT**: baseline global linear transformation approach;
- **DNN**: DNN based approach;
- **DBLSTM**: basic DBLSTM based approach;
- **MTL-DBLSTM**: proposed DBLSTM based multi-task learning approach.

For LT approach, the mean and standard deviation used in pitch transformation is derived from source pitch features and time-aligned (DTWed) target features. The DNN employed has three hidden layers and the number of units in each hidden layer is [81 135 81] respectively. In DBLSTM approach, two hidden layers are stacked and each hidden layer consists of one forward LSTM layer and one backward LSTM layer with 100 LSTM blocks for each layer. In MTL-DBLSTM

approach, the DBLSTMs employ the same structure as in DBLSTM approach and have 100, 25 and 25 LSTM blocks in each hidden layer for spectrum conversion task, pitch contour transformation task and V/UV prediction task respectively. Specifically, the weights assigned to the loss produced by the three tasks are 0.2, 1.0 and 0.1 respectively.

This work focuses on evaluating pitch transformation only. To ensure the same condition of evaluation, the converted spectrum generated by DBLSTM based conversion approach is used for all the four approaches. STRAIGHT vocoder is employed to synthesize converted speech waveform using the transformed pitch features from different evaluated approaches and the aforementioned converted spectrum.

4.2. Experimental Results

To evaluate pitch transformation performances of different approaches, both objective and subjective experiments are conducted. The assessment utterances are converted from the 32 source utterances in the test set.

4.2.1. Objective evaluation

The average error (Euclidean distance) between converted and time aligned (DTWed) target pitch contours is used to evaluate the performance of pitch transformation with the following equation where $\tilde{f}_y$ and f_y represent converted F0 and target F0:

$$E(\tilde{f}_y, f_y) = \frac{1}{N}\sum_{n=1}^{N} d(\tilde{f}_y(n), f_y(n)) \tag{9}$$

The lower the average error, the better the performance.

The performance index is also used to evaluate pitch transformation performance, where f_x represents source F0:

$$I_F = 1 - \frac{E(\tilde{f}_y, f_y)}{E(f_x, f_y)} \tag{10}$$

Higher value of performance index signifies better conversion performance, with the max value be 1.

V/UV error is used to evaluate the performance of V/UV prediction. Lower V/UV error indicates better performance.

The objective evaluation results are shown in Table 1, with "default" being the direct use of original source pitch features.

Table 1. The average error, performance index and V/UV error of different approaches.

	Average error [Hz]	Performance index	V/UV error [%]
Default	63.1212	-	9.21
LT	18.0993	0.7132	9.21
DNN	17.1496	0.7283	5.82
DBLSTM	15.0724	0.7612	4.92
MTL-DBLSTM	14.0560	0.7693	4.53

4.2.2. Subjective evaluation

In subjective evaluation, ABX preference test is conducted[1]. 10 participants are asked to listen and choose which sample (A or B converted by different approaches) sounds more similar to the original target speaker's utterance (X). During test, X will be first played as reference, followed by A or B randomly. The participants can also choose no preference (N/P) if they cannot distinguish between A and B.

[1] Samples are available at http://219.223.175.6/wiki/rnli

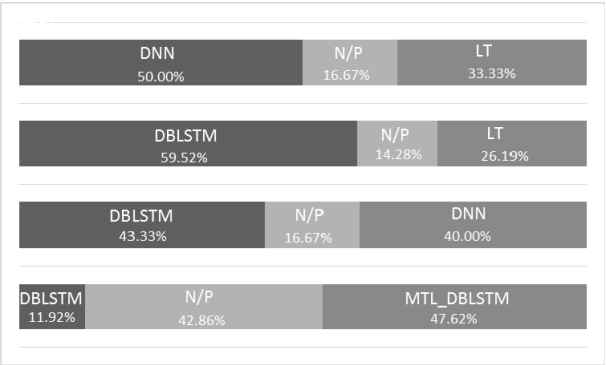

Figure 4: ABX preference test results.

4.2.3. Analysis and discussion

In objective evaluation, the results suggest the proposed MTL-DBLSTM approach outperforms conventional approaches in average error, performance index and V/UV error evaluations. From Fig. 4, DNN based approach outperforms LT approach in subjective similarity test. DBLSTM based approach shows close preferences with DNN based approach. Compared with the basic DBLSTM based approach, the proposed DBLSTM-based MTL (MTL-DBLSTM) approach achieves significantly better preferences. The results indicate the MTL-DBLSTM method achieves the best performance in pitch transformation.

In MTL-DBLSTM, the main task can be set to spectrum conversion. We set the weights of each task's loss to be 1.0, 0.1 and 0.1 to validate whether MTL fashion can bring further benefit to spectrum conversion. However, the result shows no significant difference between basic DBLSTM based approach and MTL-DBLSTM based approach. This infers the spectrum contributes to pitch conversion but not contrarily.

5. Conclusion

It has been approved there are correlations between spectral features and fundamental frequency in pitch perception of speech. To explore the possibility in capturing and modeling such correlations, this paper proposes a MTL-DBLSTM for pitch transformation by using deep bidirectional long short-term memory (DBLSTM) in conjunction with multi-task learning (MTL). In MTL-DBLSTM, the main task of pitch contour transformation is combined with two secondary tasks involving spectrum conversion and V/UV flag prediction. Moreover, the pitch contour transformation task also directly models the connections between the spectrum and the pitch contour with DBLSTM. Experimental results suggest that the proposed MTL-DBLSTM based pitch transformation method outperforms conventional approaches including LT, DNN and basic DBLSTM in both objective evaluation (average error, performance index and V/UV error) and subjective evaluation (ABX preference of voice conversion results). In the future, we will explore the possibilities to enroll more related acoustic feature prediction tasks as secondary tasks to generate more personalized and more nature speech.

6. Acknowledgement

This work is supported by Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15), National Social Science Foundation of China (13&ZD189) and NSFC (61375027, 61433018).

7. References

[1] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” [in] Proc. ICASSP, 2009.

[2] A. F. Machado, and M. Queiroz, “Voice conversion: A critical survey,” [in] Proc. Sound and Music Computing, pp. 1–8, 2010.

[3] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, “Multimodal voice conversion based on non-negative matrix factorization,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2015, no. 1, 2015.

[6] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Voice conversion in high-order eigen space using deep belief nets,” [in] Proc. INTERSPEECH, 2013.

[7] T. Nakashika, T. Takiguchi, and Y. Ariki, “High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion,” [in] Proc. INTERSPEECH, 2014.

[8] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” [in] Proc. ICASSP, 2015.

[9] B. Roubeau, N. Henrich, and M. Castellengo, “Laryngeal vibratory mechanisms: The notion of vocal register revisited,” Journal of Voice, vol. 23, pp. 425-438, 2009.

[10] J. Kuang, and M. Libermann, “The effect of spectral slope on pitch perception,” [in] Proc. INTERSPEECH, 2015.

[11] K. Yutani, Y. Utoi, Y. Nankaku, T. Toda, and K. Tokuda, “Simultaneous conversion of duration and spectrum based on statistical models including time-sequence matching,” [in] Proc. INTERSPEECH, pp. 1072-1075, 2008.

[12] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, “Spectral conversion based on statistical models including time-sequence matching,” [in] Proc. Speech Synthesis Workshop, pp. 333-338, 2007.

[13] A. R. Toth, and A. W. Black, “Incorporating durational modification in voice transformation,” [in] Proc. INTERSPEECH, pp. 1088-1091, 2008.

[14] R. Srikanth, B. Bajibabu, and K. Prahallad, “Duration modelling in voice conversion using artificial neural networks,” [in] Proc. IWSSIP, pp. 556-559, 2012.

[15] Z. Hanzlicek, and J. Matousek, “F0 transformation within the voice conversion framework,” [in] Proc. INTERSPEECH, pp. 1961-1964, 2007.

[16] A. Kunikoshi, Y. Qian, F. K. Soong, and N. Minematsu, “Improved F0 modeling and generation in voice conversion,” [in] Proc. ICASSP, 2011.

[17] B. Bollepalli, J. Beskow, and J. Gustafson, “Nonlinear pitch modification in voice conversion using artificial neural networks,” Advances in Nonlinear Speech Processing Lecture Notes in Computer Science, vol. 7911, pp. 97-103, 2013.

[18] F. Xie, Q. Yao, F. K. Soong, and H. Li, “Pitch transformation in neural network based voice conversion,” [in] Proc. ISCSLP, 2014

[19] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong and E. S. Chng, "High quality voice conversion using prosodic and high-resolution spectral features," Multimedia Tools and Applications, vol. 75, pp. 5265-5285, 2016.

[20] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997

[21] A. Graves, and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” Neural Networks, vol. 18, no. 5, pp. 602–610, 2005.

[22] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” Neural Computation, vol. 12, no. 10, pp. 2451–2471, 2000.

[23] M. Schuster, and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol. 45, pp. 2673–2681, Nov 1997.

[24] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” [in] Proc. ICASSP, 2013.

[25] R. Caruana, Multitask learning, Springer, 1998

[26] M. L. Seltzer, and J. Droppo, “Multi-task learning in deep neural networks for improved phoneme recognition,” [in] Proc. ICASSP, 2013.

[27] 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. ICASSP, 2015.

[28] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” [in] Proc. ICASSP, 2000.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Cognitive modeling, 1988.

[30] P. J. Werbos, “Backpropagation through time: what it does and how to do it,” [in] Proc. IEEE, vol. 78, no. 10, pp. 1550–1560, 1990.

[31] D. P. Kingma, and J. L. Ba, “Adam: A method for stochastic optimization,” [in] Proc. ICLR, 2015.

[32] F. Chollet, Keras [OL]. [2016-03-18]. GitHub repository. https://github.com/fchollet/keras.

[33] J. Kominek, and A. W. Black, “The CMU Arctic speech databases,” [in] Proc. Speech Synthesis Workshop, 2004.

[34] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds,” Speech communication, vol. 27, no. 3, pp. 187–207, 1999.