# 대규모 음성인식 모델을 이용한 마비말 장애 환자 음성의 쉼 구간 탐지 및 설명 제공 파이프라인

이지현 ○ 최예린 ○ 구명완

서강대학교 인공지능학과

{jhlee22,lakahaga,mwkoo}@sogang.ac.kr

# Detection of Break Index in Dysarthric Speech Using Large-Scale Speech Recognition and Explanation Pipeline

Jeehyun Lee○ Yerin Choi○ Myoung-Wan Koo

Department of Artificial Intelligence, Sogang University

**요 약**

본 논문에서는 마비말 장애 환자 쉼 구간 예측을 음성인식 태스크로 정의하고, 음성 기반 Seq2Seq 모델을 적용하여 음성을 입력으로 받아 쉼 구간 태그를 포함하는 텍스트를 출력하는 구조를 제안한다. 대용량 음성인식 모델의 사전 학습 지식을 활용해 쉼 구간 예측 태스크에 맞게 task-adaptive fine-tuning을 진행하였다. 실험 결과 음성인식-텍스트 모델 파이프라인과 비교하여 효과적으로 문장 내 쉼 구간의 위치를 탐지할 수 있음을 보였고, 특히 마비말 장애 환자 음성에서의 활용 가능성을 보였다. 제안한 방법은 쉼 구간 라벨링을 단순화하여 효율적인 마비말 장애 환자 음성 코퍼스 라벨링 방법론의 구축을 가능하게 했다. 또한, 환자 음성에 대한 성능을 고도화하기 위해 여러가지 훈련 방법에 대한 실험을 진행하였다. 마지막으로 탐지된 쉼 구간은 정상인 음성의 결과와 비교하여 환자에게 적절한 설명을 제공할 수 있는 파이프라인을 제안한다. 이를 통해 마비말 장애 환자의 언어 훈련 및 음성 인터페이스 기술 발전에 기여할 수 있기를 기대한다.

## 1. 서 론

뇌졸중으로 인한 마비말 장애(dysarthria)는 중추 및 말초신경의 이상으로 호흡, 발성, 조음, 운율 등의 산출 과정에서 말 조절 근육의 마비, 약화, 불협이 나타나는 것으로, 언어 이해와 말 명료도가 저하되는 현상을 말한다. 뇌졸중 환자들의 약 1/3~1/2에서 언어 장애가 발생[1]하는 만큼, 뇌졸중 치료와 재활을 위해 마비말 장애 정도를 평가하는 것은 매우 중요하다. 마비말 장애의 구어 평가에는 모음 연장 과제, 단어 과제, 문맥 발화 과제 등이 있다. 본 논문에서는 그 중 문맥 발화 과제, 특히 표준 문단 읽기 과제에 초점을 맞추었다. 문단 읽기 과제에서는 올바른 쉼의 구간과 비교하여 부적절한 쉼 여부 및 빈도를 분석한다. 분석한 결과는 환자에게 다시 전달되어 더 훈련해야 할 부분에 대하여 피드백을 제공한다. 이를 통해 환자의 지속적인 훈련이 이루어진다. 이와 같이 피드백을 제공하는 것은 마비말 장애 치료 과정에서 중요한 부분이며 언어치료사가 대면으로 진행하고 있다. 본 논문에서는 인공지능 모델이 언어치료사의 보완재로 쓰일 수 있도록 쉼 구간 탐지 - 설명 제공의 파이프라인을 제안한다.

부적절한 쉼은 문단 읽기 시 발생하는 일반적인 쉼 구간과 비교해 부적절한 쉼 여부와 빈도를 분석하는 지표이다. 마비말 장애 환자는 발화 시 쉼의 위치나 양을 적절히 사용하지 못해 말 명료도가 저하된다[2]. 이러한 쉼 구간은 가을문단[3]을 통해 평가된다. 환자가 문단을 읽을 때 어절을 한 호흡에 읽지 못하고 쉬는지 여부를 확인한다. 쉼 구간 적절성 판단을 위해서는 음성을 이용해 문장 내 쉼 구간을 탐지하는 작업을 진행하고 이에 대한 설명을 제공하는 것이 필요하다.

TIMIT[4] 등 쉼 구간을 라벨링한 기존 코퍼스는 쉼 구간 위치를 음성에서의 시간으로 나타낸다. 이러한 쉼 구간 라벨은 문장에서의 위치를 파악하기 어렵다는 한계가 있다. 문장 내 위치를 판단하려면 음소 구간 라벨링이 필요한데 이는 시간 및 비용이 많이 든다. 특히 마비말 장애 환자 데이터는 일반 데이터에 비해 라벨링이 어렵기 때문에 음소, 쉼 구간의 구간 라벨링에서 더 많은 비용이 필요하며, 대량의 코퍼스 구축이 어렵다.

본 논문에서는 위와 같은 문제를 해결하기 위해 쉼 구간을 찾는 문제를 음성인식 문제로 정의한다. 쉼 구간을 언어모델에 하나의 토큰(token)으로 등록하여, 음성을 입력으로 받아 쉼 구간 표시를 포함하는 텍스트를 출력한다. 이를 통해 쉼 구간 라벨링 방법을 단순화하면서도 효과적인 쉼 구간 탐지가 가능하도록 하였다. 다음으로 환자에게 적절한 설명을 제공하기 위하여 이를 정상인의 발화와 비교하여 시각화하는 방법을 제안한다.

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

본 논문의 주요 기여점은 다음과 같다. 먼저, 쉼 구간 탐지를 음성인식 태스크로 정의하여 효과적인 예측 모델을 개발하였다. 이를 통해 쉼 구간의 라벨링 방법을 단순화 하였다. 또한, 대규모 음성 모델의 사전 학습된 음성 지식을 사용하여 소량 데이터로도 환자 음성에서 문장 단위의 쉼 구간 위치를 탐지할 수 있도록 하였다. 또한 해당 태스크에서 마비말 장애 환자 음성에서도 좋은 성능을 낼 수 있도록 여러가지 실험을 진행하였다. 쉼 구간 탐지와 더불어 결과에 대한 설명성 제공하는 파이프라인을 구성하였다. 이는 현재 언어치료사의 음성 듣기 - 분석 - 피드백 제공을 자동화하는 파이프라인으로, 이를 통해 환자들의 지속적인 훈련의 시간적, 공간적 제약을 완화하는 데 일조하기를 기대한다.

## 2. 쉼 구간 탐지를 위한 코퍼스

### 2.1 쉼 구간 탐지를 위한 정상인 및 마비말 장애 환자 코퍼스

기존 쉼 구간 라벨링은 주로 쉼 구간의 시점을 기록하는 방식이다. 이러한 음성 수준에서의 시간 단위 라벨링은 비용이 많이 들고, 문장 내 쉼 구간 위치를 탐지하기 위해서는 음성과 텍스트의 정렬(alignment)이 필요하다. 반면, 본 코퍼스는 표 1에 나타난 예시와 같이 텍스트 수준에서 쉼 구간이 나타나는 위치를 [bi] 태그로 표시하여 라벨링을 단순화하고, 쉼 구간 예측을 보다 효율적으로 수행하도록 하였다.

**표 1 쉼 구간 라벨링 예시**

| | 데이터 예시 |
|---|---|
| 정상인 | 성격이 변했다니 다행이구나, [bi] 커서도 어릴 때처럼 자기만 알면 어쩌나 걱정했어. [bi] |
| 마비말 장애 환자 | 무엇보다도 탄에저 [bi] 오른디 댁 더욱 [bi] 기쁘다 [bi] 그 빼어난 아름다움이 느그진다. [bi] |

정상인 코퍼스는 여성 단일 화자 발화 코퍼스이고, 발화 기준 총 13,000개다. 환자 코퍼스는 뇌졸중을 가진 환자가 가을문단을 읽은 음성으로 구성되어 있으며, 발화 기준 2,251개이다. 뇌졸중 환자의 마비말 장애는 그 심각도를 3개 척도로 나눈다. 0은 뇌졸중 환자 중 마비말 장애가 나타나지 않은 경우다. 1,2 는 뇌졸중 환자 중 마비말 장애가 있으며 2의 심각도가 가장 높다. 해당 척도 별 음성 수량은 표 2에 나타난다. 정상인은 비공개 코퍼스이며, 환자 코퍼스는 공개 예정이다.[1]

**표 2 마비말 장애 환자 코퍼스의 심각도 척도 별 데이터 수량**

| 심각도 | 0 | 1 | 2 | Total |
|---|---|---|---|---|
| 발화 수 | 72 | 1985 | 194 | 2251 |

## 3. 쉼 구간 탐지 방법

### 3.1 쉼 구간 탐지 문제 정의

언어 재활사가 부적절한 쉼을 판단하기 위해 환자 음성을 듣고 쉼 구간을 확인하는 것에 착안하여 쉼 구간 예측을 음성인식 태스크로 설정하였다. 음성 발화가 입력으로 들어오면 쉼 구간 표시를 포함한 텍스트를 출력한다. 실제 언어 재활사의 평가 과정과 태스크를 유사하게 설정하여 마비말 장애 환자의 문단 읽기 과제에서의 활용 가능성을 확인하고자 했다.

### 3.2 모델 구조

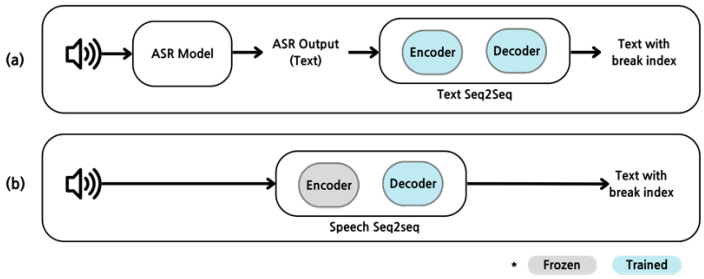


**그림 1 (a) 음성인식 후, 그 결과를 텍스트 기반 시퀀스 투 시퀀스 모델에 넣어 쉼 구간을 탐지하는 파이프라인, (b) 음성 기반 시퀀스 투 시퀀스 모델을 이용하여 음성에서 바로 쉼 구간을 포함하는 텍스트를 생성하는 모델**

본 논문에서 제안하는 최종 모델 구조는 그림 1(b)와 같은 시퀀스-투-시퀀스(Seq2Seq) 구조로, 음성이 입력되면 변형된 텍스트(쉼 구간 태그를 포함하는 텍스트)를 생성한다. Seq2Seq 구조를 활용할 경우 음성인식 결과와 텍스트 라벨의 길이가 다른 경우에도 효과적인 평가가 가능하며, 환자 데이터 라벨링에도 활용할 수 있다. 구체적으로 Huggingface의 Whisper For Conditional Generation 구조를 이용하였으며, 인코더는 freeze하고 디코더만 학습하여 쉼 구간을 포함한 텍스트 생성이라는 목적에 집중했다. 단순한 구조를 사용하여 음성에서 바로 쉼 구간을 탐지할 수 있도록 하였다.

### 3.3 음성인식-텍스트 파이프라인과 음성 Seq2Seq 비교

음성을 입력으로 받아 변형된 텍스트를 출력하는 방법에는 그림 1(a)와 같이 음성인식 모델을 이용하여 전사문을 추출한 뒤, 텍스트 기반 Seq2Seq 모델을 적용하는 파이프라인 형식도 있다. 하지만 이 구조는 두 가지 모델을 거치기 때문에 오류가 전파된다. 또한, 가장 중요한 태스크인 쉼 구간 탐지에서 텍스트만 활용하기 때문에 음성적 특징을 반영하지 못한다. 5.3절에서 음성인식-텍스트 모델 파이프라인과 음성 기반 Seq2Seq 모델의 비교 실험을 진행하였다.

### 3.4 마비말 장애 환자 음성에 대한 성능 고도화

마비말 장애 환자의 음성은 정상인 음성과 확연한 차이를

[1] 이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "No.2022-데이터-위81, 인공지능 학습을 위한 응급실 임상 대화 및 구음장애인 명령어 데이터 수집 사업"을 활용하여 수행된 연구임. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받을 수 있음.

보인다. 마비말 장애 환자는 그 심각도에 따라 정도는 다르지만 공통적으로 다음과 같은 특징을 보인다. 발음이 부정확하며 발화 속도가 정상인 발화에 비해 느리다. 또한, 음고(Pitch)가 너무 높거나 낮은 이상치가 나타나기도 하고 목이 쉰 소리, 흡인된 음성(wet voice)가 나타난다. 이와 같이 정상인 음성과 마비말 장애 환자의 음성은 서로 다른 특징을 가진 데이터이기 때문에 성능을 고도화하기 위해서는 마비말 장애 환자 음성에 대한 추가 학습이 필요하다.

환자 음성에 대한 성능을 높이기 위하여 다음과 같은 훈련방법을 진행하였다. 먼저, 정상인 데이터를 학습한 모델의 파라미터를 환자 음성에 대해 업데이트하였다. 정상인 데이터를 학습할 때와 동일하게 파인튜닝(Fine-tuning)하는 방식, 어댑터(Adapter)를 이용하는 방식에 대한 실험을 진행하였다. 어댑터를 이용하는 방식은 기존 모델의 파라미터는 고정하고 새로운 구조를 추가하여 Parameter Interference를 최소화하면서도 새로운 지식을 추가적으로 학습할 수 있는 방법이다. 다음으로는 정상인 데이터 없이 환자 음성만을 이용하여 Whisper의 파라미터를 업데이트하는 방식에 대한 실험을 진행하였다. 해당 실험 결과는 5.4절에 나타난다.

## 4. 쉼 구간 탐지 결과를 이용한 설명 제공

쉼 구간 탐지 결과를 정상인 발화에 대한 결과와 비교하여 환자에게 설명을 제공한다. 정상인 발화와 비교하여 어느 부분이 다른 지에 대한 것을 시각화한다. 환자 음성에 대한 결과 뿐만 아니라 정상인 발화에 대한 결과, 즉 정답을 같이 제공한다. 환자가 훈련을 통해 나아가야 하는 방향을 알려주어 더욱 효율적인 훈련 과정이 되도록 한다. 해당 파이프라인 및 시각화 예시는 그림 2와 같다.

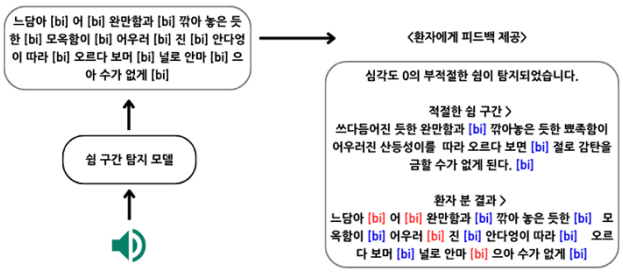


**그림 2 쉼 구간 탐지 – 설명 제공 파이프라인 예시**

## 5. 실험

### 5.1 실험 환경

학습에 사용한 코퍼스는 훈련, 검증, 테스트셋을 8:1:1의 비율로 나누었다. 수량은 훈련, 검증, 테스트 각각 10,400개, 1,300개, 1,300개다.

실험에 사용한 모델은 다음과 같다. 텍스트 모델은 KETI의 KE-T5-base[5] 모델을, 음성인식, 음성 기반 Seq2Seq 모델은 모두 OpenAI의 Whisper-small[6] 모델을 사용하였다. 훈련에 사용한 하이퍼 파라미터는 다음과 같다. AdamW optimizer에 learning rate는 5e-4로 설정하였다. 이는 음성인식-텍스트 파이프라인과 음성 기반 Seq2Seq 모델 모두 동일하게 사용하였다. 텍스트 기반 SeqSeq 모델은 454 epoch 학습하였고, 음성인식 결과에 추가 fine-tuning은 259 epoch까지 진행하여 텍스트 기반 모델은 총 713 epoch 학습했다. 음성 기반 Seq2Seq 모델은 총 226 epoch 학습하였다.

### 5.2 평가 지표

본 태스크를 평가하기 위해서 6가지 평가지표를 사용하였다. 텍스트 생성에 사용되는 ROUGE-1 precision, recall, f1-score 및 음성인식에 사용되는 문자 오류율(CER)을 이용하였다.

또한, 음성인식 결과와 무관하게 쉼 구간 탐지의 정확성을 판단하기 위해서 Break DTW를 제안한다. 텍스트 시퀀스에서 각 어절이 쉼 구간 태그 [bi]에 해당하면 1, 그렇지 않으면 0으로 변환하였다. 변환 예시는 표 3에 나타난다. 변환된 정답 라벨과 생성된 텍스트의 동적 시간 워핑(Dynamic Time Warping, DTW) 거리를 측정하였고 수식은 (1)과 같다. 정답 라벨과 생성된 텍스트가 길이가 다른 경우에도 유사성을 평가할 수 있도록 DTW를 사용하였다. 측정된 거리가 낮을수록 정답에 가깝게 쉼 구간을 예측한 것이다.

**표 3 쉼 구간 시퀀스 변환 예시**

| | 데이터 예시 |
|---|---|
| 기존 텍스트 | 성격이 변했다니 다행이구나, [bi] 커서도 어릴 때처럼 자기만 알면 어쩌나 걱정했어. [bi] |
| 변환된 시퀀스 | 0 0 0 1 0 0 0 0 0 0 0 1 |

$$BreakDTW = DTW(BISeq_{GT}, BISeq_{pred}) \quad (1)$$

### 5.3 음성인식-텍스트 파이프라인과 음성 Seq2Seq 실험

본 논문에서 제안하는 음성 Seq2Seq 모델의 성능을 평가하기 위해 비교 실험을 진행하였다. 평가 데이터는 5.1절에서 나눈 정상인 테스트셋, 마비말 장애 환자 코퍼스다. 이는 모두 모델이 훈련 과정에서 학습하지 않은 데이터다. 정상인 데이터에 대한 결과는 표 4에, 환자 데이터에 대한 결과는 표 5에 나타난다. '텍스트(1)'은 음성인식 결과에 텍스트 SeqSeq를 바로 추론한 결과이고, '텍스트(2)'는 음성인식 결과에 텍스트 SeqSeq를 추가 fine-tuning한 결과이다. 그리고 '음성'은 음성 기반 SeqSeq 모델의 결과이다.

먼저, 표 4와 같이 정상인 코퍼스에서 텍스트 기반 파이프라인 구조와 음성 Seq2Seq 모델의 성능을 비교하였다. 음성, 텍스트(2), 텍스트(1) 순으로 높은 성능을 보였다. 텍스트 모델은 음성인식 결과에 더 많은 epoch으로 추가 fine-tuning을 진행해도 음성 모델보다 낮은 성능을 보였다.

다음으로 환자 데이터에서의 추론 효과를 알아보기 위해 표 5와 같이 비교하였다. 마찬가지로 음성, 텍스트(2), 텍스트(1) 순으로 높은 성능을 보였다. 특히 Break DTW가 텍스트(1) 대

비 약 2, 텍스트(2) 대비 약 1.6 높은 성능을 보였다. 이는 정상인 코퍼스에서 보다 큰 차이로, 본 모델이 정상 코퍼스 뿐 아니라 환자 데이터에서 쉼 구간 예측이라는 태스크를 효율적으로 수행함을 보였다.

**표 4 정상인 코퍼스에 대한 모델 실험 결과**

| 모델<br>(학습 epoch) | 텍스트(1)<br>(454) | 텍스트(2)<br>(454+259) | 음성<br>(226) |
|---|---|---|---|
| ROUGE-1 precision | 90.65 | 95.75 | **98.47** |
| ROUGE-1 recall | 79.28 | 93.21 | **98.11** |
| ROUGE-1 F1 | 80.83 | 92.83 | **97.77** |
| CER | 56.76 | 12.07 | **7.75** |
| Break DTW | 0.67 | 0.28 | **0.0854** |

**표 5 마비말 장애 환자 코퍼스에 대한 모델 실험 결과**

| 모델<br>(학습 epoch) | 텍스트(1)<br>(454) | 텍스트(2)<br>(454+259) | 음성<br>(226) |
|---|---|---|---|
| ROUGE-1 precision | 91.08 | 91.23 | **91.54** |
| ROUGE-1 recall | 46.06 | 60.49 | **82.81** |
| ROUGE-1 F1 | 54.42 | 66.43 | **83.49** |
| CER | 76.74 | 63.48 | **47.07** |
| Break DTW | 4.28 | 3.83 | **2.25** |

### 5.4 마비말 장애 환자 음성에 대한 성능 고도화 실험

2장에서 설명한 마비말 장애 환자 코퍼스를 8:1:1로 나누어 훈련, 검증 그리고 평가에 사용하였다. 각 수량은 1800개, 225개, 226개이며, 심각도별로 그 비율을 유지하여 나누었다.

성능 고도화 실험은 음성 기반 Seq2Seq 모델에 대하여 진행하였다. 표 6에 3.4절에서 설명한 환자 음성에 대한 성능 고도화를 위한 각 학습 방법을 환자 음성 평가셋에 평가한 결과가 나타난다. Baseline은 정상인 데이터만을 학습한 모델이다. 정상인 데이터만을 학습하고 환자 음성에 대해 추론한 결과다. "**+ 환자 음성 추가 학습**" 이라고 표시한 모델은 Baseline 모델에 추가적으로 환자 음성을 동일한 파인튜닝 방법으로 학습한 모델이다. 다음으로는 Baseline 모델의 디코더, 인코더 각각에 어댑터를 추가하여 환자 음성을 학습한 모델을 "**+ Adapter (Decoder)**", "**+ Adapter (Encoder)**" 로 표시하였다. 마지막으로 "**환자 음성 학습**"은 정상인 데이터 학습 없이 환자 음성만으로 Whisper를 쉼 구간 탐지 태스크에 파인튜닝한 모델이다.

ROUGE-1 recall을 제외하고는 모든 지표에 대해 정상인 음성 데이터 없이 환자 음성만을 학습한 모델이 가장 성능이 좋았다. 특히, 정상인 데이터를 학습한 모델에 단순히 환자 음성을 추가 학습하면 성능이 오히려 떨어지는 결과를 관찰하였다. 어댑터를 추가하여 훈련하였을 때는 디코더보다는 인코더를 추가 학습하였을 때 더 좋은 성능을 냈다. 특히, 인코더에 어댑터를 추가하여 학습하였을 때 ROUGE-1 recall이 가장 좋았다. 이는 정상인 음성과는 다른 마비말 장애 환자 음성에 음향적, 내용적 특징을 더 학습할 필요가 있음을 시사한다.

### 6. 결론

본 논문에서는 마비말 장애의 쉼 구간 탐지를 음성인식 태스크로 규정하고, 대규모 음성인식 모델에 task-adaptive finetuning을 수행하였다. 이를 통해 보다 단순한 라벨링 작업으로도 환자 음성에 대한 쉼 구간 탐지를 효과적으로 수행함을 보였다. 제안된 라벨링 방법은 텍스트 수준에서 진행하기 때문에 그 난이도가 기존보다 낮아졌다. 또한 환자 음성에 대한 성능 고도화를 위해 여러가지 실험을 진행하였다. 마지막으로, 쉼 구간 탐지 모델에서 얻은 결과를 통해 환자에게 언어 훈련에 필요한 설명을 제공할 수 있는 파이프라인을 구성하였다. 앞으로는 소량의 환자 데이터에서 마비말 장애 환자의 음향적, 내용적 특징을 효과적으로 학습할 수 있는 방법이 더 연구되어야 할 것이다.

### 참 고 문 헌


[1]https://www.snubh.org/dh/main/index.do?DP_CD=BCD8&MENU_ID=005003

[2]한지연, 이옥분, 심이슬. 구강기류 분석에 근거한 정상 성인의 문단 읽기 시 호흡그룹의 특징. 음성과학. Dec;15(4):135-46. 2008

[3]김향희. 마비말장애 평가. 한국언어청각임상학회 언어장애 여름 연수회. 23~28. 2005.

[4]Garofolo, John S. "Timit acoustic phonetic continuous speech corpus." Linguistic Data Consortium, 1993 (1993).

[5]Kim, San, et al. "A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems." Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.

[6]Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *arXiv preprint arXiv:2212.04356* (2022).


**표 6 마비말 장애 환자 음성에 대한 성능 고도화 실험 결과**

| 모델 | Baseline | + Adapter (Decoder) | + Adapter (Encoder) | + 환자 음성 추가 학습 | 환자 음성 학습 |
|---|---|---|---|---|---|
| ROUGE-1 precision | 91.54 | 74.29 | 90.28 | 72.31 | **92.39** |
| ROUGE-1 recall | 82.81 | 73.37 | **92.84** | 87.66 | 90.31 |
| ROUGE-1 F1 | 83.49 | 69.47 | 90.07 | 73.51 | **90.35** |
| CER | 47.07 | 67.58 | 16.71 | 84.6 | **11.46** |
| Break DTW | 2.25 | 2.40 | 1.21 | 2.47 | **0.66** |