Inappropriate Pause Detection In Dysarthric Speech Using Large-Scale Speech Recognition

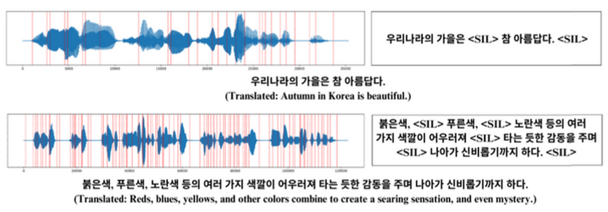
Jeehyun Lee,[1*] Yerin Choi,[1*] Tae-Jin Song,[2] Myong-Wan Koo[1]
Department of Artificial Intelligence, Sogang University, South Korea[1]
Ewha Womans University College of Medicine, Seoul, Republic of Korea[2]
* Equal Contribution

Overview

Introduction

- **Post-stoke dysarthria** is a common issue among stroke patients, severely impairing speech control.
- **Inappropriate pause (IP)** refers to delays that occur in untypical locations. Pauses occur unexpectedly, such as in the middle of a noun phrase, resulting in **reduced speech intelligibility.**
- **Inappropriate pause** is one of the factors in assessing dysarthria severity.

Existing Methods



- Current assessment relies on time-consuming auditory evaluation, lacking automatic inappropriate pause detection.
- Phoneme segmentation is costly, but a significant amount of labeled corpus is necessary to ensure reliable performance.
- Forced alignment requires precise transcription, leading to a two-step process.

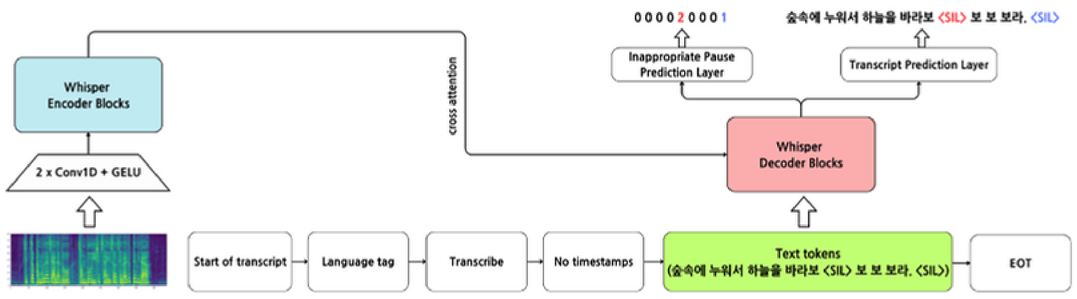
Our Approach

- We improve pause detection in dysarthric speech by **treating pauses as distinct tokens in the E2E ASR model** and extending it with an **IP prediction layer**.

Inappropriate Pause Detection Methods

Model Architecture

$$Z = \text{Seq2Seq}(X)$$
$$V' = \text{TranscriptPrediction}(Z)$$
$$IP' = \text{IPPrediction}(Z)$$
$$L = \text{Cross-entropy}(V') + \text{Soft-dtw}(IP')$$



- We extend a large-scale ASR model for inappropriate pause detection in dysarthric speech.
- Two separate layers:
  - **Transcript prediction layer:** We utilize a **Seq2Seq** architecture to transform speech input into **text with pause tags.**
    - Pause is added as a special token in Whisper
  - **Inappropriate pause prediction layer:** It determines the appropriateness of each token, classifying them as **appropriate** pauses, **inappropriate** pauses, or **non**-pauses (words).

Labeling Inappropriate Pauses

Original Text	느담아 <SIL> 어 <SIL> 완만함과 <SIL> 깎아 놓은 듯한 <SIL> 모옥함이 <SIL> 어우러 <SIL> 진 <SIL> 안다엉이 따라 <SIL> 오르다 보머 <SIL> 널로 안마 <SIL> 으아 수가 없게 <SIL>
Yale Romanization	nutama <SIL> e <SIL> wanmanhamkwa <SIL> kkakka nohun tushan <SIL> mookhami <SIL> ewule <SIL> cin <SIL> antaengi ttala <SIL> oluta pome <SIL> nello anma <SIL> ua swuka epskey <SIL>

- We add <SIL> tags to indicate pause locations in the text level.
- For each pause, we annotate its appropriateness
- Our method is much simpler than other time segmentation annotation, and can be applied other languages.

- Inappropriate Pauses Annotation Criteria
  - Pause within a noun phrase
    - a patient cannot say a word in a single breath
  - Pauses following vocalic surplus expressions such as “uh” or “um”
    - only when accompanying unclear wording or incorrect pronunciation and excessively long pauses (longer than three seconds)
  - Pauses that occur to rectify mispronunciation

Experimental Results

Dataset

Severity	w/o dysarthria	Mild-to-Moderate	Severe	Total
# of Utterances	72	1985	194	2251

- Korean dysarthric speech corpus with the *Autumn paragraph,* containing all necessary consonants and vowels for evaluation.
- NIH Stroke Scale

Evaluation Metrics

- ASR: WER, CER
- Pause Detection: PauER, IPER
  - **PauseSeq / IPSeq**
  - Seprate ASR performance from Pause / IP Detection
  - CER with PauseSeq, IPSeq

Original Text	느담아 <SIL> 어 <SIL> 완만함과 <SIL> 깎아 놓은 듯한 <SIL> 모옥함이 <SIL> 어우러 <SIL> 진 <SIL> 안다엉이 따라 <SIL> 오르다 보머 <SIL> 널로 안마 <SIL> 으아 수가 없게 <SIL>
Yale Romanization	nutama <SIL> e <SIL> wanmanhamkwa <SIL> kkakka nohun tushan <SIL> mookhami <SIL> ewule <SIL> cin <SIL> antaengi ttala <SIL> oluta pome <SIL> nello anma <SIL> ua swuka epskey <SIL>
PauseSeq	0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 1
IPSeq	0 2 0 2 0 1 0 0 0 1 0 1 0 2 0 1 0 0 1 0 0 1 0 0 2 0 0 0 1

Baselines

- Forced Alignment
  - MFA-GT: aligns the GT (human) transcription & speech
  - MFA-Whisper: aligns the ASR transcription from Whisper & speech
  - MFA-Dysarthric-Whisper: aligns the ASR transcription from fine-tuned Whisper & speech
- Conformer-RNNT with Pause tags

Results

	WER(%)	CER(%)	PauER(%)
MFA-GT	-	-	11.14
MFA-Whisper	54.89	27.35	22.49
MFA-Dysarthric-Whisper	32.21	22.38	17.27
Conformer-RNNT	64.52	49.99	22.81
Ours	**25.31**	**11.96**	**3.077**

- Our method outperforms others in both ASR & Pause detection, with lower PauER compared to MFA-GT.
- Higher ASR performance correlates with improved pause detection, indicating a complementary relationship between the two tasks.

Severity	WER(%)	CER(%)	PauER(%)	IPER(%)
Total	25.31	11.96	3.07	14.47
w/o dysarthria	6.93	2.89	2.48	20.69
Mild-to-Moderate	22.38	10.20	3.03	15.53
Severe	57.44	30.47	3.60	13.40

- IP (Inappropriate Pause) detection operates robustly across severity levels.
- As severity increases, ASR performance declines, suggesting the data was sufficient for IP detection but insufficient for ASR.

Conclusion

1. **An End-to-End Solution**: Our model can detect pauses, predict inappropriate pauses, and transcribe dysarthric speech at the same time without any post-processing.
2. **Collaborative Criteria Establishment for Inappropriate Pauses:** Collaboration with speech-language pathologists ensures the annotated data aligns closely with clinical expertise.
3. **Real-world Applicability:** Our approach demonstrates superior pause detection compared to baselines, highlighting the method’s effectiveness across different dysarthria severity levels.

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00621,Development of artificial intelligence technology that provides dialog-based multi-modal explainability