applied sciences

MDPI

Article

Grammar-Supervised End-to-End Speech Recognition with Part-of-Speech Tagging and Dependency Parsing

Genshun Wan [1,2,*], Tingzhi Mao [2], Jingxuan Zhang [2], Hang Chen [1], Jianqing Gao [2] and Zhongfu Ye [1]

[1] National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230088, China
[2] iFLYTEK Research, iFLYTEK Co., Ltd., Hefei 230088, China
* Correspondence: gswan@mail.ustc.edu.cn

Abstract: For most automatic speech recognition systems, many unacceptable hypothesis errors still make the recognition results absurd and difficult to understand. In this paper, we introduce the grammar information to improve the performance of the grammatical deviation distance and increase the readability of the hypothesis. The reinforcement of word embedding with grammar embedding is presented to intensify the grammar expression. An auxiliary text-to-grammar task is provided to improve the performance of the recognition results with the downstream task evaluation. Furthermore, the multiple evaluation methodology of grammar is used to explore an expandable usage paradigm with grammar knowledge. Experiments on the small open-source Mandarin speech corpus AISHELL-1 and large private-source Mandarin speech corpus TRANS-M tasks show that our method can perform very well with no additional data. Our method achieves relative character error rate reductions of 3.2% and 5.0%, a relative grammatical deviation distance reduction of 4.7% and 5.9% on AISHELL-1 and TRANS-M tasks, respectively. Moreover, the grammar-based mean opinion score of our method is about 4.29 and 3.20, significantly superior to the baseline of 4.11 and 3.02.

Keywords: speech recognition; grammar knowledge; multiple evaluation methodology of grammar; grammatical deviation distance

Citation: Wan, G.; Mao, T.; Zhang, J.; Chen, H.; Gao, J.; Ye, Z. Grammar-Supervised End-to-End Speech Recognition with Part-of-Speech Tagging and Dependency Parsing. *Appl. Sci.* **2023**, *13*, 4243. https://doi.org/10.3390/app13074243

Academic Editor: Javier Hernando

Received: 24 February 2023
Revised: 22 March 2023
Accepted: 23 March 2023
Published: 27 March 2023

1. Introduction

With the inexorable logarithmic growth of computing power and the development of deep learning, intelligent speech technology has developed rapidly and vigorously in recent years. The automatic speech recognition (ASR) system, one of the most mature and advanced techniques in artificial intelligence, has been applied in many fields, improving work efficiency and lifestyle changes [1–3]. Especially when the end-to-end system becomes the hotspot, the performance of large-vocabulary continuous speech recognition (LVCSR) systems has reached a new historic height [4–6].

However, along with the extensive application of ASR systems, the higher demand for the performance of LVCSR is increasing, which leads to more comprehensive evaluation indicators [7–9]. Even though the large-scale test shows a relatively low error rate, many unacceptable hypothesis errors still make the recognition results absurd and difficult to understand for most common application scenarios. These errors break some natural language expression limitations, such as grammatical and semantic rules, which adversely affect the recognition result to generalize for some post-processing tasks, for example, text summarization (TS) [10], spoken language understanding (SLU) [11] and machine translation (MT) [12]. A keyword error may utterly destroy the readability and reliability of the whole utterance [8]. Therefore, enhancing the rationality and robustness of speech recognition systems from a grammatical point of view is important for promoting it to large-scale applications.

Grammar information is a prescriptive set of specifications that allow people to combine words into phrases and sentences during speech interaction. Recognition errors in

the basic grammatical construction greatly impair the meaning of the whole utterance. On the contrary, correct grammatical knowledge can contribute to the understanding of the speech recognition results. A reasonable grammatical structure may aid users' understanding by linking prior knowledge in a more direct way. Although there may exist some speech recognition errors, most people can really understand what the saying means quickly. Accordingly, the grammar criteria and word error rate (WER) deserve the same attention that can adapt to the impending technological trends and user requirements. Although grammar knowledge can be founded in the language model used in the ASR system, explicit grammar constraints apparently are a more effective way to help the ASR system achieve better recognition performance and improve the readability of the speech hypothesis.

As is known to all, grammar knowledge is an essential part of natural language processing (NLP) and is widely used in many tasks in automatic knowledge acquisition. In [13], dependency parsing (DP) was introduced to design the syntax-related pre-training tasks to improve the performance of the language model. Moreover, Refs. [14,15] employed part-of-speech (POS) embedding to reinforce the word-embedding expression in the MT task. In addition, some researchers focused on correcting grammatical and semantic errors in the speech recognition hypothesis with downstream tasks in order to improve the fluency and readability of the ASR output [16–18]. An independent NLP task may suffer from the limited decoding hypothesis from the ASR output or inconsistency with the original speech expression. However, little attention has been given to the combination of automatic speech recognition tasks and grammar knowledge, which provides more possibility for the one-pass decoding process. Unlike the common NLP task, speech recognition cannot take advantage of the full context in advance to acquire relatively accurate grammar knowledge during the first decoding process. Thus, incorporating grammar knowledge into the speech recognition task will bring enormous challenges.

Compared with the traditional hidden Markov model (HMM)-based and neural network (NN)-based frame-to-frame systems [19,20], end-to-end speech recognition systems can directly convert speech input into the corresponding text with the joint optimization of both acoustic and language models. In this sense, the end-to-end structure is better suited to using grammar constraints as supplementary information. At present, end-to-end methods mainly include connectionist temporal classification (CTC) [21], the recurrent neural network transducer (RNN-T) [22], and an encoder–decoder based on the attention mechanism [23]. As a typical non-autoregressive method, CTC can realize the mapping between input speech and target labels without an additional alignment operation while it depends strongly on the conditional independent assumptions. RNN-T consists of an encoder, a prediction network, and a joint network, and it has gained a lot of attention because of its streaming mode. The encoder–decoder framework, especially, which is mainly divided into three parts—encoder, decoder, and the attention module—is often used together with the CTC model within the multi-task learning framework to improve the recognition robustness [24]. The structure paradigm provides a good supplement or alternative to restrict the output hypothesis within the grammatical category.

In order to evaluate the reasonableness of the speech recognition hypothesis with grammar knowledge, a series of methods are proposed to improve the recognition performance in this work. Based on the mainstream end-to-end ASR system, POS and DP are regarded as the two operational dimensions of grammar knowledge. First, we integrate the grammar knowledge into the additional representation of word embedding to further reinforce the grammar rules of the final recognition hypothesis. The grammar knowledge is extracted from the hypothesis of the CTC branch. Next, a text-to-grammar downstream task is introduced in the training stage to avoid the mismatch of access to grammar knowledge between the training and inference stage. With a multi-task learning method, the accuracy and the intelligibility of speech recognition can be learned jointly. Finally, a complete framework for the multiple evaluation methodology of grammar (MEMG) is proposed. In order to adapt to practical scenarios, MEMG shares most of the decoder param-

eters of end-to-end systems with little training cost. Meanwhile, an evaluation method for the grammatical deviation distance (GDD) is defined to evaluate the reasonable degree of ASR output and overcome the one-sidedness of the traditional WER criterion. Finally, the grammar-based mean opinion score is analyzed based on a subjective evaluation experiment to help us better evaluate our proposed methods.

We briefly summarize the contributions of this study as follows:

1. A method for incorporating grammar knowledge into end-to-end ASR systems is proposed, which can improve the accuracy of speech recognition and reduce hypothesis errors.
2. Several grammar constraint methods, such as the reinforcement of word embedding with grammar embedding and adding the auxiliary text-to-grammar downstream task, are proposed, which can increase the readability of the hypothesis and improve the user experience of ASR systems.
3. For providing various evaluation methods for grammar knowledge in recognition outputs, we introduced an evaluation method for the grammatical deviation distance and a grammar-based mean opinion score to supplement existing common evaluation criteria such as WER and CER for exploring an expandable usage paradigm with grammar knowledge.

Our proposed framework has been demonstrated to be effective in the publicly available dataset. Its incorporation of grammar knowledge into end-to-end ASR systems not only provides a valuable avenue for future research, but also enables increased reproducibility in this field. The rest of the paper is organized as follows: Section 2 gives a brief introduction to the ASR framework. Section 3 elaborates on the proposed method of incorporating grammar knowledge into speech recognition tasks. In Section 4, we report and analyze the experimental results. Finally, we conclude our findings in Section 5.

2. The Framework of the ASR System

The baseline framework of the ASR system considered in this paper is presented in this section.

The attention-based encoder–decoder ASR system has enabled state-of-the-art speech recognition performance over a wide range of tasks. In particular, it consists of a "conformer" encoder module and a "transformer" decoder module [25] as shown in Figure 1. In the conformer module, the input sampling rate is reduced from 10 ms to 40 ms with a convolutional subsampling block, and the rate is kept constant through the entire multi-blocked stacked architecture. Each encoder block consists of the following components: a sequence of feed-forward modules, a multi-head self-attention module, a convolution module, and another feed-forward module. The conformer block gives consideration to the diversity of both global interactions and local features, and obtain better accuracy with fewer parameters.

During encoder–decoder training, the CTC criteria, whose non-autoregressive decoding method benefits the inference latency, is used to maintain the training robustness and achieve a fast convergence as the following loss function formula. In addition, the tunable parameter λ is set between 0 and 1:

$$Loss_{asr} = \lambda Loss_{ce} + (1 - \lambda) Loss_{ctc} \quad (1)$$

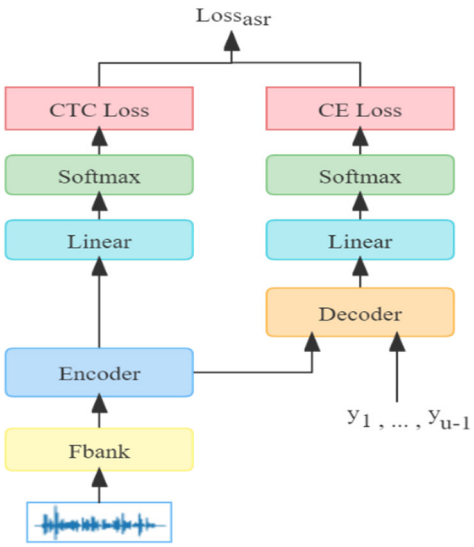

Figure 1. The framework of the ASR system.

3. Integration of Grammar Knowledge

The attention-based encoder–decoder ASR system uses grammar knowledge such as POS and DP to supplement the information representation. In this section, the proposed methods are introduced in detail, including the reinforcement of word embedding with grammar embedding, the auxiliary text-to-grammar task, and a complete framework for the multiple evaluation methodology of grammar.

3.1. Reinforcement of Word Embedding with Grammar Embedding

A language model is essential in many natural language processing research fields, especially in information retrieval, question answering, speech recognition, etc. Moreover, data-driven word embedding in an encoder–decoder ASR system can be used to describe the probability distribution of different syntactic units, such as words, statements, and even whole documents. However, many grammatical errors in the speech recognition hypothesis indicate that there is still plenty of room for language model representation improvement. Despite being extremely important, the grammar information in recognition system rarely gets the attention it deserves. The critical point to changing this situation is that we should capture grammar information more explicitly to produce a more acceptable quality of the speech recognition hypothesis.

Inspired by the ideas about incorporating grammar tags into the transformer neural machine translation (NMT) framework [14], we could consider enriching the grammatical representation of word embedding with POS. According to the characteristics of different languages, POS helps to identify the correct interpretation of the word. Therefore, the POS grammar information can serve as a complemental input of the ASR system to further enhance the performance of the baseline system.

Since each word can be assigned to a POS tag considering the context of the word in a sentence, we can pre-process all the training data to obtain POS tags with the help of an open-source language technology platform (LTP) tool. Firstly, we realize the word segment for all the ground-truth labels in the training data based on the language model, and then obtain the POS labels for each word in the utterance to align the grammar label sequence with the character label sequence. Finally, the extracted grammar information is taken as a part of the input embedding to combine with the word embedding by direct concatenation. The grammatical constraints will become more normative through this pattern of self-reinforcement. The following formula describes the process of the reinforcement of word embedding with grammar knowledge, where the FusionLayer is the splice or addi-

tion operation. In addition, the entire flow of the reinforcement of word embedding with grammar embedding is illustrated in Figure 2.

$$Y_{emb-ori} = EmbedLayer(y) \tag{2}$$

$$Y_{emb-pos} = EmbedLayer(y_{pos}) \tag{3}$$

$$Y_{emb} = \text{FusionLayer}(Y_{emb-ori}; Y_{emb-pos}) \tag{4}$$

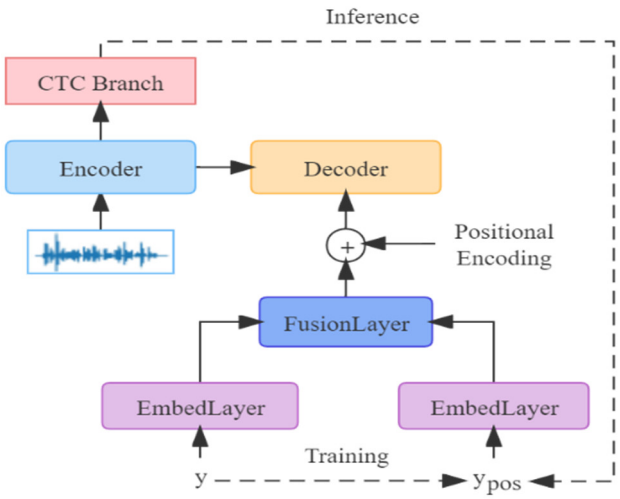

Figure 2. Reinforcement of word embedding with grammar embedding.

We have to consider the mismatch between the training and testing conditions for the grammar knowledge acquisition mode. While decoding, no ground-truth labels can be used to ensure the accuracy of the grammar information. As an alternative, grammar labels can be extracted from the hypothesis of the CTC branch. Although there may exist some inevitable speech recognition errors, the expand space of the POS tags based on the hypothesis is relatively limited.

3.2. The Auxiliary Text-To-Grammar Downstream Task

Although the CTC branch makes the reinforcement of word embedding with grammar embedding possible, the grammar knowledge extracted from the CTC branch and the final hypothesis of the encoder–decoder branch support and interfere with each other. This situation may increase the risk of low robustness in complex scenes. An auxiliary text-to-grammar downstream task can be added to the ASR system based on the performance evaluation with different types of grammatical knowledge to constrain the final hypothesis directly. Under this circumstance, joint training patterns of speech recognition accuracy and intelligibility may be more effective.

In this study, we propose a new architecture for text-to-grammar tasks based on the grammatical knowledge of part-of-speech and dependency parsing, a vital task format in natural language processing. The text-to-grammar task is a downstream task to constrain the primary ASR model, especially in grammatical representation. In order to enhance the effective efficiency of the text-to-grammar task and further influence the ASR output, the top hypothesis from the standard encoder–decoder ASR system is equally regarded as an additional input via a new embedding layer.

The details of the auxiliary text-to-grammar task are shown in Figure 3. Here, we mark the grammar-label sequence as y_{gram}. In addition, the h_u is used to express the hidden states of the decoder module and the h_e is the representation of the one best decoding

result encoded by the embedding layer. After obtaining h_u and h_e, the addition operation is employed to merge them as the input of the text-to-grammar task. The encoder of the transformer is used as the backbone network of the text-to-grammar. The following formula describes the final loss function at the training stage and θ is introduced as a hyper-parameter to balance the two tasks.

$$Loss_{t2g} = CE([h_e : h_u], y_{gram}) \tag{5}$$

$$Loss_{final} = Loss_{asr} + \theta Loss_{t2g} \tag{6}$$

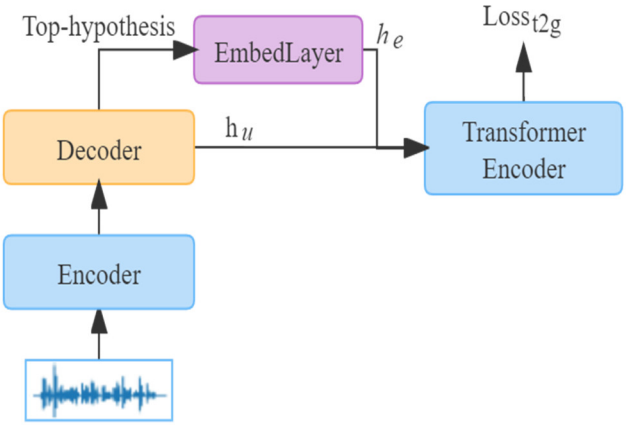

Figure 3. Text-to-grammar downstream task.

3.3. Multiple Evaluation Methodology of Grammar

In order to further explore an expandable usage paradigm with grammar knowledge for the ASR system, we extend a complete framework and make it more robust to the overall model structure.

As shown in Figure 4, the architecture of the multiple evaluation methodology of grammar consists of the original ASR loss, the part-of-speech loss, and the dependency parsing loss. In order to combine these extended tasks into the main architecture, we also employ a multi-task training method. Nevertheless, unlike the auxiliary text-to-grammar downstream task, the new one is a parallel constraint network structure. The three tasks share most parts of the decoder parameters. In addition, an independent project layer was added before the output layers to learn the representation of different types of grammar knowledge, respectively. Specifically, consistent with the format of the inputs and outputs for the original ASR loss, the part of speech is evaluated with the sequence grammar labels y_{pos}, and the representation after passing the linear layer of POS is recorded as h_{pos}. However, the dependency parsing task is mainly based on the utterance structure, so we cannot migrate from the sequence method. Inspired by [26], a bilinear function is introduced to imitate the structure information, which consists of two parts: the arc and head, respectively, represent the relationship and location information between words in the utterance. Here, they are called h_{arc} and h_{head}. The following formula describes the process of the multiple evaluation methodology of grammar.

$$Loss_{pos} = CE(h_{pos}, y_{pos}) \tag{7}$$

$$Loss_{dp} = \gamma CE(h_{arc}, y_{arc}) + (1 - \gamma)\, CE(h_{head}, y_{head}) \tag{8}$$

$$Loss_{final} = Loss_{asr} + \delta Loss_{pos} + (1 - \delta) Loss_{dp} \tag{9}$$

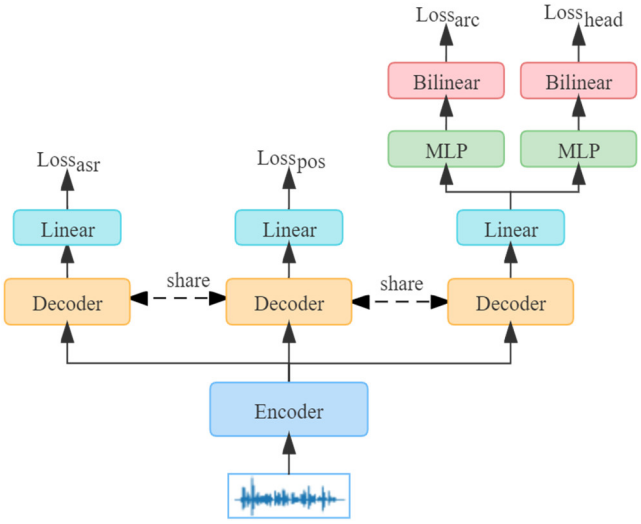

Figure 4. Multiple evaluation methodology of grammar.

4. Experiments

In this section, the datasets and the experimental setup are described in detail, firstly. Then, a new evaluation method for the grammatical deviation distance is defined to evaluate the reasonable degree of the ASR output. Meanwhile, we also introduce the grammar-based mean opinion score to evaluate the quality of the hypothesis subjectively. Finally, the performance of the three methods is evaluated from different angles to give a more comprehensive analysis of the recognition performance.

4.1. Datasets

We evaluated the performance of the proposed approach on both the small open-source Mandarin speech corpus AISHELL-1 [27] and the sizeable private-source Mandarin speech corpus TRANS-M; the details of the two speech recognition corpus are illustrated in Table 1.

Table 1. The details of the speech recognition corpus.

Dataset Name	Subset	Duration (h)	Utterance
AISHELL-1	Training	150	120,098
	Development	10	14,326
	Test	5	7176
TRANS-M	Training	12,000	12,225,244
	Development	3	4834
	Test	10	27,952

The AISHELL-1 task is a benchmark for large-Mandarin-vocabulary continuous speech recognition, so we choose it as the main task. As an elaborately prepared public dataset, AISHELL-1 contains over 170 h of recorded speech data, including 400 speakers. For each speaker, around 360 utterances are provided. In addition, the domain of AISHELL-1 mainly includes finance, science and technology, sports, entertainment, and news.

The TRANS-M task combines multiple data styles, including lessons, conferences, interviews, and television programs. Obviously, the vocabulary is varied and covers more fields. Unlike AISHELL-1, TRANS-M turns to a more colloquial and dialogic linguistic style. On the whole, the automatic speech recognition on this dataset is challenging and highly applicably valuable.

4.2. Experimental Setup

All experiments in our study were performed using the WeNet toolkit [28] and run on a server equipped with 8 Tesla V100 GPUs.

Based on the conformer ASR system described in Section 2, an 80-dimension log-Mel filter bank was used as input features with a 25 ms Hamming window with a 10 ms fixed frame rate, and all the training features were processed with mean and variance normalization. Speed perturbation and SpecAugment [29] were used simultaneously for additional data augmentation. Specifically, speech was perturbed on speed with the factors of 0.9, 1.0, and 1.1; SpecAugment was implemented, with two frequency masks with the maximum frequency mask (F = 10) and two times masks with the maximum time mask (T = 50). The architecture of the encoder model is composed of a convolutional subsampling module and 12-layer conformer blocks. The convolutional subsampling module contains 2D convolutional layers with stride 2, resulting in a 40 ms frame-rate output. For the conformer blocks of the encoder module, each layer is configured with an 8-head attention of 512-dim and 2048 feed-forward hidden nodes. For the transformer blocks in the decoder module, each layer is also configured with an 8-head attention and 2048 feed-forward hidden nodes. Finally, 6728 tokens served as the decoder outputs, including the Chinese characters in GB2312, a start symbol, an end symbol, and an unknown symbol used to indicate 'out of vocabulary'.

In order to prevent over-fitting, label smoothing [30] is applied in the training process, and the penalty is set to 0.1. The Adam algorithm [31] is adopted to avoid falling into the local minimum and ensure the algorithm's stability with regularization terms. The learning rate firstly warms up linearly to 0.002 in 25,000 steps and then decreases proportionally to the inverse square root of the step number. To the framework of the ASR system, the CTC weight is set to 0.3 in training while it is set to 0.5 in decoding.

4.3. Evaluation Metric of Hypothesis

The character error rate (CER), word error rate (WER), and sentence error rate (SER) are the three main evaluation methods used to measure the performance of the speech recognition hypothesis. These methods can objectively reflect the corresponding relationship between the recognition result and ground truth involved in the deletion (D), insertion (I), and substitute (S) errors. As is well-known, we can guess the real meaning of the hypothesis even if some recognition errors may exist. However, there is another possibility that some sentences are difficult to understand even if there is only one recognition error in it. To some degree, the common evaluation criteria, such as CER, WER, and SER, cannot reflect the legibility of the final recognition result in a more direct way. In some scenarios, the WER or CER improvement may no longer correlate with a practical value. It is far from being enough to take the above evaluation criteria as the only measure of the quality of an ASR system.

In order to give a more objective and comprehensive evaluation of the speech recognition hypothesis, another evaluation method involved in the grammatical knowledge is defined in this work as follows:

Grammatical deviation distance (GDD): The study of its characteristics is to objectively and accurately redefine a new evaluation metric for the hypothesis, show concern for grammar knowledge, and pay attention to the grammatical deviation distance between the hypothesis and ground truth. GDD is proposed as a complementary method of the speech recognition evaluation criteria. Specifically, we first use the LTP toolkit to obtain the linguistic tags for the hypothesis and ground truth, such as part-of-speech (POS) and dependency parsing (DP), etc. Then, we align each of them on the linguistic tag level. Finally, we calculate the GDD of each utterance according to the weights corresponding to the linguistic tags. We believe that different linguistic tag errors have different effects on the GDD. Moreover, we also prove that the design of linguistic tag weights (LTW) is almost

consistent with the manual review results. The following formula shows the details of the calculation for the grammatical deviation distance.

$$U_{GDD} = \frac{LTW_{error}}{LTW_{total}} \tag{10}$$

$$\text{GDD} = \frac{\sum_{i=1}^{M} U_{GDD}}{M} \times 100\% \tag{11}$$

where U_{GDD} is the grammatical deviation distance of each utterance, LTW_{error} refers to the sum of the linguistic label weights corresponding to the error linguistic label position, and LTW_{total} is the sum of all linguistic tag weights. In addition, M is the number of utterances in the data set.

4.4. Experiment Results

Establishment of baseline: In this section, we intend to build a competitive baseline compared with the prior work to evaluate the effectiveness of the proposed system. As described in the dataset, we use AISHELL-1 as the primary training set to establish the baseline model and make sure the training process is fair and reasonable. As shown in Table 2, the CER in both Dev and Test are competitive enough and robust enough compared with the other two typical speech recognition systems based on the small open-source Mandarin speech corpus AISHELL-1.

Table 2. The CER of the baseline system.

Model	CER (%)	
	Dev	Test
ESPnet [32]	-	5.1
K2 [33]	4.55	5.1
Ours	4.22	4.87

4.4.1. Reinforcement of Word Embedding with Grammar Embedding (RWE)

The extracted grammar information is taken as a part of input embedding to enrich the representation of each token embedding. In this part, we actually only use the part-of-speech embedding to combine with the word embedding because the sequence of the POS can be consistent with the character sequence. The POS sequence was obtained from the ground truth in the training process while it was obtained from the hypothesis of the CTC branch in the testing process. In order to maintain the unity on the scale of the modeling unit and combine the two types of embedding, the BIES labeling criteria was used to make POS tags for different words and expressions.

As shown in Table 3, when the POS sequence was obtained from the ground truth, the theoretical limit seemed impressive. For AISHELL-1, the CER dropped from 4.87% to 4.49% while the GDD dropped from 6.16% to 5.50%. There has been a similar improvement in TRANS-M. When the POS sequence was obtained from the hypothesis from the CTC branch, the improvement of the CER and GDD might seem modest on both the AISHELL-1 and TRANS-M dataset. In actual environments, the performance of the ASR system may be degraded significantly because of the mismatch between the training and testing conditions on the POS information. Although the potential of the grammar information is enormous, it has not been fully developed.

Table 3. The CER and GDD of the proposed system on two test sets.

Model	AISHELL-1 (%)		TRANS-M (%)	
	CER	GDD	CER	GDD
Baseline	4.87	6.16	18.14	24.02
RWE (ground truth)	4.49	5.50	15.30	21.02
RWE (hypothesis from CTC branch)	4.84	6.16	18.03	24.00
AT2G	4.83	6.08	17.73	23.06
MEMG	4.71	5.87	17.23	22.60

4.4.2. The Auxiliary Text-to-Grammar Downstream Task (AT2G)

Since mismatch is a long-standing and complex problem, we take the joint training of the ASR task and text-to-grammar task as a new configuration. In this experiment, the loss of the text-to-grammar falls at a high speed, and the joint training tasks have a fast convergence rate and good stability in the training process.

As reported in Table 3, the AT2G slightly reduces the CER and GDD on both datasets. For TRANS-M, the CER and GDD of AT2G are 17.73% and 23.06%, respectively, representing a relative CER reduction of 2.26% and a relative GDD reduction of 4.00% over the baseline. However, AT2G adds extra parameters and the cost of model training, and weakens the constraints of the hidden features in the ASR decoder.

4.4.3. Multiple Evaluation Methodology of Grammar (MEMG)

To make the best use of the grammar information, we combine both the part-of-speech loss and the dependency parsing loss into the original loss simultaneously. As described in Table 3, the results on both AISHELL-1 and TRANS-M show that the multiple evaluation methodology of grammar obtains a very significant performance improvement in terms of the CER and GDD. Specifically, the AISHELL-1 of the CER decreased from 4.87% to 4.71% for the test set that has a 3.29% relative reduction, while TRANS-M of the CER decreased from 18.14% to 17.23% with a 5.0% relative reduction. Meanwhile, compared with the baseline model, the relative GDD reduction of AISHELL-1 and TRANS-M can reach 4.7% and 5.9%, respectively, which indicates the improvement of readability of the final speech recognition hypothesis.

In order to know the influence of the different evaluation methodologies on the ASR system, we designed a series of ablation experiments with different hyper-parameters to balance the tasks. As shown in Table 4, different grammatical weights are set to make certain the constraining function of the different evaluation methodologies.

Table 4. The ablation experiments of MEMG.

Evaluation Methodology	Weight	AISHELL-1 (%)			
		Dev		Test	
		CER	GDD	CER	GDD
DP	0	4.42	5.48	4.87	6.16
	0.1	4.50	5.57	4.88	6.03
	0.3	4.42	5.43	4.86	6.06
	0.5	4.40	5.43	4.82	6.02
	0.7	4.45	5.47	4.93	6.17
POS	0	4.42	5.48	4.87	6.16
	0.1	4.31	5.26	4.78	5.84
	0.3	4.31	5.16	4.72	5.84
	0.5	4.34	5.26	4.69	5.80
	0.7	4.27	5.17	4.80	5.83

Compared with the DP task, the POS task is more effective for the ASR system within the grammar constraint. It can be explained that learning the structure of a whole utterance is challenging and usually requires adequate extended information. Meanwhile, we also note that the lower CER does not necessarily mean a lower GDD, which is almost consistent with our guess and observation. When the weight is 0.5, we find the model has almost achieved consistency improvement.

4.4.4. Model Comparison with External Language Model

In order to prove the practicability and generalization of the proposed model, the results from the different approaches with the 4-gram language model on the AISHELL-1 test are described in Table 5.

Table 5. Model comparison with language model on AISHELL-1.

Model	No LM (%)		4-gram LM (%)	
	CER	GDD	CER	GDD
Baseline	4.87	6.16	4.59	5.79
MEMG	4.71	5.87	4.47	5.51

Compared with the baseline with the external language model, the multiple evaluation methodology of grammar can also obtain a relative CER reduction of 2.61% and a relative GDD reduction of 4.84%. Overall, the degree of improvement is only slightly decreased on CER while stable on GDD. For the attention-based encoder–decoder ASR system which involves the acoustic model and language models, the proposed method could pay more attention to syntactic information during the training process.

4.5. Grammar-Based Mean Opinion Score of Speech Recognition Hypothesis

As an objective statistical indicator based on the LTP toolkit [34], GDD is defined to evaluate the reasonable degree of the speech recognition hypothesis. In order to reflect the proper degree of intelligibility for all the hypotheses of the different methods, we introduce a new method to listen to the natural feeling of the users and pay more attention to their satisfaction. Inspired by the application of the mean opinion score (MOS) in text-to-speech [35,36], the grammar-based mean opinion score (GMOS) is analyzed based on a subjective evaluation experiment. Three volunteers with linguistic ability are chosen to rate the candidate hypothesis on factors including reading fluency, grammar changes, and semantic deviation. After obtaining the GMOS of the three volunteers, the mean value is calculated. From Table 6, it can be observed that the multiple evaluation methodology of grammar has a significant improvement. Compared with the baseline system, the MEMG increased the GMOS from 4.11 to 4.29 on AISHELL-1, while it increased the GMOS from 3.02 to 3.20 on TRANS-M. These results indicate that the proposed method is conducive to improving the subjective experience.

Table 6. The GMOS of the proposed system on two test sets.

Model	GMOS	
	AISHELL-1	TRANS-M
Baseline	4.11 ± 0.06	3.02 ± 0.08
MEMG	4.29 ± 0.06	3.20 ± 0.08

In Table 7, we also list some examples from the test sets. Our proposed methods improve the readability of the speech recognition results compared with the baseline model, which proves the improvement of the grammatical level again.

Table 7. The examples of two test sets.

ID	Model	Utterance
1	Ground-Truth	Chinese: 何 为 爱 你 知 道 嘛 English: You know what love is
	Baseline	Chinese: 何 为 哎 你 知 道 嘛 English: You know what it is
	MEMG	Chinese: 何 为 爱 你 知 道 嘛 English: You know what love is
2	Ground-Truth	Chinese: 实 行 全 口 径 统 计 和 动 态 检 测 English: Implement full caliber statistics and dynamic detection
	Baseline	Chinese: 实 行 权 口 净 统 计和 动 态 检 测 English: Implement net power statistics and dynamic detection
	MEMG	Chinese: 实 行 全 口 径 统 计和 动 态 检 测 English: Implement full caliber statistics and dynamic detection

Both the objective statistical indicators and subjective experience indicators above show that the grammar-supervised end-to-end speech recognition system may be more suitable for improving the readability and accuracy of the speech recognition hypothesis. By introducing the grammar knowledge such as part-of-speech tagging and dependency parsing, the information representation of the attention-based encoder–decoder ASR system can correlate well with an ability to understand the syntactic structure. Compared with the implicit learning of grammatical knowledge in the decoder module, the explicit use of grammar knowledge is more conducive to correlating with practical value. The expandable usage paradigm with grammar knowledge can be used in other intelligent speech systems, especially those involved in speech and text transition such as text-to-speech systems [35,36].

5. Conclusions

In this study, we have proposed a method for incorporating grammar knowledge into automatic speech recognition systems. We have shown that explicit grammar constraints are an effective way to improve the recognition performance of ASR systems and increase the readability of the speech hypotheses. Specifically, we have proposed several grammar constraint methods, including the reinforcement of word embedding with grammar embedding and adding the auxiliary text-to-grammar downstream task, which can intensify the grammar rules of the final recognition hypothesis.

For providing various evaluation methods for grammar knowledge in recognition outputs, we introduced an evaluation method for the grammatical deviation distance and a grammar-based mean opinion score to supplement the existing common evaluation criteria such as WER and CER for exploring an expandable usage paradigm with grammar knowledge.

To evaluate the effectiveness of our approach, we conducted experiments on the small open-source Mandarin speech corpus AISHELL-1 and large private-source Mandarin speech corpus TRANS-M to verify the effectiveness of our method. Compared with the baseline, our approach achieved relative CER reductions of 3.2% and 5.0% and relative GDD reductions of 4.7% and 5.9% on AISHELL-1 and TRANS-M tasks, respectively. Meanwhile, the grammar-based mean opinion score of our method is about 4.29 and 3.20, while the baseline can only reach 4.11 and 3.02.

Overall, our study provides a framework for future research on incorporating grammar knowledge into ASR systems. We believe that this approach has great potential for improving speech recognition technology and enhancing the user experience in various applications such as virtual assistants, voice-controlled devices, and automated customer service systems.

Future work will evaluate the performance of deeper integration with the grammar information, and we will also incorporate more downstream tasks to increase the role of multiple evaluation methods.

Author Contributions: Conceptualization, G.W.; methodology, G.W. and J.Z.; project administration, J.G.; software, T.M.; supervision, J.G. and Z.Y.; validation, G.W. and T.M.; writing—original draft, T.M.; writing—review & editing, G.W., J.Z. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Key R & D Program of China (Open Innovation Platform on Intelligent Voice, 2020AAA0103600).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets are analyzed in this study. These data can be found here: (http://openslr.org/33, accessed on 4 July 2019).

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

References

1. Dong, P.; Wang, S.; Niu, W.; Zhang, C.; Lin, S.; Li, Z.; Gong, Y.; Ren, B.; Lin, X.; Tao, D. RTMobile: Beyond Real-Time Mobile Acceleration of RNNs for Speech Recognition. In Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 20–24 July 2020; pp. 1–6.
2. Chenxuan, H. Research on Speech Recognition Technology for Smart Home. In Proceedings of the 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 19–21 December 2021; pp. 504–507.
3. Sathyendra, K.M.; Muniyappa, T.; Chang, F.-J.; Liu, J.; Su, J.; Strimel, G.P.; Mouchtaris, A.; Kunzmann, S. Contextual Adapters for Personalized Speech Recognition in Neural Transducers. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8537–8541.
4. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proceedings of the Advances in Neural Information Processing Systems 33, Virtual, 6–12 December 2020; pp. 12449–12460.
5. Li, B.; Chang, S.-Y.; Sainath, T.N.; Pang, R.; He, Y.; Strohman, T.; Wu, T. Towards Fast and Accurate Streaming End-To-End ASR. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6069–6073.
6. Chen, X.; Wu, Y.; Wang, Z.; Liu, S.; Li, J. Developing Real-Time Streaming Transformer Transducer for Speech Recognition on Large-Scale Dataset. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 5904–5908.
7. Kim, S.; Le, D.; Zheng, W.; Singh, T.; Arora, A.; Zhai, X.; Fuegen, C.; Kalinli, O.; Seltzer, M.L. Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. *arXiv* **2021**, arXiv:2110.05376.
8. Kim, S.; Arora, A.; Le, D.; Yeh, C.F.; Fuegen, C.; Kalinli, O.; Seltzer, M.L. Semantic Distance: A New Metric for ASR Performance Analysis towards Spoken Language Understanding. *arXiv* **2021**, arXiv:2104.02138.
9. Roy, S. Semantic-WER: A Unified Metric for the Evaluation of ASR Transcript for End Usability. *arXiv* **2021**, arXiv:2106.02016.
10. Weng, S.; Lo, T.; Chen, B. An effective contextual language modeling framework for speech summarization with augmented features. In Proceedings of the European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 316–320.
11. Sari, L.; Thomas, S.; Hasegawa-Johnson, M. Training Spoken Language Understanding Systems with Non-Parallel Speech and Text. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8109–8113.
12. Hrinchuk, O.; Noroozi, V.; Ganesan, A.; Campbell, S.; Subramanian, S.; Majumdar, S.; Kuchaiev, O. NVIDIA NeMo Offline Speech Translation Systems for IWSLT 2022. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), Dublin, Ireland, 26–27 May 2022.
13. Sun, B.; Wang, B.; Che, W.; Wu, D.; Chen, Z.; Liu, T. Improving Pretrained Language Models with Syntactic Dependency Prediction Task for Chinese Semantic Error Recognition. *arXiv* **2022**, arXiv:2204.07464.
14. Perera, R.; Fonseka, T.; Naranpanawa, R.; Thayasivam, U. Improving English to Sinhala Neural Machine Translation using Part-of-Speech Tag. *arXiv* **2022**, arXiv:2202.08882.
15. Hlaing, Z.Z.; Thu, Y.K.; Supnithi, T.; Netisopakul, P. Improving neural machine translation with POS-tag features for low-resource language pairs. *Heliyon* **2022**, *8*, e10375. [CrossRef] [PubMed]
16. Liao, J.; Eskimez, S.E.; Lu, L.; Shi, Y.; Gong, M.; Shou, L.; Qu, H.; Zeng, M. Improving Readability for Automatic Speech Recognition Transcription. *arXiv* **2020**, arXiv:2004.04438. [CrossRef]
17. Chen, Y.-C.; Cheng, C.-Y.; Chen, C.-A.; Sung, M.-C.; Yeh, Y.-R. Integrated Semantic and Phonetic Post-Correction for Chinese Speech Recognition. In Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING), Taoyuan, Taiwan, 15–16 October 2021; pp. 95–102.

18. Mani, A.; Palaskar, S.; Meripo, N.V.; Konam, S.; Metze, F. ASR Error Correction and Domain Adaptation Using Machine Translation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6344–6348.
19. Naik, A. HMM-based phoneme speech recognition system for the control and command of industrial robots. *arXiv* **2020**, arXiv:2001.01222. [CrossRef]
20. Chiu, C.-C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
21. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006.
22. Radfar, M.H.; Barnwal, R.; Swaminathan, R.V.; Chang, F.J.; Strimel, G.P.; Susanj, N.; Mouchtaris, A. ConvRNN-T: Convolutional Augmented Recurrent Neural Network Transducers for Streaming Speech Recognition. *arXiv* **2022**, arXiv:2209.14868.
23. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4960–4964.
24. Miao, H.; Cheng, G.; Gao, C.; Zhang, P.; Yan, Y. Transformer-Based Online CTC/Attention End-to-End Speech Recognition Architecture. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6084–6088.
25. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 5036–5040.
26. Kiperwasser, E.; Goldberg, Y. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 313–327. [CrossRef]
27. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Republic of Korea, 1–3 November 2017; pp. 1–5.
28. Zhang, B.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; Lei, X. WeNet: Production First and Production Ready End-to-End Speech Recognition Toolkit. *arXiv* **2021**, arXiv:2102.01547.
29. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
30. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
31. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2015**, arXiv:1412.6980.
32. Zhang, H.; Yuan, T.; Chen, J.; Li, X.; Zheng, R.; Huang, Y.; Chen, X.; Gong, E.; Chen, Z.; Hu, X.; et al. PaddleSpeech: An Easy-to-Use All-in-One Speech Toolkit. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, Online, 10–15 July 2022.
33. Tian, J.; Yu, J.; Weng, C.; Zou, Y.; Yu, D. Improving Mandarin End-to-End Speech Recognition with Word N-Gram Language Model. *IEEE Signal Process. Lett.* **2022**, *29*, 812–816. [CrossRef]
34. Che, W.; Feng, Y.; Qin, L.; Liu, T. N-LTP: An Open-source Neural Language Technology Platform for Chinese. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021.
35. Chen, L.-W.; Rudnicky, A. Fine-grained style control in transformer-based text-to-speech synthesis. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022.
36. Saba, R.; Bilal, M.; Ramzan, M.; Khan, H.U.; Ilyas, M. Urdu Text-to-Speech Conversion Using Deep Learning. In Proceedings of the 2022 International Conference on IT and Industrial Technologies (ICIT), Chiniot, Pakistan, 3–4 October 2022; pp. 1–6. [CrossRef]