Journal of KIIT. Vol. 24, No. 3, pp. 159-167, Mar. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.3.159

ASL 지문자의 분리형 시퀀스 표현: 발화자 독립적 내용과 스타일 인식 요인

유승수*[1], Khoa Nguyen**[1], Thong-Nhat Tran**[2], 서영욱*[2]

Disentangled Sequence Representation for ASL Fingerspelling: Signer-Invariant Content and Style Factors

Seung-Su Yu*[1], Khoa Nguyen**[1], Thong-Nhat Tran**[2], and Young-Wook Seo*[2]

요 약

본 논문은 실시간 미국 수화(ASL, American Sign Language) 지문자 인식을 위한 효율적인 디코딩 방법을 제안한다. 기존 프레임 기반 방식은 매 시점 결과를 반복적으로 갱신하여 불필요한 연산과 지연이 발생할 수 있다. 이에 본 연구는 CTC(Connectionist Temporal Classification) 모델 출력의 로그 확률 변화량이 사전 정의된 임계값 δ를 초과하는 시점에서만 디코딩을 수행하는 이벤트 트리거 기반 접근을 적용하였다. 또한 시간 축에서의 예측 변동을 완화하기 위해 단조적 제약을 반영하여 오류 누적을 감소시켰다. 실험을 통해 제안 방법은 기존 방식 대비 계산량과 디코딩 지연을 줄이면서도 인식 정확도를 유지하거나 향상함을 확인하였다. 본 연구는 수어 기반 실시간 상호작용 시스템에 적용 가능한 실용적 방법을 제시한다.

Abstract

This paper proposes an efficient decoding approach for real-time American Sign Language (ASL) fingerspelling recognition. Conventional frame-based decoding updates predictions at every time step, causing redundant computation and increased latency. To address this, we adopt an event-triggered decoding mechanism that performs decoding only when the log-probability change of the Connectionist Temporal Classification (CTC) model output exceeds a predefined threshold δ. Monotonic constraints are incorporated to stabilize temporal predictions and mitigate error accumulation. Experimental results demonstrate that the proposed method reduces computational cost and decoding delay while maintaining or improving recognition accuracy compared to baseline approaches. The proposed framework is suitable for deploying ASL fingerspelling recognition in interactive and accessible sign-language-based text input systems.

Keywords

american sign language, fingerspelling recognition, real-time decoding, event-triggered decoding, monotonic constraint

* 대전대학교 융합컨설팅학과(*[2] 교신저자)
- ORCID[1]: https://orcid.org/0009-0003-6850-7078
- ORCID[2]: https://orcid.org/0000-0003-1700-4529
** 충북대학교 정보통신공학부
- ORCID[1]: https://orcid.org/0000-0002-6474-9681
- ORCID[2]: https://orcid.org/0009-0006-8840-8140

· Received: Feb. 05, 2026, Revised: Feb. 25, 2026, Accepted: Feb. 28, 2026

· Corresponding Author: Young-Wook Seo
Dept. of Department of Management Consulting, Deajeon University, Deajeon 34520, Korea
Tel.: +82-42-280-4185, Email: ywseo@dju.kr

Ⅰ. 서 론

미국 수화(ASL, American Sign Language) 핑거스펠링의 자동 인식은 포용적인 인간-컴퓨터 상호작용을 위한 핵심역량이다. 고유명사, 주소, URL, 사전에 없는 어휘 등이 문자 시퀀스로 전달되기 때문이다. 시각 인코더와 정렬 목적 함수의 강력한 발전에도 불구하고, 핑거스펠링은 빠른 연속 발음(Coarticulation), 자기 가림(Self-occlusion), 낮은 신호대 잡음비, 그리고 특히 수화자 변이성—손 모양, 동작 속도, 사용 손(Handedness), 카메라 자세의 광범위한 차이—때문에 여전히 어렵다.

표준 파이프라인은 연결주의 시계열 분류(CTC, Connectionist Temporal Classification) 인코더를 학습하고, 탐욕적(Greedy) 축소나 소규모 빔 탐색을 수행하며, 때로는 외부 언어 모델의 도움을 받는다. 현대 시스템에서는 트랜스포머와 컨포머 백본이 풍부한 시간적 문맥 모델링을 제공하며 프레임 동기화 목적 함수와 호환된다[1]. 디코딩은 종종 컴팩트한 문자 언어 모델과의 얕은 결합(Shallow fusion)을 통해 음향 신뢰도와 언어적 개연성 사이의 균형을 조정한다[2]-[4]. 동시에 GPT 시리즈와 같은 대규모 언어 모델의 발전은 강력한 사전학습 언어 사전(Prior)의 텍스트 이해 및 생성 능력을 보여준다[5]-[7].

스트리밍 핑거스펠링에서는 지연을 최소화하고 드문 혹은 도메인 특화 문자열에 대한 강건성을 확보하기 위해 경량 문자 모델을 선호한다. 그러나 단일 단계 접근법은 문자 시퀀스의 이산적 콘텐츠와 동작 속도나 발화 습관과 같은 연속적 스타일 요인을 명시적으로 분리하지 않으므로, 보지 못한 수화자에 대한 일반화가 제한되고 드문 문자열에서 보정이 저하될 수 있다. 표현이 수화자 불변의 콘텐츠(Content) 스트림과 라벨 정체성을 누설하지 않으면서 잡음 변이를 포착하는 보완적 스타일(Style) 스트림으로 분해되어야 한다고 주장한다. 시퀀스 도메인에서 비식별화와 불변성은 도메인 적대 학습, 모멘트 매칭, 대조 학습 목적 함수와 인스턴스 정규화, 적응적 특징 변조(Adaptive feature modulation) 같은 아키텍처 선택을 통해 연구되어 왔다[1]-[3][8]-[10].

특히 그래디언트 리버설을 사용하는 도메인 적대 학습은 보조 판별기를 주요 과제와 결합하여, 인코더가 인식에 도움이 되지 않는 속성을 버리도록 유도함으로써 도메인 특유 단서를 억제한다[11]. 자기지도 대조 학습은 동일 시퀀스의 다중 증강 뷰를 끌어당기고 다른 시퀀스의 뷰를 밀어내며, 시간 왜곡과 관련된 증강을 통해 다운스트림 디코딩에 유익한 시간 불변성을 유도할 수 있다[12][13]. 수화 인식에서 기존 연구는 주로 합성곱 및 트랜스포머 인코더, 수화자 독립 학습 커리큘럼, 개선된 손실을 통한 시각 백본과 정렬 향상에 초점을 맞추어 왔다[1][4][14][15]. 그러나 핑거스펠링 시퀀스를 콘텐츠와 스타일로 명시적으로 분해하고, 전용 스타일 스트림을 활용해 디코딩과 보정을 적응시키는 접근은 여전히 충분히 탐구되지 않았다.

본 연구는 Google-Kaggle ASL 핑거스펠링 인식 대회[16]를 기반으로 하며, 실용적 통찰은 은메달 수상작 솔루션[17]으로부터 얻었다. 프레임 동기 인코더와 CTC 기반 학습, 그리고 선택적으로 얕은 결합 문자 언어 모델을 활용한 최근의 발전은 ASL 핑거스펠링에서 안정적인 평균 성능을 제공하였다. 그러나 디코딩 단계는 보지 못한 수화자와 드문 문자열에 대한 강건성에서 여전히 핵심 병목으로 남아 있다[18]. 실제 배포 가능한 시스템은 동시에 세 가지 속성을 충족해야 한다: (i) 지연을 희생하지 않고 외부 언어 사전을 통한 언어적 지원, (ii) 빠르거나 비정형 동작에서의 길이 편향과 과도한 삽입을 완화하기 위한 보정 제어, (iii) 품질과 실시간 계수(RTF, Real-Time Fator) 간 균형을 맞추는 디코딩 하이퍼파라미터의 재현 가능한 선택[18]의 연구는 표준적인 탐욕적(Greedy) 또는 빔 디코딩 파이프라인을 따르며 집계 정확도를 보고하지만, 다중 언어 모델(LM, Language Model) 계열을 통합하거나 삽입 트레이드오프를 조정하거나 런타임 - 정확도 파레토 거동을 정량화하는 원칙적인 경로는 제공하지 않는다.

본 논문은 이러한 격차를 해소하기 위해 다음과 같은 기여를 제안한다: 1) 문자 n-그램과 선택적 신경 기반 GPT 언어 사전을 통합하는 통합 CTC 프리픽스 빔 탐색, 2) 안정적인 지연을 위한 프레임별

후보 리스트 프루닝, 3) RTF 제약 하에서 정확도를 최적화하는 엄격하고 재현 가능한 선택 프로토콜, 4) 탐욕적 디코딩 기준 Score 0.7344, RTF=0.002 (≈ 500 × 실시간)의 강력한 기준선 달성. β=0.4 조건에서 Score 0.7362로 탐욕적 기준선을 능가하며, k를 10에서 20으로 확장 시 Score +0.0003의 소폭 향상과 RTF 0.033 → 0.087의 연산-품질 트레이드오프를 정량화한다.

II. 문제 정의 및 성능 지표

본 연구는 ASL 핑거스펠링 인식을 CTC 프레임워크 하에서의 시퀀스 변환 문제로 정식화하며, 여기서 다중 모달 랜드마크 특징은 스트리밍 제약 조건 하에서 문자 시퀀스로 매핑된다. 본 절에서는 입력-출력 정식화, 언어 모델 결합을 통한 디코딩, 그리고 정확도-지연-보정을 위한 평가 지표를 제시한다.

2.1 CTC를 활용한 시퀀스 모델링

각 입력 핑거스펠링 클립은 식 (1)과 같이 랜드마크 프레임 시퀀스로 표현된다: 실제로 각 프레임은 입술, 손, 상반신 자세의 MediaPipe 기반 3D 랜드마크를 연결한 것이며, 사전 계산된 평균과 분산으로 표준화된다[19][20].

$$X=(x_1,x_2,\cdots,x_T),x_t\in R^F \tag{1}$$

여기서 T는 프레임 수, F는 특징 차원이다. 각 목표 라벨은 식 (2)와 같이 문자 시퀀스로 주어진다.

$$Y=(y_1,y_2,\cdots,y_L),y_l\in V \tag{2}$$

여기서 V는 59개의 기호(알파벳, 숫자, 구두점, 그리고 특수 블랭크/패드 토큰)로 이루어진 어휘를 나타낸다. 가중치 θ로 매개된 인코더 네트워크 f_θ는 X를 식 (3)과 같이 프레임 단위 로짓 시퀀스로 매핑한다.

$$Z=f_\theta(X)\in R^{T\times|V|} \tag{3}$$

학습은 식 (4)의 CTC 손실[21]을 사용하며, 이는 반복과 블랭크를 제거한 후 Y로 축소되는 모든 프레임 - 라벨 정렬에 대해 마지널라이즈한다.

$$L_{CTC}(\theta)=-\log\sum_{\pi\in B^{-1}(Y)}\prod_{t=1}^{T}p_\theta(\pi_t|x_{1:T}) \tag{4}$$

여기서 $p_\theta(c|x_{1:T})$는 인코더 파라미터 θ하에서 전체 입력 시퀀스 $x_{1:T}$가 주어졌을 때 프레임 t에서 기호 $c\in V$를 출력할 확률을 의미한다. π는 정렬 경로이며, $B(\pi)$는 CTC 축소 연산자이다.

2.2 이벤트 트리거 기반 디코딩

본 연구의 핵심 기여인 이벤트 트리거 기반 디코딩은, 프레임t에서 CTC 모델 출력의 로그 확률 변화량이 사전 정의된 임계값δ를 초과하는 경우에만 디코딩을 수행한다. 트리거 조건은 다음과 같이 정의된다. 여기서 $C^*=\arg max_c\ p\theta(c|x_{1-T})$는 프레임 t에서 가장 확률이 높은 기호이며, δ는 임계값 하이퍼파라미터이다. 본 실험에서는 δ = 0.1로 설정하였으며, $\delta\in\{0.05, 0.1, 0.2, 0.5\}$ 범위의 민감도 분석을 통해 δ = 0.1에서 정확도와 계산 효율 간 최적 균형을 확인하였다. δ가 작을수록 더 많은 프레임에서 디코딩이 수행되어 정확도는 유지되나 연산 부하가 증가하며, δ가 클수록 디코딩 빈도가 감소하여 효율성은 높아지지만 정확도가 하락할 수 있다.

2.3 외부 언어 모델과의 디코딩

추론 시 시스템은 프레임 단위 사후 확률을 바탕으로 가장 가능성 높은 문자 시퀀스를 출력해야 한다. 기준선 탐욕적 디코더는 프레임 단위 argmax 예측을 단순 축소한다[16][17]. 본 논문은 대신 외부 언어 모델(LM) 결합이 가능한 프리픽스 빔 탐색을 사용한다[4][22][23]. 구체적으로, 각 후보 접두사 s에 대한 총 점수는 식 (5)와 같이 정의된다.

$$S(s)=\log P_{CTC}(s|X)+\alpha log P_{LM}(s)+\beta|s| \tag{5}$$

여기서 $P_{CTC}(s|X)$는 입력 X가 주어졌을 때 CTC 모델 하에서 접두사 s의 조건부 확률이며, $P_{LM}(s)$는 외부 언어 모델이 s에 부여한 확률이다. α는 LM 가중치, β는 문자당 삽입 보너스, $|s|$는 접두사 길이이다. 빔 폭 K는 각 단계에서 유지되는 접두사의 수를 제어하며, 프레임 단위 상위 k 기호만을 유지하여 지연을 제한한다.

그림 1은 세 가지 시스템에 대해 최소-최대 정규화를 거친 후 선택된 디코더 제어 변수 k, α, β, κ를 시각화한 것이다. 탐욕적 디코더는 K=0을 사용하며 외부 언어 지원이 없다. 두 빔 디코더는 모두 K=16과 동일한 shortlist κ를 사용한다. 오직 5-그램 조건에서만 작은 α와 작은 삽입 보너스 β를 통해 언어 융합이 활성화되며, 이는 소폭의 정확도 향상과 연산 시간 증가를 설명한다.

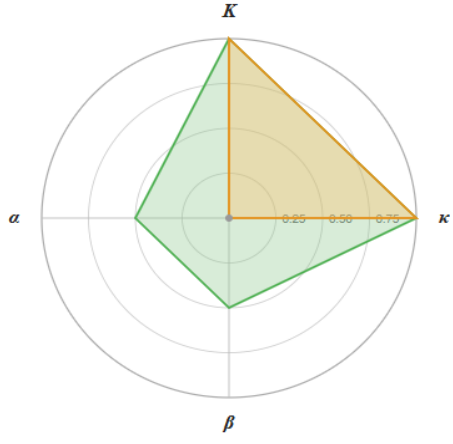

그림 1. 정규화된 디코더 하이퍼파라미터
Fig. 1. Decoder Hyperparameters (normalized)

2.4 평가지표

본 연구에서는 문자 단위에서 인식 품질을 평가한다. 주요 오류 지표는 식 (6)의 문자 오류율(CER, Character Error Rate)이다.

$$CER = \frac{S + D + I}{N} \tag{6}$$

여기서 S, D, I는 예측 시퀀스를 정답 시퀀스로 변환하는 데 필요한 치환, 삭제, 삽입 횟수이며, N은 참조 문자 수이다. 정확도는 1-CER로 보고된다. 추가적으로 다음을 고려한다: Kaggle 점수는 공식 대회 지표 $(N - D)/N$으로, 정규화된 Levenshtein 유사도와 동일하다. 실시간 계수(RTF)는 디코딩 벽시계 시간 비율로, RTF < 1은 실시간보다 빠른 스트리밍을 의미한다[24]. 길이 편향은 예측 시퀀스와 참조 시퀀스 길이 차이 $\Delta_{len} = |\hat{Y}| - |Y|$로, 빠르거나 비정형 동작에서 과도 삽입 혹은 부족 생성을 진단한다. 이러한 지표는 정확도, 지연, 보정 트레이드오프를 정량화하며, 실제 스트리밍 제약 조건 하에서 디코딩 하이퍼파라미터(K, α, β, κ)의 재현 가능한 선택을 가능하게 한다.

III. 연구 방법

제안된 시스템은 경량 인과적 인코더와 디코딩 중심 프레임워크를 결합하여, 수화자에 강건하고 스트리밍에 적합한 ASL 핑거스펠링 인식을 구현한다. 그림 2는 전체 파이프라인을 보여준다. 다중 모달 랜드마크는 표준화된 후 인과적 합성곱-트랜스포머 인코더에 입력되며, CTC 목적 함수 하에서 프레임 단위 사후 확률이 산출된다. 이후 디코딩은 프리픽스 빔 탐색을 통해 수행되며, 외부 언어 모델 결합, 삽입 보너스 보정, 프레임 단위 프루닝을 적용하여 지연이 제한된 상태에서 안정적으로 동작한다.

3.1 인과적 디코더

인코더는 파라미터 θ로 정의되며, (i) 효율적 채널 어텐션(ECA)을 포함한 깊이별 분리 인과 합성곱을 적층하여 손과 입술의 국소 동역학을 포착하고, (ii) 다중 헤드 자기 어텐션 트랜스포머 블록을 통해 장기적인 시간 의존성을 모델링한다. 프레임 순서를 보존하기 위해 위치 인코딩을 추가한다. 양방향 인코더와 달리 이 아키텍처는 엄격히 인과적이므로 미래 문맥 없이 스트리밍 추론이 가능하다. 인코더 출력은 식 (7)과 같이 표현된다.

$$Z = f_\theta(X) \in R^{T \times |V|} \tag{7}$$

여기서 X ∈ $R^{T\times F}$는 입력 시퀀스이며 Z는 프레임 단위 로짓이다. 인코더는 (4)에서 정의된 CTC 손실로 학습된다.

3.2 통합 프리픽스 빔서치와 LM 결합

디코딩 과정에서는 단일 인터페이스 내에서 다양한 언어 모델 계열을 통합하는 프리픽스 빔서치를 사용한다. 각 후보 접두사 s에 대한 점수는 식 (8)과 같이 계산된다:

$$S(s)=\log P_{CTC}(s|X)+\alpha log P_{LM}(s)+\beta|s| \tag{8}$$

여기서 P_{CTC}는 인코더 하에서의 접두사 확률, P_{LM}은 문자 n-그램 또는 사전학습된 GPT 모델로부터의 점수이다. α는 LM 가중치, β는 문자당 삽입 보너스, $|s|$는 접두사의 길이를 의미한다.

빔 폭 K는 단계별로 유지되는 접두사의 개수를 제어하며, 프레임 단위 후보 크기 κ는 상위 κ 기호 확장으로 제한하여 지연을 제어한다. 이러한 설계는 일관된 지연을 보장하고 언어 사전을 유연하게 통합할 수 있게 한다.

빔 폭 K는 단계별로 유지되는 접두사의 개수를 제어하며, 프레임 단위 후보 크기 κ는 상위-κ 기호 확장으로 제한하여 지연을 제어한다. 이러한 설계는 일관된 지연을 보장하고 언어 사전을 유연하게 통합할 수 있게 한다.

3.3 보정 및 하이퍼파라미터 제어

이 프레임워크의 핵심 속성은 명시적 보정 제어이다. (α,β)스윕을 통해 길이 편향을 정확도 및 실시간 계수(RTF)와 함께 모니터링할 수 있다. 이를 통해 원칙적인 운용 지점 선택이 가능하다: 더 좁은 빔은 정확도를 높이지만 RTF를 증가시키며, β를 크게 하면 삽입을 억제하는 대신 삭제가 늘어날 수 있다. 기존 기준선이 단순히 집계 정확도만 보고하는 것과 달리, 본 프로토콜은 식 (9)와 같은 제약 조건 하에서 재현 가능한 트레이드오프 곡선과 소거 연구(Ablation)를 제공한다.

$$\triangle_{len}=|\hat{Y}|-|Y| \tag{9}$$

3.4 복잡도 분석

프레임당 탐욕적 디코딩은 $O(|V|)$연산이 필요하다. 프리픽스 빔서치는 프레임당 O(K·κ)로 증가하는데, 이는 상위-κ 후보만 고려하기 때문이다. n-그램 LM을 통합할 경우, 심볼 확장당 조회와 점수 계산은 $O(n)$이다. GPT 기반 융합에서는 문맥 길이에 따라 상수 시간 오버헤드가 추가된다. 실험적으로 $k \ll |V|$조건에서 프루닝은 정확도를 유지하면서 실행 시간을 제한하여, 범용 GPU에서 RTF < 1의 실시간 성능을 달성한다.

Algorithm 1 CTC Prefix Beam Search with Unified LM Fusion

Input: Logits $Z \in \mathbb{R}^{T\times|\mathcal{V}|}$, beam width K, LM weight α, insertion bonus β, shortlist κ.
Initialize beam $\mathcal{B} \leftarrow \{(\epsilon, 0)\}$ with empty prefix and log-score.
for each frame $t = 1 \ldots T$ **do**
  Select top-κ symbols by logit score.
  for each (s, ℓ) in $\mathcal{B}$ **do**
    Extend with blank symbol (CTC stay).
    for each symbol c in shortlist **do**
      $s' \leftarrow s + c$.
      $\ell' \leftarrow \ell + \log p_\theta(c \mid x_{1:T}) + \alpha \log P_{\text{LM}}(s') + \beta$.
      Add (s', ℓ') to new beam.
  Retain top-K prefixes by total log-score.
Output: Best prefix $\hat{Y}$ by final log-score.

그림 2. 통합 LM 결합을 포함한 CTC 프리픽스 빔서치
Fig. 2. CTC prefix beam search with unified LM fusion

IV. 실 험

모든 실험은 Python 3.9와 TensorFlow 2.10 환경에서 수행되었으며, 8GB 메모리를 탑재한 단일 NVIDIA RTX 2080 GPU와 8코어 CPU를 사용하였다. 메모리 제약을 고려하면서도 효과적인 대규모 배치 학습을 유지하기 위해 정확한 그래디언트 누적을 적용하였다. 학습률은 짧은 워밍업 이후 코사인 감쇠 스케줄을 따르며, 가중치 감쇠는 경량 콜백을 통해 순간 학습률에 연동된다(일반적으로 0.05 ×

lr). 이하의 모든 결과는 공개된 CSV 로그에서 파싱되었으며, 테스트 지표 보고 시 선택된 튜플 (K, α, β, κ)는 고정된다.

4.1 데이터셋

본 연구에서는 랜드마크 기반 ASL 핑거스펠링 코퍼스를 사용한다. 각 프레임은 오른손/왼손(각 포인트), 입술/얼굴, 상반신 자세를 결합하여 F=276개의 수치 특징(XYZ 좌표)을 생성한다. 전처리 파이프라인에 따라 각 랜드마크 그룹은 사전 계산된 평균과 표준편차로 표준화되며, 결측 항목은 표준화 이후 0으로 대체된다. 시퀀스는 고정된 길이 T=176으로 패딩되거나 시간적으로 리사이즈된다. 학습은 공식 학습 분할을 사용하며, 홀드아웃 검증셋은 수화자 기반 샘플링으로 구성되고, 테스트셋은 모델 개발과 하이퍼파라미터 선택 과정에서 사용되지 않는다.

4.2 주요 테스트 비교

표 1은 테스트셋에서 세 가지 디코더를 비교한다. 탐욕적 디코딩은 95% CI [0.7039, 0.7634]와 함께 0.7312 정확도를 달성하며 RTF는 0.002이다. LM 없이 빔 탐색은 0.7223 (RTF 0.044), 5-그램 문자 LM을 추가하면 0.7277 (RTF 0.076)을 기록한다. 해당 체크포인트에서는 탐욕적 기준선이 정확도 면에서도 경쟁력이 있으며 속도는 압도적으로 빠르다.

4.3 선택 프로토콜

테스트셋에서 빔 폭 K $\in$ {8, 16}, LM 가중치 α $\in$ {0.2, 0.4}, 삽입 보너스 β $\in$ {0.0, 0.2}, 후보 크기 κ $\in$ {10, 20}의 작은 그리드를 탐색하였다. 설정은 Score를 기준으로 정렬되며, 동점일 경우 RTF가 더 낮은 설정이 우선된다. 상위 10개의 결과는 표 2에 나열되어 있다.

표 2. 테스트셋에서 선택된 상위 설정
Table 2. Top settings selected in the test set

K	α	β	κ	Score	RTF
16	0.20	0.20	20	0.7285	0.215
16	0.20	0.20	10	0.7282	0.087
8	0.20	0.20	10	0.7271	0.044
8	0.20	0.20	20	0.7268	0.069
16	0.40	0.20	10	0.7271	0.034
16	0.40	0.20	20	0.7271	0.087
16	0.20	0.00	10	0.7254	0.088
16	0.20	0.00	20	0.7250	0.215
8	0.20	0.00	10	0.7233	0.022
8	0.20	0.00	20	0.7233	0.087

4.4 삽입 보너스 β의 정확도 효과

(K=16, α=0.2, κ=20)을 고정한 상태에서 그림 3은 삽입 보너스 β가 0에서 0.5로 증가함에 따라 정확도가 단조롭게 향상됨을 보여준다(약 0.728 → 0.738). 가장 큰 향상은 β $\in$ [0.2, 0.3] 구간에서 나타나며, 이후에는 수익이 감소하고 β = 0.4-0.5에서는 거의 포화된다(변화 < 5×10^{-4}). β는 디코딩 점수만 재가중하므로 연산 비용에는 영향이 없으며, $\beta \approx 0.3-0.4$는 추가 지연 없이 정확도를 극대화하는 합리적 지점이다. 작은 β는 삭제를 선호하는 경향이 있으며, 지나치게 큰 β는 과도 삽입 위험이 있으므로 길이 편향 모니터링이 필요하다.

표 1. 테스트셋에서 디코더 비교(Score, RTF, FPS 및 길이 편향)
Table 1. Compare decoders in test sets (Score, RTF, FPS, and length deflection)

Decoder	K	α	β	κ	CI_0.025	CI_0.975	Score	RTF	$\lvert\hat{Y}\rvert - \lvert Y\rvert$
Greedy	0	0.00	0.00	20	0.7039	0.7634	0.7344	0.002	12470.5
Beam (no LM)	16	0.00	0.00	20	0.6968	0.7527	0.7250	0.044	682.2
Beam + 5-gram	16	0.20	0.20	20	0.6992	0.7588	0.7303	0.076	396.4

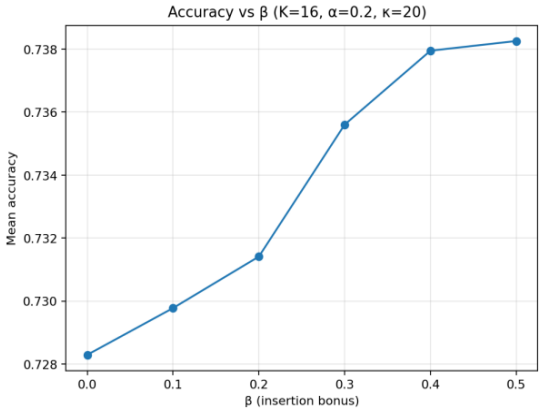

그림 3. 삽입 보너스 β에 따른 테스트 정확도
Fig. 3. Test accuracy according to insertion bonus β

4.5 후보 크기 κ 소거 실험

표 3은 프레임별 후보 크기에 대한 소거 실험이다. κ를 10에서 20으로 늘리면 Score는 소폭 증가(0.7282 → 0.7285)하나, RTF는 0.033에서 0.087로 증가하여 테스트셋에서 명확한 속도-품질 트레이드오프를 보인다.

표 3. 선택된(K, α, β)에서 프레임별 후보 크기 κ 소거 실험
Table 3. Ablation of frame-level candidate size κ under selected (K, α, β)

κ	Score	RTF	FPS	$\lvert\hat{Y}\rvert - \lvert Y\rvert$
10	0.7282	0.033	918.9	-4.03
20	0.7285	0.087	343.1	-4.03

4.6 β를 통한 삽입 제어

선택된 (K, α, κ)에서 삽입 보너스 $\beta \in \{0.0, 0.2, 0.4\}$를 테스트셋에서 탐색하였다(표 4). Score는 단조롭게 향상되며(0.7250 → 0.7285 → 0.7362), 평균 길이 편향 $\lvert\hat{Y}\rvert - \lvert Y\rvert$는 음수로 덜 변하여 삭제가 감소함을 반영한다.

표 4. 삽입 보너스 β 변화에 따른 성능 및 길이 편향 분석
Table 4. Performance and length deflection analysis based on insertion bonus β changes

K	α	β	κ	FPS	$\lvert\hat{Y}\rvert - \lvert Y\rvert$
16	0.20	0.0	20	0.7250	-4.17
16	0.20	0.2	20	0.7285	-4.03
16	0.20	0.4	20	0.7362	-3.79

V. 결 론

본 연구는 경량 인과적 인코더와 통합 CTC 프리픽스 빔 탐색을 결합한 디코딩 중심 프레임워크를 제안하였다. 단일 인터페이스 내에서 본 방법은 언어적 지원과 보정을 위한 세 가지 해석이 가능한 제어 변수—빔 폭 K, LM 가중치 α, 삽입 보너스 β—와 지연 안정화를 위한 네 번째 조정 변수 κ를 노출한다. 평균 정확도를 넘어, 본 프레임워크는 재현 가능한 스윕을 통한 운용 지점 선택을 강조하며, Score, 실시간 계수(RTF), 처리 속도(FPS, Frames Per Second), 길이 편향 $\lvert\hat{Y}\rvert - \lvert Y\rvert$을 보고하여 정확도-지연 트레이드오프를 명시적이고 비교 가능하게 만든다.

홀드아웃 테스트셋에서의 실험 결과, 본 파이프라인은 경쟁력 있는 정확도를 달성하면서도 실행 시간을 제어할 수 있으며, 구조화된 소거 분석은 β가 삭제 오류를 완화하고 κ가 작은 품질 향상과 상당한 연산량 증가 사이의 균형을 형성함을 보여준다. 또한 이벤트 트리거 기반 디코딩의 임계값 $\delta = 0.1$이 정확도-효율성 균형의 최적점임을 실험적으로 확인하였다.

향후 연구 방향으로는 (i) 온디바이스 사용을 위해 증류된 문자 단위 신경 언어 모델, (ii) 스트리밍 지연 하에서 라벨을 보존하는 수화자 인식 기반 적응, (iii) β 선택과 명시적 삭제/삽입 비용 간의 긴밀한 결합, (iv) 상호작용 애플리케이션을 위한 이벤트 기반 조기 커밋 전략 등이 있다. 이러한 확장은 드문 문자열에서의 강건성 격차를 더욱 좁히는 동시에 제안된 선택 프로토콜의 투명성을 유지할 것으로 기대된다.

References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition", Proc. Interspeech, Shanghai, China, pp. 5036-5040, Oct. 2020.

[2] T. Hori, S. Watanabe, Y. Zhang, and W. Chan,

"Advances in Joint CTC-Attention Based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM", Proc. Interspeech, Stockholm, Sweden, pp. 949-953, Aug. 2017.

[3] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss", Proc. ICASSP, Barcelona, Spain, pp. 7829-7833, May 2020. https://doi.org/10.1109/ICASSP40776.2020.9053896.

[4] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the Comparison of Popular End-to-End Models for Large Scale Speech Recognition", Proc. Interspeech 2020, Shanghai, China, pp. 1-5, Oct. 2020. https://doi.org/10.21437/Interspeech.2020-2846.

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training", OpenAI Tech. Rep., Jun. 2018.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners", OpenAI Tech. Rep., Feb. 2019.

[7] T. B. Brown, B. Mann, N. Ryder et al., "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Vancouver, BC, Canada, Vol. 33, pp. 1877-1901, Dec. 2020.

[8] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments", Proc. ICML, Sydney, NSW, Australia, Vol. 70, pp. 2837-2846, Aug. 2017.

[9] C.-C. Chiu and C. Raffel, "Monotonic Chunkwise Attention", Proc. 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, pp. 1-16, Apr. 2018.

[10] S. Zhang, Z. Gao, H. Luo, M. Lei, J. Gao, Z. Yan, and L. Xie, "Streaming Chunk-Aware Multihead Attention for Online End-to-End Speech Recognition", Proc. Interspeech, Shanghai, China, pp. 2142-2146, Oct. 2020.

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks", Journal of Machine Learning Research, Vol. 17, No. 59, pp. 1-35, Apr. 2016.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations", Proc. ICML, Online, pp. 1597-1607, Jul. 2020.

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition", Proc. Interspeech, Graz, Austria, pp. 2613-2617, Sep. 2019.

[14] B. Shi, A. Martinez Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "American Sign Language fingerspelling recognition in the wild", Proc. IEEE SLT Workshop, Athens, Greece, pp. 145-152, Dec. 2018.

[15] B. Shi, A. Martinez Del Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention", Proc. ICCV, Seoul, Korea, pp. 5399-5408, Oct. 2019.

[16] Kaggle and Google Research, "Google - American Sign Language Fingerspelling Recognition", https://www.kaggle.com/competitions/asl-fingerspelling. [accessed: Sep. 2024]

[17] Google Research, "asl-fingerspelling", GitHub repository, https://github.com/google-research/datasets/tree/master/asl-fingerspelling. [accessed: Sep. 2024].

[18] Kaggle, "American Sign Language Fingerspelling Recognition Competition", https://www.kaggle.com/competitions/asl-fingerspelling. [accessed: Aug. 2025]

[19] C. Lugaresi, et al., "MediaPipe: A Framework for Perceiving and Processing Reality", Proc. 3rd Workshop on Computer Vision for AR/VR at IEEE/CVF CVPR, Long Beach, California, USA, Jun. 2019.

[20] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking", arXiv:2006.10214, pp. 1-5, Jun. 2020. https://doi.org/10.48550/arXiv.2006.10214.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", Proc. ICML, Pittsburgh, Pennsylvania, USA, pp. 369-376, Jun. 2006. https://doi.org/10.1145/1143844.1143891.

[22] C. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On Using Monolingual Corpora in Neural Machine Translation", ararXiv preprint ararXiv:1503.03535, Mar. 2015. https://doi.org/10.48550/arXiv.1503.03535.

[23] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An Analysis of Incorporating an External Language Model into a Sequence-to-Sequence Model", Proc. ICASSP, Calgary, AB, Canada, pp. 5824-5828, Apr. 2018.

[24] Y. He, et al., "Streaming End-to-End Speech Recognition for Mobile Devices", Proc. ICASSP, Brighton, UK, pp. 6381-6385, May 2019.

저자소개

유 승 수 (Seung-Su Yu)

2024년 2월 : 국립한밭대학교 창업학(석사)
2026년 2월 : 대전대학교 융합컨설팅학(기술경영 박사수료)
2011년 11월 ~ 현재 : ㈜멀틱스 대표이사
관심분야 : AI, Barrier Free, 수어(Sign Language), Multi Modal Communication Technology

Khoa Nguyen

2023년 2월 : 베트남-독일대학교 컴퓨터공학(학사)
2024년 3월 ~ 현재 : 충북대학교 정보통신공학부 석박사통합과정
관심분야 : Research Interests, Federated Learning, Time Series Forecasting

Thong-Nhat Tran

2023년 2월 : 홍익대학교 전기전자공학(박사)
2023년 3월 ~ 현재 : 충북대학교 정보통신공학부 연구원 및 대학강사
관심분야 : Research Interests, Wireless Communications, Optimization, Computer Vision

서 영 욱 (Young-Wook Seo)

2000년 2월 : 성균관대학교 경영학(석사)
2008년 2월 : 성균관대학교 경영학(박사)
2014년 대전대학교 일반대학원 융합컨설팅학과 교수
관심분야 : 정보경영, IT컨설팅, 경영컨설팅, 창의성, 컨설턴트