AASP-P8.1

CONFIDENCE-BASED FILTERING FOR SPEECH DATASET CURATION WITH GENERATIVE SPEECH ENHANCEMENT USING DISCRETE TOKENS

Kazuki Yamauchi[1,2], Masato Murata[1], Shogo Seki[1]
[1]CyberAgent, Japan, [2]The University of Tokyo, Japan

CyberAgent AI Lab

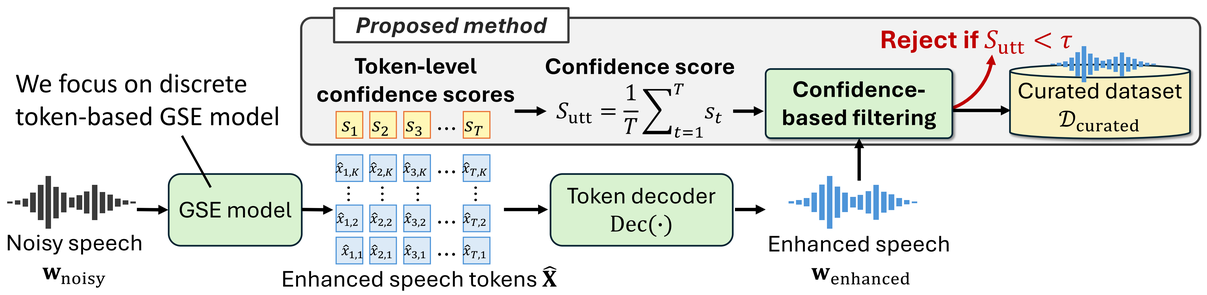
1. Introduction

- **Generative Speech Enhancement (GSE)**
  - Generative model-based SE method using a module originally developed for text-to-speech (TTS)
  - **Application:** Cleaning in-the-wild TTS dataset (e.g., LibriTTS-R [Y. Koizumi+23], FLEURS-R [M. Ma+24])
  - **Challenge:** Cause ***"hallucination errors"*** such as **phoneme omissions** and **speaker inconsistencies**
    - Speech that fails to be emphasized can **negatively impact the training of TTS models**

Propose a *non-intrusive* method for detecting and filtering hallucination errors in GSE

Not require clean reference speech and ground-truth transcriptions

2. Proposed Method



- **Step 1. Compute token-level confidence scores** s_1 s_2 s_3 … s_T
  - $s_t = \log p(x_{t,1} = \hat{x}_{t,1} \mid \mathbf{w}_{\text{noisy}};\ \theta)$ …… Log probability of each token
- **Step 2. Compute utterance-level confidence score** S_{utt}
  - $S_{\text{utt}} = \frac{1}{T}\sum_{t=1}^{T} s_t$ …… Average of token-level confidence scores
- **Step 3. Filter out the outputs based on** S_{utt}
  - $\mathcal{D}_{\text{curated}} = \{\mathbf{w}_{\text{enhanced}} \mid S_{\text{utt}} \geq \tau\}$ …… **Reject sample if S_{utt} is smaller than the threshold τ**

3. Experiments

Exp. 1: Validation of the confidence score as a GSE metric

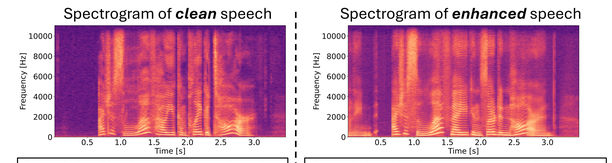
But cannot be used during cleaning in-the-wild dataset because the clean reference does not exist

➢ Measure SRCC between the confidence score and **intrusive metrics capable of detecting hallucination**

- **GSE model**: Genhancer [H. Yang+24]
  - Pretrained on LibriTTS-R, noise and impulse response data
- **Dataset**: EARS-WHAM [J. Richter+24]
  - Simulated noisy and clean speech (100 hours)
- **Non-intrusive metrics for comparison**:
  - UTMOS-out/in, DNSMOS-out/in, Whisper confidence-out/in
  - **"-in"** … for noisy input, **"-out"** … for enhanced output
- **Intrusive metrics for evaluation**:
  - SI-SDR, PESQ, SpeechBERTScore (SBS), phoneme similarity (LPS), Word Accuracy (WAcc), speaker similarity (SpkSim)

SRCC between non-intrusive and intrusive metrics on EARS-WHAM

Metric	SI-SDR	PESQ	SBS	LPS	WAcc	SpkSim
UTMOS-out	0.540	0.606	0.656	0.737	0.610	0.512
UTMOS-in	0.179	0.604	0.491	0.467	0.423	0.501
DNSMOS-out	0.181	0.338	0.427	0.330	0.323	0.383
DNSMOS-in	0.381	0.720	0.614	0.569	0.546	0.639
Whisper confidence-out	0.529	0.676	0.736	0.770	**0.766**	0.636
Whisper confidence-in	0.277	0.420	0.438	0.500	0.504	0.365
Confidence score (ours)	**0.590**	**0.883**	**0.892**	**0.788**	0.730	**0.790**



✓ Genhancer confidence has **the highest correlation with all intrusive metrics** except WAcc

✓ Whisper confidence, which has a high WAcc, has **a low correlation with other indicators**

✓ **Our method can detect hallucinations** that are difficult to detect with conventional methods (e.g., UTMOS-out)

Exp. 2: Applicability to in-the-wild TTS dataset curation

➢ Compare TTS models trained on curated dataset ***with and without*** **the confidence-based filtering**

- **TTS model**: Matcha-TTS [S. Mehta+24]
- **Dataset**: TITW-hard [J.-W. Jung+25]
  - In-the-wild noisy speech from YouTube (189 hours)

	#samples	UTMOS ↑	DNSMOS ↑	WER (%) ↓
Source (noisy)	280,130	2.73	2.74	21.31
Enhanced (unfiltered)	280,130	3.64	3.10	20.45
Enhanced (top 90%)	252,117	3.76	3.14	19.29
Enhanced (top 80%)	224,104	**3.80**	**3.17**	18.79
Enhanced (top 70%)	196,091	3.76	3.15	**18.14**
Enhanced (top 60%)	168,078	3.73	3.12	18.78
Enhanced (top 50%)	140,065	3.68	3.11	20.18

✓ Models trained on datasets filtered with confidence score outperformed models trained on unfiltered datasets
  ➢ **Enhancement errors negatively impact TTS training, and the proposed method can remove them**

4. Conclusion

- **Confidence-based filtering for discrete token-based GSE**
  - Propose a non-intrusive method for detecting and filtering hallucination errors in GSE
- **Future work:**
  - Extend our method to **a general approach not limited to discrete token-based GSE**