# CGAM : CTC Knowledge Distillation 기반 Cross-Attention Masking을 통한 Whisper Hallucination 저감

박우성, 문찬영, 김세민, 한진모, 이윤형, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실
wsp156@snu.ac.kr, {cymoon, smkim21, jmhan, yhlee}@hi.snu.ac.kr, nkim@snu.ac.kr

# CGAM: Reducing Whisper Hallucination via Cross-Attention Masking with CTC Knowledge Distillation

Wooseong Park, Chanyeong Moon, Semin Kim, Jinmo Han, Yoonhyeong Lee and Nam Soo Kim
Human Interface Laboratory
Department of Electrical and Computer Engineering and INMC,
Seoul National Univ.

## Abstract

OpenAI Whisper is a powerful speech recognition model, but it suffers from hallucination, generating meaningless text for non-speech inputs such as environmental sounds and music. We propose training a lightweight CTC adapter (4.2MB) on top of the Whisper Encoder through Knowledge Distillation from Wav2Vec2ForCTC, then applying the adapter output as a frame-level mask to the Decoder cross-attention. Experiments show that the proposed method "reduces hallucination rate on UrbanSound8K from 100.0% to 1.14% while preserving identical WER (4.07%) on LibriSpeech test-clean.

## I. 서론

OpenAI의 Whisper[1]는 680,000시간의 대규모 웹 데이터로 학습된 Encoder-Decoder 구조의 음성 인식 모델로, 다국어 전사 및 번역에서 뛰어난 성능을 보인다. 그러나 Whisper는 환경음, 음악, 노이즈 등 비음성(non-speech) 입력에 대해서도 텍스트를 생성하는 환각(hallucination) 문제가 보고되고 있다[3].

이러한 환각 문제를 해결하기 위한 다양한 접근이 제안되었다. 가장 일반적인 방법은 Silero VAD[4] 등 외부 음성 활동 감지(VAD) 모델을 전처리로 적용하여 비음성 구간을 필터링하는 것이나, Whisper와 독립적인 별도 모델이 필요하고 feature 공간의 불일치가 발생한다. Whisper 내부에서 환각을 억제하려는 최근 시도로 CalmWhisper[5], Whisper-CD[6] 등이 제안되었지만, 음성 인식률이 낮아지는 trade-off를 갖게 되거나 VAD를 적용한 것 만큼의 효과를 볼 순 없었다. 한편, CTC 기반 모델(Wav2Vec2[7], Conformer-CTC 등)은 비음성 입력에서 대부분 blank/PAD 토큰을 출력하여 환각에 본질적으로 강건한 특성을 보인다.

본 연구에서는 CTC 모델의 이러한 강건성을 Whisper에 접목하는 방법을 제안한다. Whisper Encoder 위에 경량 adapter를 학습시키고, 이 adapter의 출력을 frame-level speech mask로 변환하여 Decoder의 cross-attention에 적용한다. 이를 통해 별도의 외부 모델 없이 Whisper 내부에서 통합적으로 환각을 억제한다.

## II. 본론

### 2.1 전체 구조

제안 시스템은 세 단계로 구성된다. 첫째, Whisper Encoder의 hidden states 위에 Conv1d 기반 Adapter 모델(이하 CGAM)를 배치하고 CTC teacher로부터 KD를 통해 학습한다. 둘째, 학습된 CGAM의 CTC 출력에서 PAD token 확률을 기반으로 frame-level speech mask를 생성한다. 셋째, 생성된 mask를 Whisper Decoder의 cross-attention에 적용하여 비음성 프레임에 대한 attention을 차단한다.

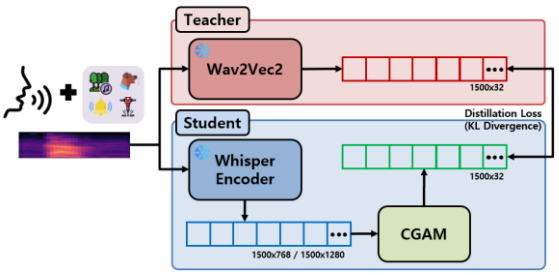


그림 1. CGAM 학습 구조

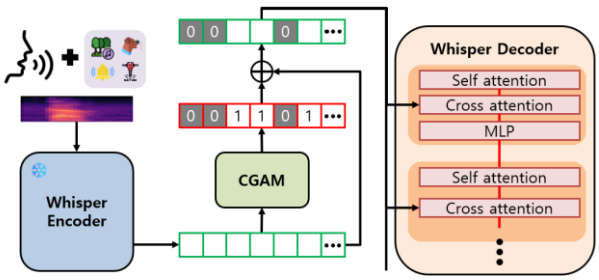


그림 2. CGAM 추론 구조

## 2.2 CTC-KD Adapter

Student는 Whisper Encoder의 hidden states(768차원, frozen)이며, Teacher는 Wav2Vec2ForCTC(facebook/wav2vec2-large-960h, frozen)이다. CGAM은 Conv1d(768→256, kernel=5) → GELU → GroupNorm → Conv1d(256→128, kernel=3) → GELU → GroupNorm → Conv1d(128→32, kernel=1)의 3층 구조로, 전체 학습 파라미터는 약 4.2MB이다. Whisper는 오디오 길이에 관계없이 1500 프레임을 출력하는 반면, Wav2Vec2는 실제 오디오 길이에 비례한 프레임을 출력한다. 따라서 Whisper hidden states를 teacher의 실제 프레임 수(T_teacher)로 trim하여 정렬한 후, CGAM의 출력과 teacher CTC logits 간 KL Divergence를 최소화한다.

## 2.3 학습 데이터

CGAM 학습에는 음성과 비음성 데이터를 혼합하여 총 3,001개의 오디오 파일을 사용하였다. 음성 데이터로는 LibriSpeech의 dev-clean(2,703개)을 사용하여 영어 발화 샘플을 포함하였다. 비음성 데이터로는 MUSAN[9] 데이터셋에서 free-sound(103개), 클래식 음악(77개), sound-bible(86개)을 사용하였으며, DEMAND[10] 데이터셋에서 주방 소음(16개)과 세탁기 소음(16개)을 추가하여 총 298개의 비음성 샘플을 포함하였다. 음성과 비음성 데이터를 함께 학습함으로써, CGAM이 음성 프레임에서는 teacher의 문자 토큰 분포를, 비음성 프레임에서는 PAD 토큰 분포를 모방하도록 학습된다.

## 2.4 학습 설정

학습 하이퍼파라미터는 다음과 같다: 에폭 5, 배치 크기 4, 학습률 1e-4(AdamW), 최대 오디오 길이 10초. 10초를 초과하는 오디오는 랜덤 크롭을 적용하였다. Loss는 KL Divergence에 프레임별 가중치를 적용하여, teacher의 argmax가 PAD(0)이 아닌 음성 프레임에 10배의 가중치를 부여하였다. 이는 CTC 출력의 대부분이 PAD인 특성상, 음성 프레임 학습이 소수 클래스로 무시되는 것을 방지하기 위함이다.

## 2.5 Cross-Attention Masking

학습된 CGAM은 각 프레임의 CTC 토큰 분포를 출력한다. CTC의 PAD token(인덱스 0) 확률이 threshold 이상인 프레임은 비음성으로 판단하여 mask=0, 나머지는 mask=1로 설정한다. 생성된 mask를 Whisper의 encoder attention mask로 전달하여, 비음성 프레임이 Decoder cross-attention에서 참조되지 않도록 차단한다. 입력이 전적으로 비음성인 경우 모든 프레임이 masking되어 Decoder는 자연스럽게 EOT(End of Transcript) 토큰만 생성한다.

# Ⅲ. 실험

## 3.1 실험 데이터셋 및 설정

평가에는 학습에 사용되지 않은 데이터셋을 사용하였고, 평가는 총 2가지로 음성 평가와 비음성 평가를 진행하였다. 비음성 평가에는 UrbanSound8K[11]의 4개 환경음 클래스(drilling, car_horn, engine_idling, jackhammer), 총 3,429개 클립을 사용하였다. UrbanSound8K의 나머지 클래스(children_playing, street_music 등)는 사람

목소리가 포함된 경우가 있어, 해당 발화가 정상적으로 인식되더라도 hallucination으로 오분류될 수 있어 제외하였다. 따라서 순수 비음성 입력에 대한 hallucination 억제 성능을 엄밀히 측정하기 위해, 음성 성분이 없는 기계음 및 충격음 클래스만을 평가에 사용하였다. 음성 평가에는 LibriSpeech test-clean(2,620개 발화)을 사용하였고, Baseline으로 Vanilla Whisper와 Silero VAD[4]+Whisper를 비교하였다.

### 3.2 Whisper Hallucination 제거 성능

표 1에서 보듯이, Vanilla Whisper는 비음성 샘플 전체에서 'so', 'oh', 'um' 등의 환각을 생성한다. Silero VAD를 통해 필터링을 하게 되면, 0.73%로 대부분의 환각을 제거하게 된다. 제안 방법도 Silero VAD와 유사 수준인 1.14%로 대부분의 환각을 제거하는 것을 볼 수 있다.

| 방법 | 환각 수 | 환각률 |
|---|---|---|
| Vanilla Whisper | 3,429 | 100.0% |
| Silero VAD | 25 | 0.73% |
| Proposed | 39 | 1.14% |

표 1. UrbanSound8K 환각률 비교

### 3.3 음성 인식 성능 유지

표 2에서 보듯이, LibriSpeech test-clean 2,620개 발화에 대해 제안 방법은 Vanilla Whisper와 동일한 WER(4.07%)을 달성하였다. Silero VAD도 동일하게 4.07%로 일치를 보여, 제안 방법이 음성 성능을 해치지 않음을 확인하였다.

| 방법 | WER |
|---|---|
| Vanilla Whisper | 4.07% |
| Silero VAD | 4.07% |
| Proposed | 4.07% |

표 2. LibriSpeech WER 평가 결과

## Ⅳ. 결론

본 연구에서는 Whisper의 비음성 환각 문제를 해결하기 위해, CTC 모델(Wav2Vec2)로부터 Knowledge Distillation을 통해 Whisper Encoder 위에 4.2MB 경량 adapter를 학습하고, cross-attention masking을 적용하는 방법을 제안하였다. UrbanSound8K에서 비음성 환각률을 100%에서 1.14%로 대폭 감소시키면서, LibriSpeech에서 WER 4.07%의 동일 출력을 유지하여 음성 인식 성능의 저하 없이 환각을 억제할 수 있음을 확인하였다. 제안 방법은 Whisper의 기존 Encoder feature를 재활용하여 최소한의 추가 비용으로 동작하며, 외부 VAD 모델 없이 Whisper 내부에서 통합적으로 환각을 억제한다는 장점이 있다. 향후 낮은 SNR 환경에서의 강건성 향상과 음향 기반 환각(sound-grounded hallucination)에 대한 추가 연구가 필요하다.

## 참고문헌


[1] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *International conference on machine learning*. PMLR, 2023.

[2] Barański, Mateusz, et al. "Investigation of whisper asr hallucinations induced by non-speech audio." *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025.

[3] Silero, V. A. D. "pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier." *URL: https://github.com/snakers4/silero-vad (accessed: April 15, 2025)*.

[4] Yingzhi Wang, Anas Alhmoud, Saad Alsahly, Muhammad Alqurishi, and Mirco Ravanelli. CalmWhisper: Reduce Whisper Hallucination On Non-Speech By Calming Crazy Heads Down. In Interspeech 2025.

[5] Ahn, Hoseong, et al. "Whisper-CD: Accurate Long-Form Speech Recognition using Multi-Negative Contrastive Decoding." *arXiv preprint arXiv:2603.06193* (2026).

[6] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

[7] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." NeurIPS Workshop, 2015.

[8] Snyder, David, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus." *arXiv preprint arXiv:1510.08484* (2015).

[9] Thiemann, Joachim, Nobutaka Ito, and Emmanuel Vincent. "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments." *(No Title)* (2013).

[10] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." *Proceedings of the 22nd ACM international conference on Multimedia*. 2014.