

Article

Sequence-to-Sequence Acoustic Modeling with Semi-Stepwise Monotonic Attention for Speech Synthesis

Xiao Zhou ¹, Zhenhua Ling ^{1,*}, Yajun Hu ² and Lirong Dai ¹

¹ National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230026, China; xiaozh@mail.ustc.edu.cn (X.Z.); lrdai@ustc.edu.cn (L.D.)

² iFLYTEK Research, Hefei 230088, China; yjhu@iflytek.com

* Correspondence: zhling@ustc.edu.cn

Abstract: An encoder–decoder with attention has become a popular method to achieve sequence-to-sequence (Seq2Seq) acoustic modeling for speech synthesis. To improve the robustness of the attention mechanism, methods utilizing the monotonic alignment between phone sequences and acoustic feature sequences have been proposed, such as stepwise monotonic attention (SMA). However, the phone sequences derived by grapheme-to-phoneme (G2P) conversion may not contain the pauses at the phrase boundaries in utterances, which challenges the assumption of strictly stepwise alignment in SMA. Therefore, this paper proposes to insert hidden states into phone sequences to deal with the situation that pauses are not provided explicitly, and designs a semi-stepwise monotonic attention (SSMA) to model these inserted hidden states. In this method, hidden states are introduced that absorb the pause segments in utterances in an unsupervised way. Thus, the attention at each decoding frame has three options, moving forward to the next phone, staying at the same phone, or jumping to a hidden state. Experimental results show that SSMA can achieve better naturalness of synthetic speech than SMA when phrase boundaries are not available. Moreover, the pause positions derived from the alignment paths of SSMA matched the manually labeled phrase boundaries quite well.

Keywords: speech synthesis; sequence-to-sequence; attention; phrase boundary



Citation: Zhou, X.; Ling, Z.; Hu, Y.; Dai, L. Sequence-to-Sequence Acoustic Modeling with Semi-Stepwise Monotonic Attention for Speech Synthesis. *Appl. Sci.* **2021**, *11*, 10475. <https://doi.org/10.3390/app112110475>

Academic Editor: Javier Hernandez

Received: 20 August 2021

Accepted: 25 October 2021

Published: 8 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Statistical parametric speech synthesis (SPSS) [1] is a mainstream approach to speech synthesis currently. It consists of three main components: text analysis, acoustic modeling, and waveform reconstruction. Text analysis [2] extracts linguistic features, such as phone transcriptions and prosodic structures, from input texts. Acoustic modeling aims to represent the mapping relationship between linguistic and acoustic features using statistical models [3]. Vcoders [4,5] are utilized to reconstruct speech waveforms from the predicted acoustic features at the synthesis time. Recently, neural-network-based sequence-to-sequence (Seq2Seq) acoustic models such as Tacotron [6] and Tacotron2 [7], and neural vocoders such as WaveNet [5] have been proposed and improved the naturalness of SPSS significantly.

The attention mechanism imitates the human brain. For example, human vision can quickly scan the image to obtain the target area that needs to be focused on, and then put more attention on this area to obtain more detailed information about the target. The Seq2Seq acoustic models also uses an attention mechanism to bridge the encoder and decoder, then the decoder pays attention to different parts of the input text to generate the corresponding acoustic features of each frame. One issue with the original Tacotron is that its attention mechanism is not robust enough, which may lead to errors in predicted acoustic features, such as repeating, skipping, and attention collapse. Repeating refers to the fact that too much attention stays on a certain input, resulting in a stuttering feeling in the synthetic speech. Skipping refers to the fact that too little attention stays on a certain input, resulting in missing words in the synthetic speech. Attention collapse refers to the

fact that the attention mechanism does not know which part in the text to pay attention to, resulting in unintelligible synthetic speech. One approach to alleviate this issue is to modify the attention mechanism utilizing the monotonic property of the alignment between phone sequences and acoustic feature sequences. Some improved attention techniques, such as forward attention [8] and stepwise monotonic attention (SMA) [9], have been proposed. In SMA, alignment paths were constrained to be strictly stepwise and monotonic, which meant that the attention at each decoding step can only choose staying at the same phone or to moving forward to the next phone, without moving backward and skipping. This strategy improved the robustness of Seq2Seq speech synthesis effectively.

On the other hand, phrase boundaries [10] are important prosodic labels for speech synthesis and they are usually indicated by pauses in continuous speech. Figure 1 shows the waveform and the aligned transcription for an example sentence. We can see that each phrase boundary (<pb>) corresponds to a short pause (sp) in this sentence. However, the phone sequences derived by grapheme-to-phoneme (G2P) conversion [11] may not contain the pauses at the phrase boundaries in utterances, considering the costs of labeling phrase boundaries at the training stage and predicting them at the synthesis time, especially for low-resource languages. The lack of explicit pause positions challenges the assumption of strictly stepwise alignment in SMA and may degrade the quality of synthetic speech when phrase boundaries are not available.

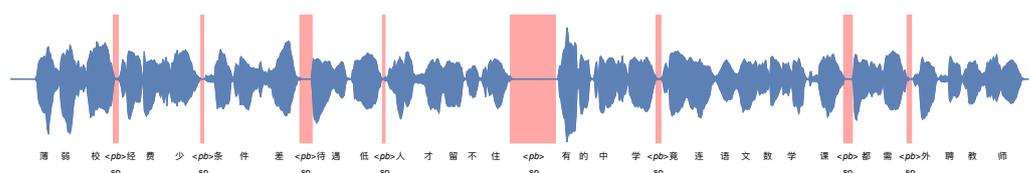


Figure 1. The waveform and aligned transcription for an example sentence, where <pb> means phrase boundary. The pink segments indicate the positions of short pauses (sp) in the waveforms. The English translation of this sentence is “The weak schools with low funding, poor conditions and low salaries are unable to retain talents. Some middle schools even need external teachers for Chinese and mathematics classes”.

Therefore, this paper proposes to insert hidden states into phone sequences to deal with the situation that pauses are not provided explicitly, and designs a semi-stepwise monotonic attention (SSMA) to model these inserted hidden states as the standard SMA cannot handle these states very well. In this method, hidden states are employed to absorb the pause segments in utterances using an unsupervised way. In comparison with SMA, the attention of SSMA at each decoding frame has three options for the next frame, including moving forward to the next phone, staying at the same phone, or jumping to the hidden state. Experimental results show that SSMA outperformed SMA in both objective and subjective evaluations when phrase boundaries are not given. Furthermore, the F1 score between the pause positions derived from the alignment paths of SSMA and the manually labeled phrase boundaries was 56.22%, which demonstrated the ability of SSMA on learning phrase boundaries without supervision.

The paper is organized as follows. Section 2 briefly review the existing stepwise monotonic attention mechanism. Section 3 introduces our proposed method. Sections 4 and 5 are experimental results and conclusions.

2. Related Work

The Tacotron [6] model unified acoustic modeling and duration modeling in a single model, and adopted a simple additive attention mechanism [12] to calculate the attention weights by query and keys. Further, Tacotron2 [7] used a location-sensitive attention mechanism [13], but there were still alignment errors, especially for out-of-domain texts. The alignment errors led to robustness problems in the predicted acoustic features, such as repeating, skipping, and failing to stop. To alleviate these problems, a number of

methods have been proposed, including non-autoregressive acoustic models with explicit phone duration modeling [14,15] and improved attention mechanisms, such as forward attention [8], stepwise monotonic attention (SMA) [9] and location-relative attention [16].

Among them, SMA [9] applied strictly stepwise and monotonic constraints to alignment paths. Its mechanism is described in Algorithm 1, where W , V , U , v and b are trainable weights, G denotes weights of convolution kernels, σ is sigmoid function, θ is the Heaviside step function, and γ is a trainable weight to control the strength of Gaussian noise.

Algorithm 1: Stepwise Monotonic Attention (SMA).

Input: query vector q_t , key vectors $K = \{k_1, k_2, \dots, k_N\}$, previous attention weights a_{t-1} , $mode \in \{hard, soft\}$

Output: attention weights $a_t = \{a_{t1}, a_{t2}, \dots, a_{tN}\}$, context vector c_t

if $t = 1$ **then**

$a_{t1} \leftarrow 1; \{a_{t2}, a_{t3}, \dots, a_{tN}\} \leftarrow 0;$

else

$F = \{f_1, f_2, \dots, f_N\} \leftarrow G * a_{i-1};$

 // * means convolution operation.

for $n \leftarrow 1$ **to** N **do**

$e_{t,n} \leftarrow v^T \tanh(Wq_t + Vk_n + Uf_n) + b;$

if $mode = soft$ **then**

$p_{t,n} \leftarrow \sigma(e_{t,n} + \gamma \mathcal{N}(0, 1));$

else if $mode = hard$ **then**

$p_{t,n} \leftarrow \theta(e_{t,n});$

end

end

$a_t \leftarrow a_{t-1} \cdot p_t + [0; a_{t-1,1:N-1} \cdot (1 - p_{t,1:N-1})];$

 // \cdot means element-wise product.

 // $1:N-1$ means the index range.

 // a_t are normalized.

end

$c_t \leftarrow \sum_{n=1}^N a_{t,n} k_n;$

return $a_t, c_t;$

At the first decoding step, the attention weights are manually set as one for the first phone, and zero for the rest phones considering the phone sequence is monotonically aligned with the acoustic feature sequence. Starting from the second decoding step, the attention weights are calculated recursively. The SMA calculates energy value $e_{t,n}$ based on query value q_t in the decoder, key values $K = \{k_1, k_2, \dots, k_N\}$ in the encoder outputs and location features F . Then “selection probability” $p_{t,n}$ based on energy value is computed by a sigmoid function. To sharpen the probability of the output, SMA adds Gaussian noise to energy values before feeding them into the sigmoid function. In order to achieve a monotonic and non-skipping attention mechanism, the attention weights a_t are calculated based on the previous attention weights a_{t-1} and “selection probability” p_t as shown in Algorithm 1. SMA adopted probability-based soft alignment at the training stage, and can choose between soft and hard alignments at the synthesis stage.

3. Proposed Methods

We propose to insert a hidden state between every two adjacent phones. To deal with these inserted hidden states, we propose a semi-stepwise monotonic attention (SSMA), as shown in Figure 2. The SSMA mechanism is a modification of the SMA mechanism, that can deal with the situation that there are missing pause labels in phone sequences due to the lack of explicit phrase boundaries. These hidden states are expected to absorb the decoding frames that correspond to pauses. We name it “semi-stepwise” because these

hidden states are skippable if there is no pause between two adjacent phones, thus the alignment paths are not strictly stepwise.

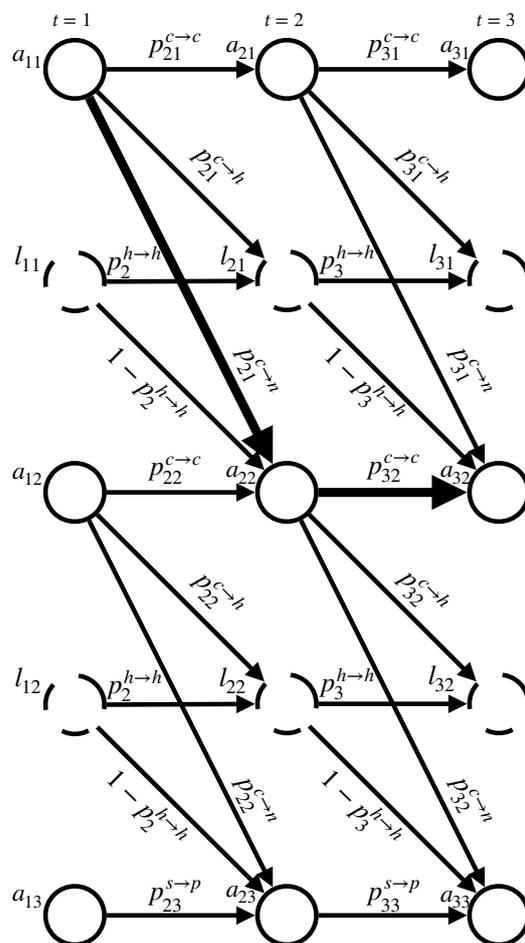


Figure 2. Schematic diagram of SSMA with $t \in \{1, 2, 3\}$ and $n \in \{1, 2\}$. The dashed circles denote hidden states at each decoding step. The definitions of the symbols in this figure can be found in Section 3.

At the t -th decoding step, a_{tn} denotes the attention weight of the n -th phone, and l_{tn} denotes the attention weight of the hidden state after the n -th phone. We have $\sum_n (a_{tn} + l_{tn}) = 1$. As shown in Figure 2, the attention at the n -th phone of the $(t - 1)$ -th decoding step has three choices to derive the attention of the t -th decoding step, staying at current phone, jumping forward to the next phone, or jumping to the hidden state after current phone. Let $p_{t,n}^{c \rightarrow c}$, $p_{t,n}^{c \rightarrow n}$ and $p_{t,n}^{c \rightarrow h}$ denote the probability of these three choices, respectively, and their sum should be 1. Moreover, the attention at a hidden state of the $(t - 1)$ -th decoding step has two choices to derive the attention of the t -th decoding step, staying at current hidden state or jumping to the next phone. Let $p_t^{h \rightarrow h}$ and $1 - p_t^{h \rightarrow h}$ denote the probability of these two choices, respectively.

Similar to Algorithm 1 for SMA, the attention weights $\mathbf{a}_t = \{a_{t1}, a_{t2}, \dots, a_{tn}\}$ and $\mathbf{l}_t = \{l_{t1}, l_{t2}, \dots, l_{tN-1}\}$ are calculated in a recursive way. The detailed pseudo-code is shown in Algorithm 2, where $\mathbf{W}, \mathbf{V}, \mathbf{v}, b, \mathbf{W}^l, \mathbf{V}^l, \mathbf{v}^l, b^l, \mathbf{U}, \mathbf{V}'$ and \mathbf{b}' are trainable weights.

When $t = 1$, the attention weights are manually set as one for the first phone, and zero for the rest phones and hidden states. Considering that a phone at the t -th decoding step has three options to choose the attention for the next decoding step, a two-level prediction strategy based on location-sensitive attention [13] is adopted to calculate the probabilities of these options. Specifically, we first predict the staying probability $p_t^{c \rightarrow c}$ and then predict the proportion of $p_t^{c \rightarrow n} / (p_t^{c \rightarrow n} + p_t^{c \rightarrow h})$, where $./$ denotes element-wise division. In order

to enhance the query effectiveness, a query matrix q'_t is calculated using a deep neural network (DNN). The DNN model accepts three inputs, i.e., the query q_t , the keys K and the location vectors F . The probability $p_{t,n}^{c \rightarrow c}$ is calculated by DNN using $q'_{t,n}$ and k_n . The probability $p_{t,n}^{c \rightarrow n} / (p_{t,n}^{c \rightarrow n} + p_{t,n}^{c \rightarrow h})$ is calculated by the DNN using $q'_{t,n}$ and k_{n+1} . An end-of-sentence (EOS) embedding vector is taught to replace k_{n+1} when n is index of the last phone. Then, $p_t^{c \rightarrow n}$ and $p_t^{c \rightarrow h}$ can be derived from $p_t^{c \rightarrow c}$ and $p_t^{c \rightarrow n} / (p_t^{c \rightarrow n} + p_t^{c \rightarrow h})$. In our preliminary experiments, we found that this strategy can obtain higher naturalness of synthetic speech than using Gumbel-Softmax [17]. To calculate the staying probability on hidden states $p_t^{h \rightarrow h}$, we simply use a DNN. The DNN model accepts two inputs, i.e., the query vector q_t and the hidden state embedding k_l . The DNN first predicts e_t^l , and then the trick of adding-noise and a sigmoid function are applied again to calculate $p_t^{h \rightarrow h}$. Finally, the attention weights a_t and l_t are calculated recursively and the context vector c_t is updated accordingly.

Algorithm 2: Semi-Stepwise Monotonic Attention.

Input: query vector q_t , key vectors $K = \{k_1, k_2, \dots, k_N\}$, hidden state embedding vector k_l , previous attention weights a_{t-1} and l_{t-1} , $mode \in \{hard, soft\}$

Output: attention weights $a_t = \{a_{t1}, a_{t2}, \dots, a_{tN}\}$ and $l_t = \{l_{t1}, l_{t2}, \dots, l_{t,N-1}\}$, context vector c_t

if $t = 1$ **then**
 | $a_{t1} \leftarrow 1; \{l_{t1}, a_{t2}, l_{t2}, \dots, a_{tN}\} \leftarrow 0;$

else
 | $F = \{f_1, f_2, \dots, f_N\} \leftarrow G * a_{t-1};$
 | $e_t^l \leftarrow (v^l)^\top \tanh(W^l q_t + V^l k_l) + b^l;$
 | **for** $n \leftarrow 1$ **to** N **do**
 | | $q'_{t,n} = V'^\top \tanh(W q_t + V k_n + U f_n) + b';$
 | | $e_{t,n}^s \leftarrow v^\top \tanh(q'_{t,n} + V k_n) + b;$
 | | $e_{t,n}^j \leftarrow v^\top \tanh(q'_{t,n} + V k_{n+1}) + b;$
 | | **if** $mode = soft$ **then**
 | | | $p_{t,n}^{c \rightarrow c} \leftarrow \sigma(e_{t,n}^s + \gamma \mathcal{N}(0, 1));$
 | | | $p_{t,n}^{c \rightarrow n} \leftarrow (1 - p_{t,n}^{c \rightarrow c}) \sigma(e_{t,n}^j + \gamma \mathcal{N}(0, 1));$
 | | | $p_{t,n}^{c \rightarrow h} \leftarrow (1 - p_{t,n}^{c \rightarrow c}) (1 - \sigma(e_{t,n}^j + \gamma \mathcal{N}(0, 1)));$
 | | | $p_t^{h \rightarrow h} \leftarrow \sigma(e_t^l + \gamma \mathcal{N}(0, 1));$
 | | **else if** $mode = hard$ **then**
 | | | $p_{t,n}^{c \rightarrow c} \leftarrow \sigma(e_{t,n}^s);$
 | | | $p_{t,n}^{c \rightarrow n} \leftarrow (1 - p_{t,n}^{c \rightarrow c}) \sigma(e_{t,n}^j);$
 | | | $p_{t,n}^{c \rightarrow h} \leftarrow (1 - p_{t,n}^{c \rightarrow c}) (1 - \sigma(e_{t,n}^j));$
 | | | $\{p_{t,n}^{c \rightarrow c}, p_{t,n}^{c \rightarrow n}, p_{t,n}^{c \rightarrow h}\} \leftarrow \text{binary}(p_{t,n}^{c \rightarrow c}, p_{t,n}^{c \rightarrow n}, p_{t,n}^{c \rightarrow h});$
 | | | $p_t^{h \rightarrow h} \leftarrow \theta(e_t^l);$
 | | **end**
 | **end**
 | $a_t \leftarrow a_{t-1} \cdot p_t^{c \rightarrow c} + [0; a_{t-1,1:N-1} \cdot p_{t,1:N-1}^{c \rightarrow n}] + [0; (1 - p_t^{h \rightarrow h}) l_{t-1}];$
 | $l_t \leftarrow p_t^{h \rightarrow h} l_{t-1} + a_{t-1,1:N-1} \cdot p_{t,1:N-1}^{c \rightarrow h};$
 | // $\{a_t; l_t\}$ are normalized.

end
 $c_t \leftarrow \sum_{n=1}^N a_{t,n} k_n + k_l \sum_{n=1}^{N-1} l_{t,n};$
return $a_t, l_t, c_t;$

At the training stage, $p_{t,n}^{c \rightarrow c}$, $p_{t,n}^{c \rightarrow n}$, $p_{t,n}^{c \rightarrow h}$ and $p_t^{h \rightarrow h}$ are computed in the probability-based soft mode. In the inference stage, the hard mode is adopted. In the hard mode, the

maximum value among $p_{t,n}^{c \rightarrow c}$, $p_{t,n}^{c \rightarrow n}$ and $p_{t,n}^{n \rightarrow h}$ is set to 1 and the other two values are set to 0, as indicated by the *binary* function in Algorithm 2. $p_t^{h \rightarrow h}$ also becomes 0 or 1.

4. Experiments

4.1. Experimental Setup

A Chinese corpus pronounced by a female speaker was used in our experiments. The scripts were selected from newspapers, and the recordings were sampled at 16 kHz with 16 bits resolution. The total 12,319 utterances (≈ 17.51 h) were split into a training set of 11,608 utterances, a validation set of 611 utterances and a test set of 100 sentences. The training set was used to train acoustic models and the validation set was used to tune hyperparameters.

A publicly available implementation of Tacotron2 (<https://github.com/NVIDIA/tacotron2>, accessed on 12 June 2020) was utilized as the basis of our implementation. When training the model, 80-band mel-spectrograms were used as the acoustic features. The frame length was 64 ms and the frame shift was 15 ms. Phone sequences were adopted as model input and the initials and finals of Mandarin Chinese were treated as phones for simplification. A phone embedding vector, a tone embedding vector and a prosodic position embedding vector were concatenated to represent each phone. The Adam optimizer [18] was used, the training epochs were 200 and the training batch size was 80. The initial learning rate was 1×10^{-3} , and then the learning rate exponential decayed by 0.9 times every 10 epochs. A WaveNet vocoder was built to reconstruct waveforms in our experiments.

Finally, three Tacotron2-based acoustic models were built for comparison. (Audio samples are available at https://xiaozhah.github.io/SSMA_demos, accessed on 17 April 2021).

SMA-PB The attention mechanism was stepwise monotonic attention (SMA) [9]. In both training and synthesis stages, the phrase boundaries of texts were not available. The initial bias b was 3.5 and the noise scale γ was 2.0. The soft mode was used at the training stage, while the hard mode was used at the synthesis stage.

SSMA-PB This model adopted SSMA instead of SMA and other model structure and hyperparameters were the same as SMA-PB. In both training or synthesis stage, this model used the same data as SMA-PB. The initial bias b and b^l was 3.5, and the noise scale γ was 2.0. The soft mode was used at the training stage, while the hard mode was used at the synthesis stage.

SMA+PB This model had the same structure and hyperparameters as SMA. In both training and synthesis stages, this model used the texts with manually labeled phrase boundaries.

4.2. Objective Evaluation

The objective performance of SMA-PB, SSMA-PB, and SMA+PB on predicting the acoustic features of test sentences was evaluated. The metrics included mel-cepstral distortion (MCD), F0 root mean square error (RMSE), F0 correlation (CORR), and unvoiced/voiced (UV) error percentage. The frame-level MCD was calculated as:

$$\text{MCD} = \frac{10}{\log 10} \sqrt{2 \sum_{m=1}^M (c_r(m) - c_s(m))^2}, \quad (1)$$

where c_r and c_s are mel-cepstral coefficients (MCCs) from natural and synthetic speech, respectively, and M is their order. The frame-level F0 RMSE was calculated as:

$$\text{RMSE} = 1200 \sqrt{(\log_2(F_r) - \log_2(F_s))^2}, \quad (2)$$

where F_r and F_s represents F0 values extracted from natural and synthetic speech, respectively. The F0 CORR was defined as the F0 values correlation coefficient between synthetic

speech and natural speech in the voiced segment. The UV error percentage was the ratio of the number of unmatched U/V frames between natural and synthetic speech to the total number of frames. For calculating the four metrics, twelve-dimensional MCCs, and F0 values were extracted from synthetic speech at 5 ms frame shift by STRAIGHT [19] analysis. The FastDTW algorithm [20] based on MCCs was adopted to align predicted acoustic features toward reference ones for calculating the four metrics. The results are shown in Table 1.

Table 1. Objective performance of SMA-PB, SSMA-PB, and SMA+PB on predicting the acoustic features of test sentences, where **MCD**, **RMSE**, **CORR**, and **UV** denote the mel-cepstral distortion, F0 root mean square error, F0 correlation, and UV error percentage, respectively.

	MCD (dB)	RMSE (Hz)	CORR	UV (%)
SMA-PB	3.73	46.49	0.83	7.77
SSMA-PB	3.65	46.13	0.83	7.78
SMA+PB	3.53	41.95	0.88	7.00

From this figure, we can see that SMA+PB achieved the best accuracy of acoustic feature prediction. This is reasonable since it utilized manually labeled phrase boundaries in both training and synthesis stages. Comparing SMA-PB with SMA+PB, it can be found that the objective performance of SMA degraded significantly when phrase boundaries were not available. Comparing SMA-PB, SSMA-PB achieved smaller MCD and comparable F0 distortion, which may be that the MCD was more related to pause segments in utterances.

4.3. Subjective Evaluation

Twenty sentences with at least one phrase boundary were randomly selected from the test set. These phrase boundaries were not used when synthesizing them with SMA-PB and SSMA-PB. The utterances synthesized using the SMA-PB, SSMA-PB, and SMA+PB systems were compared by two groups of AB preference tests on their naturalness. In each test, the synthetic utterances of two systems were evaluated in random order by 11 native listeners. The listeners were asked to judge which sentence in each pair sounded more natural or there was no preference. The average preference scores are shown in Table 2. Each row in the Table 2 compares whether there are significant differences in different systems. The percentage of the better system is shown in bold format.

Table 2. Subjective preference scores (%) among SMA-PB, SSMA-PB, and SMA+PB systems on the test set, where N/P denotes “No Preference” and p means the p -value of t -test between two systems.

SMA-PB	SSMA-PB	SMA+PB	N/P	p
21.36	54.55	-	24.09	<0.001
-	26.36	45.91	27.73	<0.001

From this table, we can see that the SSMA-PB system outperformed the SMA-PB system significantly with $p < 0.001$. This result indicates that using SSMA helped Tacotron2 to synthesize speech with better naturalness when phrase boundaries were not given in both training and synthesis stages. On the other hand, the subjective performance of SSMA-PB was still not as good as SMA+PB which utilized manually labeled phrase boundaries ($p < 0.001$).

4.4. Discussions

To explore the interpretability of our proposed SSMA method, two experiments were conducted to evaluate the consistency between the pause positions derived from SSMA-based alignment and manually labeled phrase boundaries.

4.4.1. Predicting Phrase Boundaries from Texts

This experiment evaluated the accuracy of predicting phrase boundaries from texts using the pause positions determined by SSMA at the inference stage. The sentences in the test set were synthesized by SSMA-PB with hard mode. If the hidden state between two adjacent phones was assigned more than one frame at the decoding time, a phrase boundary was predicted between these two phones. The hidden states adjacent to silence phones were not considered. Evaluation metrics included the precision, recall, and F1 score of predicting phrase boundaries. The results are shown in the first row of Table 3. We can see that SSMA-PB achieved a recall of 73.61%, which means that most true phrase boundaries can be found by SSMA-based decoding. However, its precision was much lower. One possible reason is that there were short pauses determined by SSMA, which may not correspond to true phrase boundaries. It should be noticed that SSMA-PB predicted phrase boundaries in an unsupervised way, i.e., no phrase boundary annotations were utilized at the training stage.

Table 3. Precision (%), Recall (%), and F1 score (%) of the SSMA-PB system on predicting phrase boundaries from texts and annotating phrase boundaries by forced-alignment.

	Precision	Recall	F1 Score
Prediction	32.72	73.61	45.31
Annotation	43.67	78.89	56.22

4.4.2. Annotation Phrase Boundaries by Forced Alignment

This experiment evaluated the accuracy of annotating phrase boundaries by SSMA-based forced alignment when both texts and recordings were given. In this case, the decoder was conducted in a teacher-forcing way, which means that the true history of the mel-spectrogram was used as input at each decoding step. After alignment paths were calculated, the phrase boundaries were annotated following the conditions used in previous experiment. The results of phrase boundary annotation are shown in the second row of Table 3, where the same metrics used in previous experiment were employed. We can see that SSMA-based phrase boundary annotation achieved higher precision, recall, and F1 score than SSMA-based phrase boundary prediction. This is reasonable because the former utilized both textual and acoustic information. The F1 score of SSMA-PB on annotating phrase boundaries was 56.22%, which shows that without supervision those taught hidden state positions in SSMA did have a strong correlation with manually labeled phrase boundaries.

5. Conclusions

This paper has proposed a semi-stepwise monotonic attention (SSMA) method to improve the performance of sequence-to-sequence (Seq2Seq) speech synthesis when phrase boundaries are not available in both training and synthesis stages. In this method, hidden states are added between adjacent phones to model the possible pauses between them. Thus, the attention to a phone at each decoding step has three options for the next decoding step, moving forward to the next phone, staying at the same phone, or jumping to a hidden state. Then, an algorithm was designed to calculate the attention weights of SSMA in a recursive way. Experimental results show that SSMA achieved better subjective performance than SMA when phrase boundaries are not available, which is quite suitable for low-resource languages that lack phrase boundaries. To improve the accuracy of SSMA-based unsupervised phrase boundary annotation and to evaluate our proposed method using the datasets of more languages will be the tasks of our future work.

Author Contributions: Conceptualization, Z.L.; methodology, X.Z.; implementation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, Z.L.; resources, Y.H.; project administration, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Key R&D Program of China under Grant 2019YFF0303001, the Major Program of National Social Science Foundation of China under Grant 15ZDB103 and the National Nature Science Foundation of China under Grant 61871358.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [[CrossRef](#)]
2. Liberman, M.; Church, K.W. Text Analysis and Word Pronunciation in Text-to-speech Synthesis. In *Advances in Speech Signal Processing*; Marcel Dekker: New York, NY, USA, 2013; pp. 791–832.
3. Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), Istanbul, Turkey, 5–9 June 2000; Volume 3, pp. 1315–1318.
4. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [[CrossRef](#)]
5. Den Oord, A.V.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.W.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, USA, 13–15 September 2016; p. 125.
6. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 4006–4010. [[CrossRef](#)]
7. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
8. Zhang, J.X.; Ling, Z.H.; Dai, L.R. Forward Attention in Sequence-to-sequence Acoustic Modeling for Speech Synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4789–4793.
9. He, M.; Deng, Y.; He, L. Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1293–1297. [[CrossRef](#)]
10. Sanders, E.; Taylor, P. Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis. In Proceedings of the EUROSPEECH '95, Madrid, Spain, 18–21 September 1995.
11. Bisani, M.; Ney, H. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Commun.* **2008**, *50*, 434–451. [[CrossRef](#)]
12. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
13. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 577–585.
14. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech: Fast, robust and controllable text to speech. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 3171–3180.
15. Ren, Y.; Hu, C.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech. *arXiv* **2020**, arXiv:2006.04558.
16. Battenberg, E.; Skerry-Ryan, R.; Mariooryad, S.; Stanton, D.; Kao, D.; Shannon, M.; Bagby, T. Location-Relative Attention Mechanisms For Robust Long-Form Speech Synthesis. *arXiv* **2019**, arXiv:1910.10288.
17. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv* **2016**, arXiv:1611.01144.
18. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
19. Kawahara, H.; Masuda-Katsuse, I.; De Cheveigne, A. 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 Commun.* **1999**, *27*, 187–207. [[CrossRef](#)]
20. Salvador, S.; Chan, P. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **2007**, *11*, 561–580. [[CrossRef](#)]