Multi-granularity Semantic and Acoustic Stress Prediction for Expressive TTS

Wenjiang Chi* Xiaoqin Feng* Liumeng Xue† Yunlin Chen* Lei Xie† Zhifei Li*
* Shanghai Mobvoi Information Technology Co., Ltd, China.
† Audio, Speech and Language Processing Group (ASLP)

Stress, as the perceptual prominence within sentences, plays a key role in **expressive text-to-speech** (TTS). It can be either the semantic focus in text or the acoustic prominence in speech. However, stress labels are always annotated by listening to the speech, lacking semantic information in the corresponding text, which may degrade the accuracy of stress prediction and the expressivity of TTS. This paper proposes a **multi-granularity stress prediction method** for expressive TTS. Specifically, we first build Chinese Mandarin datasets with both coarse-grained semantic stress and fine-grained acoustic stress. Then, the proposed model progressively predicts **semantic stress** and **acoustic stress**. Finally, a TTS model is adopted to synthesize speech with the predicted stress. Experimental results on the proposed model and synthesized speech show that our proposed model achieves good accuracy in stress prediction and improves the expressiveness and naturalness of the synthesized speech.

ABSTRACT

https://xqfeng-josie.github.io/stress

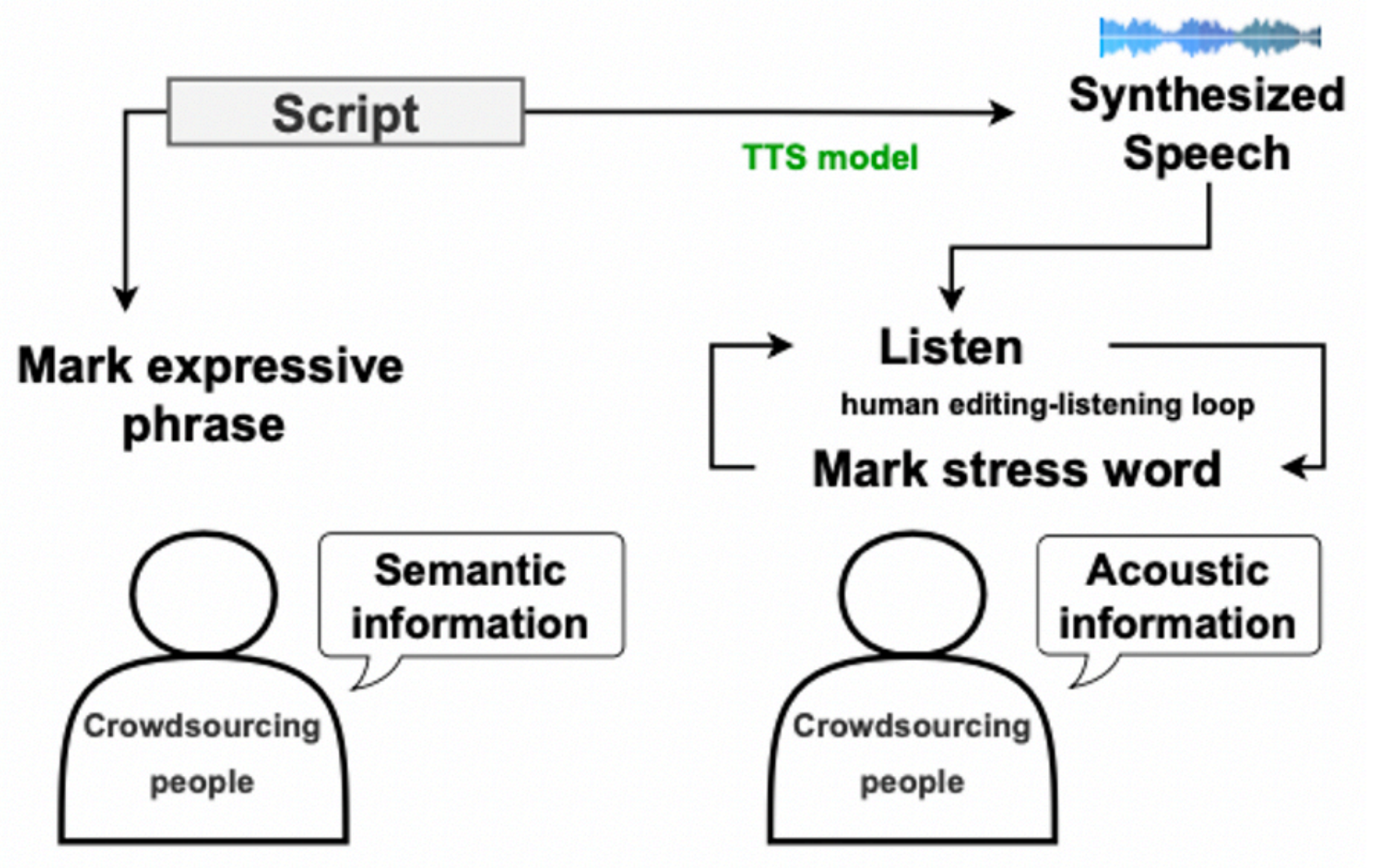
INTRODUCTION

Prosody, which is crucial and essential for expressive speech synthesis, is mainly composed of rhythm, **stress** and intonation. Therefore, accurately modeling stress is beneficial to improve the expressiveness and naturalness of TTS. With the significant improvement of pre-trained language models, such as BERT, the learned contextualized word representation is involved in rich semantic and syntactic cues of the text, which could be beneficial to the stress prediction and the downstream TTS task.

CONTRIBUTION

- We propose a **stress dataset construction** method with multi-grained stress annotations.
- we propose a **multi-granularity stress prediction model** for TTS, which combines the semantic information of text and acoustic information of speech.
- Our proposed multi-granularity stress prediction model achieves more accurate performance and the generated speech based on the predicted stress is more **expressive** and **natural**.

DATASET CONSTRUCTION



- No **voice actor** involved, affordable.
- No **limited** recording enviroment, pre-designed scripts. Suitable for **real-world** scenarios.
- **Multi-granularity** stress are considered.

年 轻 的 母 亲 暴 跳 如 雷， 竟 然 打 翻 了 桌子! Semantic-Stress
年 轻 的 母 亲 暴 跳 如 雷， 竟 然 打 翻 了 桌子! Acoustic- Stress
The young mother gets into a rage, unexpectedly knocking over the table!

EXPERIMENTS

Objective evaluation results: micro-F1, micro-precision and micro-recall

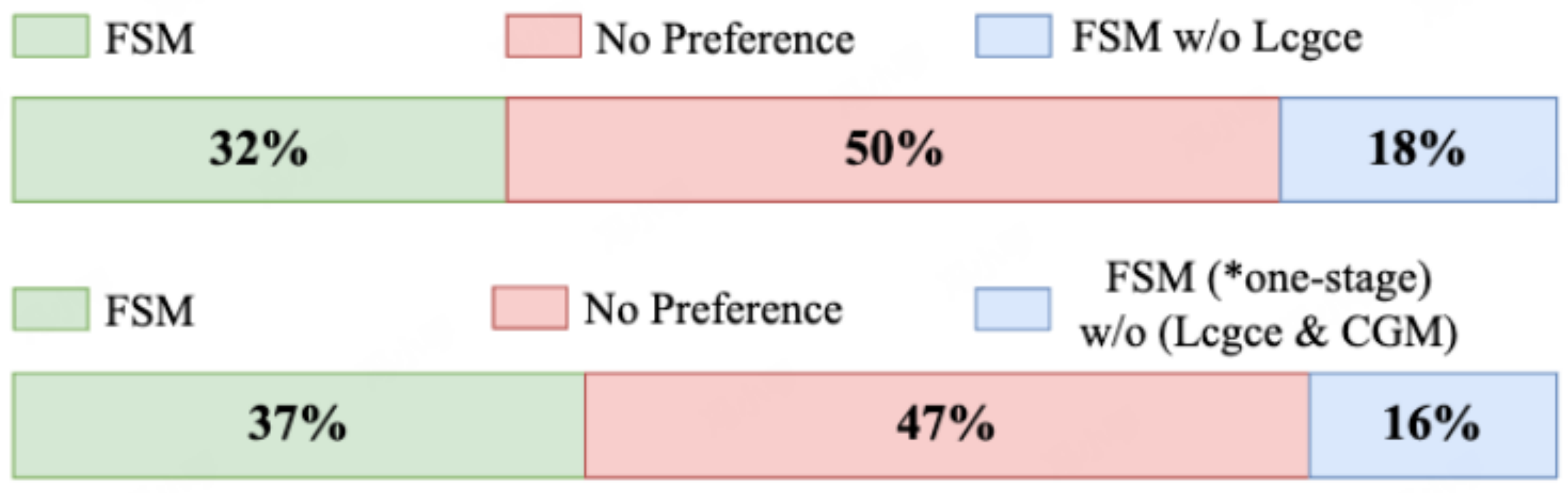
Model	Precision	Recall	F1
CGM	0.8471	0.9061	0.8756
FSM	0.7593	0.6762	0.7153
FSM (*one-stage)	0.6218	0.6284	0.6251
BERT-base	0.6050	0.5737	0.5890
3-layer BLSTM	0.0100	0.8471	0.0197
CRF	0.5278	0.0252	0.0482

Subjective evaluation results: NMOS and EMOS

Type	Original	ManualSet	CGM	CGM_{random}	FSM
NMOS	3.86	3.98	3.75	3.83	**3.95**
EMOS	3.9	4.08	3.83	3.88	**4.07**

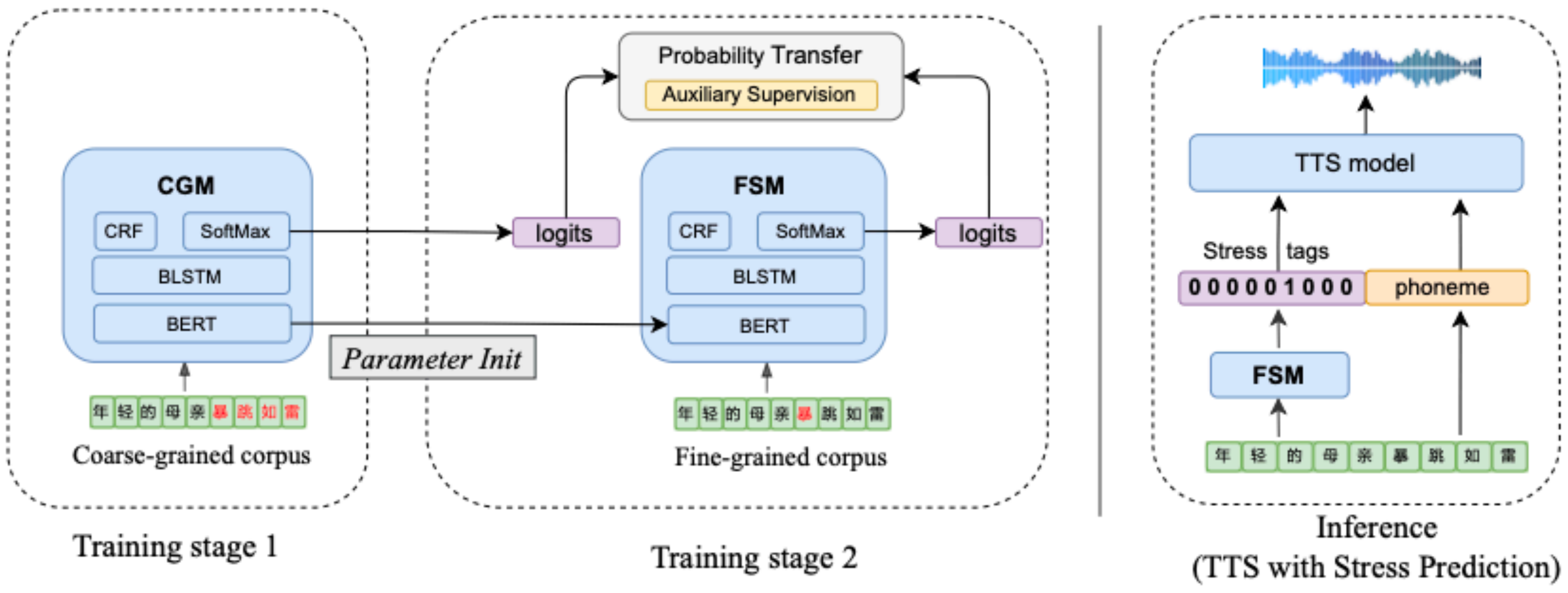
ABLATION STUDIES

- **w/o *Lcgce***: auxiliary supervised loss
- **w/o CGM**: two-stage training strategy.



It can be seen that the initialization of CGM can enhance the effect of the FSM a little bit. However, the auxiliary supervision loss of CGM **significantly** improves the performance of the FSM. Overall, the results demonstrate that the incorporation of coarse-grained information supervision is an **effective** approach to preventing stress weight dispersion during training.

MULTI-GRANULARITY STRESS PREDICTION



- The **overall architecture** of our proposed model, where the character in red is ground-truth stressed words. The illustration for the calculation of auxiliary supervised loss in training stage 2.

- **Training Stage 1: Coarse-grained Semantic Stress Prediction**
  The coarse-grained stress model (CGM) is trained on the coarse-grained corpus to identify the salient entities within a sentence.
- **Training Stage 2: Fine-grained Acoustic Stress Prediction**
  The fine-grained stress model (FSM) is trained on the fine corpus to combine with the semantic stress information and get the final fine-grained stress.
- **TTS with Predicted Stress**
  The TTS model adopts the framework of VITS, which is able to synthesize speech with stressed words by feeding the stress tags. FSM takes the text as input and outputs the stress tag for each character in the text.

The Key Point: **How to combine semantic and acoustic stress information?** We use two ways to transfer the semantic stress information and keep the acoustic stress information.

$$L_{cgce} = -\sum_{i=1}^{seq\_len}\sum_{k=1}^{2} y_{ik}\log(F2)$$

CGM → **FSM**

- **Parameter Initialization**
- **Auxiliary Supervised Loss**