An Initial Research: Towards Accurate Pitch Extraction for Speech Synthesis Based on BLSTM

Yibin Zheng\textsuperscript{a,c}, Zhengqi Wen\textsuperscript{a}, Bin Liu\textsuperscript{a}, Ya Li\textsuperscript{a}, Jianhua Tao\textsuperscript{a,b,c}

\textsuperscript{a}National Laboratory of Pattern Recognition, \\
\textsuperscript{b}CAS Center for Excellence in Brain Science and Intelligence Technology, \\
Institute of Automation, Chinese Academy Sciences, 100190, Beijing, China \\
\textsuperscript{c}School of Computer and Control Engineering, University of Chinese Academy of Sciences \\
{yibin.zheng, zqwen, bin.liu, yli, jhtao}@nlpr.ia.ac.cn

Abstract—Accurate pitch extraction from speech is important but challenging problem for speech synthesis. However, the additive nature and long-term suprasegmental property of pitch features have not been fully exploited in most of the existing pitch estimators as they are operated frame by frame. As a result, they would cause some inherent discontinuities, such as double/half F0 errors and unvoiced/voiced (U/V) error. This would adversely affect the quality of synthetic speech as well as the expressiveness of the prosody information. In this paper, we explore the novel use of multi-tasks (Task 1: U/V; Task 2: Pitch) bidirectional long short-term memory recurrent neural network (BLSTM) to model the pitch and voicing decision simultaneously in a unified framework. The features used in this study are extracted from the frequency domain. We compute the log-frequency power spectrogram and then normalize to the long-term speech spectrum to attenuate noises. A filter is then used to enhance the harmonicity. Experiments show that the proposed approach substantially outperforms RAPT, which behaves the best in clean condition. Besides, our proposed approach can even work well with a certain level of background noise.

Keywords—pitch extraction, voicing decision, BLSTM, log-frequency power spectrogram, speech synthesis

I. INTRODUCTION

Pitch or fundamental frequency (F0), is one of the most important characteristics of speech signals. An accurate F0 extraction algorithm is critical to many applications, including speaker identification [1], speech separation [2], and especially for speech synthesis [3]. Although pitch extraction algorithm has been studied for decades, it is still challenging to estimate F0 from speech where the harmonic structure is not clear.

A pitch detector performs both pitch estimation and unvoiced/voiced (U/V) detection. In general, pitch estimation can be performed using (1) time-domain, (2) frequency-domain, (3) time-frequency-domain signal processing techniques, or (4) statistical based approach. Time-domain pitch estimation exploits the signal’s temporal periodicity by computing a temporal correlation or difference function directly from the signal samples. Some well-known examples of time-domain pitch estimation algorithms are RAPT [4], YIN [5], and AMDF [6], which are known to give accurate pitch estimates for clean speech. Frequency-domain pitch estimation relies on the presence of strong harmonic peaks near integer multiples of F0 in the short-time spectral representation. Some examples of such frequency-domain pitch estimation algorithms are subharmonic-to-harmonic ratio (SHR) [7], dominant harmonics [8] and SWIPE [9]. In time-frequency-domain pitch estimation algorithms, the input signal is typically decomposed into multiple frequency sub-bands, and time-domain techniques are applied on each sub-band signal. A popular time-frequency-domain technique is the auditory-model correlogram-based algorithm inspired by Licklde’s duplex theory of pitch perception [10], in which frequency decomposition is performed using an auditory filterbank [11], followed by autocorrelation (ACR) computation on each subband signal. In statistical based approaches, some models such as, HM-MGMM [12] and DNN [13] have been tried. A typical case of these is by quantizing F0 range into some different octaves and compute the posterior probability of each time step, followed by a viterbi decoding to generate the F0 contour [13].

As for U/V detection in a pitch detector, it can be performed by either utilizing the information derived from the pitch estimation module, or using a separate module that is independent of the pitch estimation algorithm. The simplest U/V detector is one that applies a constant decision threshold on a single degree-of-voicing feature computed by the pitch estimation module, e.g., ACR or cepstral peak amplitudes [14]. To further improve detection accuracy, some other approaches have also been employed [15-16].

However, most of these methods mentioned are based on the signal processing and operated frame by frame. Thus, the additive nature and long-term suprasegmental property of F0 features have not been fully exploited, leading to double/half F0 errors and unvoiced/voiced (U/V) error (Voicing is often very irregular at voice onset and offset leading to minimal wave-shape similarity in adjacent. This would adversely affect the quality of synthetic speech as well as the expressiveness of the rhythm information.

The aim of this work is to provide an accurate pitch estimation method for speech synthesis. We define the pitch pattern of an utterance as being represented by two components, namely pitch time series and voiced segments. Thus, the pitch pattern estimation task can be decomposed into two sub tasks, pitch estimation and voicing decision. Given that speech is inherently a sequential signal and temporal dynamics is crucial...
to pitch extraction, we adopt bidirectional long short-term memory recurrent neural network (BLSTM) to directly model the pitch and voicing decision simultaneously in a unified framework by multi-task learning from power spectral density (defined in Section 3-A).

The main contribution of this work is that the novel use of BLSTM for accurate pitch extraction for speech synthesis without any complex signal processing techniques and “post-processing” approaches. As we directly model the pitch value and U/V flag, the selection of pitch candidates can be avoided effectively (It is less optimal to make a hard decision for candidate selection). Besides, as BLSTM can take into account of the past and feature context information, double/half F0 errors and unvoiced/voiced (U/V) error can well be solved, and pitch trajectory could be captured effectively.

This paper is organized as follows. The next section relates a review of BLSTM. Section III discusses the feature extraction and proposed F0 extraction method. The experimental results and comparisons are presented in Section IV. We discuss related issues and conclude the paper in Section V.

II. BIDIRECTIONAL LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORK (BLSTM)

A recurrent neural network (RNN) is able to deal with the correlation 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 step. The feedforward process of RNN is:

\[ h_t = f(W_{hi}x_t + W_{hh}h_{t-1} + b_h) \]  \hspace{1cm} (1)

\[ y_t = o_t \tanh(c_t) \]  \hspace{1cm} (2)

\[ x = (x_1, x_2, ..., x_p) \] is the input vector sequence, \( h = (h_1, h_2, ..., h_p) \) is the hidden state vector sequence computed from input vector sequence, and \( y = (y_1, y_2, ..., y_p) \) is the output vector sequence. \( f \) is the activation function for hidden state. \( W_{hi}, W_{hh}, W_{ho} \) represent the input-hidden, hidden-hidden and hidden-output weight matrices respectively. \( b_h \) and \( b_o \) denote the bias vectors for hidden state vectors and output vectors.

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 [17], as shown in Fig.1. The feedforward process of BRNN include forward \( h_f \), and backward hidden sequence \( h_b \).

In pitch estimation task, the long time span contexts in an speech utterance need to be modeled. However, the RNN and BRNN structure is only able to retain short term memory because of the vanishing gradient problem. Long short term memory (LSTM) [18] 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:

\[ i_t = \sigma(W_{ii}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \]  \hspace{1cm} (3)

\[ f_t = \sigma(W_{if}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \]  \hspace{1cm} (4)

\[ c_t = f_t c_{t-1} + i_t \tanh(W_{ic}x_t + W_{hc}h_{t-1} + b_c) \]  \hspace{1cm} (5)

\[ o_t = \sigma(W_{io}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \]  \hspace{1cm} (6)

\[ h_t = o_t \tanh(c_t) \]  \hspace{1cm} (7)

where \( \sigma \) is a logistic; \( i, f, o \) and \( c \) respectively represent input gate, forget gate, output gate and cell memory [19].

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.

III. PROPOSED APPROACH

The proposed F0 extraction algorithms first extract spectral domain features in each frame, and then employ neural networks to capture the relationship between frequency domain features and F0, U/V flag.

A. Feature extraction

The features used in this study are extracted from the spectral domain based on [20]. We compute the log-frequency
power spectrogram and then normalize to the long-term speech spectrum to attenuate noises. A filter is then used to enhance the harmonicity.

Specifically, a signal is first decomposed to the spectral domain using short time Fourier transformation. Let $X_t(f)$ denote the power spectral density (PSD) of the frame $t$ in the frequency bin $f$. The PSD in the log-frequency domain can be represented as $X_t(q)$, where $q = \log f$. Then, the normalized PSD can be computed as:

$$X'_t(q) = X_t(q) \frac{L(q)}{X_t(q)}$$  \hspace{1cm} (8)

where $L(q)$ represents the long-term average speech spectrum, and $X_t(q)$ denotes the smoothed averaged spectrum of speech, which is calculated by using a 21-point moving average filter in the log-frequency domain and averaging over the entire sentence (2-4 s duration) in the time domain in this study. With the normalized spectrum, we further enhance harmonicity for pitch tracking using a filter with broadened peaks having an impulse response defined as:

$$h(q) = \begin{cases} 
\frac{1}{\gamma - \cos(2\pi q)} - \beta & \text{if } \log(0.5) < q < \log(K + 0.5) \\
0 & \text{otherwise}
\end{cases}$$ \hspace{1cm} (9)

where $\beta$ is chosen so that $\int h(q)dq = 0$, and $\gamma$ controls the peak width which is set to 1.8.

The convolution $\tilde{x}_t(q) = X'_t(q) \ast h(q)$ contains peaks corresponding to harmonics and their multiples and submultiples. Only the spectral components in the plausible pitch frequency range (50 to 450 Hz in this study) are selected as features. So we have a spectral feature vector in frame :

$$\tilde{x}_t = (\tilde{x}_t(q_1), \ldots, \tilde{x}_t(q_d))^T$$ \hspace{1cm} (10)

Gonzalez and Brookes [20] proposed to extract the spectral feature $\tilde{x}_t$ for pitch extraction in noise. Ideally, the pitch, can be found by taking the highest peak in $\tilde{x}_t$. In [20], several highest peaks are chosen for each frame as pitch candidates, and a dynamic programming algorithm is then used to form pitch contours. Although the feature vector is designed to deal with noisy speech, pitch candidate selection may lose useful information because it simply ignores non-peak spectral information. In our study, we treat $\tilde{x}_t$ as the extracted feature and employ supervised learning to estimate pitch, i.e. to learn the mapping from the features to the pitch frequencies. We expect supervised learning to yield better results.

Since neighboring frames contain useful information for pitch tracking, we incorporate the neighboring frames into the feature vector. Therefore, the final frame-level feature vector is

$$x_t = (\tilde{x}_{t-d}, \ldots, \tilde{x}_{t+d})^T$$ \hspace{1cm} (11)

where $d$ is set to 3 in our study.

B. Multi-task BLSTM model

Instead of selecting pitch candidates, we employ supervised training approach to learn the F0 and U/V series according to the features in each frame by multi-task learning. Neural networks have recently achieved large progress in speech processing, and we propose multi-task BLSTM to model such relationship. Fig. 3 shows the processing flow of the proposed method. It is summarized as follows:

1. The input signal $s(t)$ is first framed by a short time analysis window of a fixed length of 25ms with a fixed window shift of 5ms.
2. The enhance harmonicity feature vector is then extracted by the approach introduced in Section 3-A.
3. The extracted features are then fed into the proposed multi-task BLSTM to model the mapping between frequency domain features and F0, U/V flag. F0 estimation and voicing decision are treated as two different regression tasks in this work. For all experiments in this paper, different weights vary from 0 to 1 are tried for these two sub-tasks’ loss function and the best performance among them is selected as the final result.
4. A voicing decision threshold is chosen in order to minimize the U/V error in the validation set. Then F0 contour can be generated according to the F0 and U/V series.

IV. EXPERIMENTAL RESULTS AND COMPARISONS

We evaluate the performance for the proposed approach using our recording database (which is used for speech synthesis). The database contains 1000 utterances spoken by 10 male speakers and 10 female speakers. For each speaker in the database, 950 sentences are used for training, while the remaining 50 sentences are used for testing. Ground truth F0 series for training set are extracted by RAPT [4], while ground truth F0 series for test set are extracted by RAPT [4] and then
professional annotators by observing the ground truth. The number of frames in the utterance, \( N \), indicates the percentage of frames are misclassified in terms of voicing:

\[
GPE = \frac{N_{\text{u} \rightarrow \text{v}}}{N} \times 100\%
\]

\[
VDE = \frac{N_{\text{u} \rightarrow \text{v}} + N_{\text{v} \rightarrow \text{u}}}{N} \times 100\%
\]

where \( N \) is the number of the frames in the utterance, \( N_{\text{u} \rightarrow \text{v}} \) and \( N_{\text{v} \rightarrow \text{u}} \) are the number of frames misclassified as unvoiced and voiced, respectively. \( N_{\text{v}} \) is the number of voiced frames in ground truth, and \( N_{0,20} \) is the number of frames in which F0 deviation is smaller than 20% of the ground truth frequency.

For the proposed multi-tasks BLSTM model, a 3-layer neural network consisting a single non-recurrent layer, followed by 2 stacks of bidirectional hidden shared layers (each with 256*2 LSTM hidden units) is used. All networks are trained with a momentum of 0.9, an initial learning of 0.001 for the first 10 epoch, and then decreases by 20% after each epoch.

### A. Feature Selection

**TABLE I. PERFORMANCE OF DIFFERENT PARAMETER**

<table>
<thead>
<tr>
<th>Measures</th>
<th>GPE %</th>
<th>VDE %</th>
</tr>
</thead>
<tbody>
<tr>
<td>original PSD</td>
<td>2.91</td>
<td>1.42</td>
</tr>
<tr>
<td>normalized PSD</td>
<td>2.44</td>
<td>1.29</td>
</tr>
<tr>
<td>filtered normalized PSD</td>
<td>0.98</td>
<td>0.56</td>
</tr>
</tbody>
</table>

In this study, we first compute the PSD \( X(q) \). The normalized spectral in the log-frequency domain, are then used to generate the normalized PSD \( \hat{X}(q) \). The normalized spectral features are then convolved with a filter with a broadened impulse response, resulting the final features used in our study \( \hat{X}(q) \). To reveal feature effects, we train three BLSTM models using different features. As shown in Tab. I (which is tested on clean speech), the filtered normalized PSD achieves the best performance, and the normalized PSD and the original PSD achieve comparable performance. Therefore, the filtered normalized PSD is employed for following experiments.

### B. Extension to noise condition

We compare our approach with RAPT [4], which achieves the best GPE performance in clean condition that has been reported [21]. Tab. II and III shows the performance of the test sets in white and volvo conditions, respectively. It can be seen from Tab. II and III that in clean condition, multi-task BLSTM achieves superior performance than RAPT in all two objective measures. This is quite meaningful for speech synthesis as the significance of accurate pitch estimation from clean speech for speech synthesis. We further test our approaches on noise speech to validate its robustness to noise. Though we don’t add any noise in our training corpus, our approach still behaves well in noise condition, especially in lower SNR (-5 dB) condition, our approach behaves much better than RAPT. As we our motivation is to provide an accurate pitch predictor for speech synthesis in this paper, we don’t compare with other pitch estimation algorithms which are designed for noise speech.

<table>
<thead>
<tr>
<th>SNR</th>
<th>GPR %</th>
<th>VDE %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>52.20</td>
<td>7.14</td>
</tr>
<tr>
<td>10</td>
<td>14.73</td>
<td>5.44</td>
</tr>
<tr>
<td>20</td>
<td>27.41</td>
<td>4.94</td>
</tr>
</tbody>
</table>

**TABLE II. PERFORMANCE IN WHITE NOISE CONDITION**

<table>
<thead>
<tr>
<th>SNR</th>
<th>GPR %</th>
<th>VDE %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>30.28</td>
<td>5.25</td>
</tr>
<tr>
<td>10</td>
<td>3.98</td>
<td>0.87</td>
</tr>
<tr>
<td>20</td>
<td>2.84</td>
<td>0.56</td>
</tr>
</tbody>
</table>

**TABLE III. PERFORMANCE IN VOLVO NOISE CONDITION**

<table>
<thead>
<tr>
<th>SNR</th>
<th>GPR %</th>
<th>VDE %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>2.55</td>
<td>0.56</td>
</tr>
<tr>
<td>10</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.56</td>
<td></td>
</tr>
</tbody>
</table>

### C. Speech reconstruction

One of the motivations for predicting the voicing and pitch of speech in this paper is to obtain high quality of speech analysis-synthesis. This section reports the results of informal listening tests and compares spectrograms that synthesized using the RAPT and BLSTM to predict the pitch, while line spectral pair (LSP) coefficients is kept the same during analysis-synthesis. The subjective ABX test includes 11 listeners, who compared 20 sentence pairs randomly chosen from test set (in clean condition). Test result, given in Fig. 4 shows that, multi-BLSTM achieves superior performance than RAPT. This can be explained when we compare the spectrograms between the sentence pairs synthesized from RAPT and BLSTM. Fig. 5 (marked in the black boxes) shows the spectrogram of the utterance “緌子是指帽子或旗杆上的繚子 (Rui Zi refers to a hat or a tassel on flagpole).”. The formants are seen to be well preserved in the reconstructed speech using BLSTM. Fig. 6 shows pitch series of corresponding utterance extracted by RAPT and BLSTM. The orange boxes in Fig. 6 indicates that BLSTM can alleviate the double/half F0 errors and unvoiced/voiced (U/V) error in a way, thus leading to the results in subjective test.
In this paper, we explore the novel use of multi-tasks bidirectional long short-term memory recurrent neural network (BLSTM) to model the pitch and voicing decision simultaneously in a unified framework. The features used in this study are extracted from the frequency domain. We compute the log-frequency power spectrum and then normalize to the long-term speech spectrum to attenuate noises. We evaluate our system on the corpus used for speech synthesis. Experiments show that our approach substantially outperforms RAPT, which achieves the best GPE performance in clean condition that has been reported. We conduct speech reconstruction experiments to show that our approach can alleviate the double/half F0 errors and unvoiced/voiced (U/V) error than RAPT in a way, which is useful and important for speech synthesis. Besides, our proposed approach can even work well with a certain level of background noise. Our future research will focus on trying more features and adding some noises to the training set to boost the performance of multi-tasks BLSTM in noise condition.

Fig. 4. ABX test results for Multi-task BLSTM and RAPT.

Fig. 5. Comparison of speech for the utterance “緌子是指帽子或旗杆上的缨子(Rui Zi refers to a hat or a tassel on a flagpole.)” for (a) Original, (b) reconstructed using RAPT, and (c) reconstructed using multi-task BLSTM.

V. CONCLUSIONS

ACKNOWLEDGMENT

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61305003, No.61425017, No.61403386), the Strategic Priority Research Program of the CAS (Grant XDB02080006), and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES


