[**Robust Speech Recognition via Large-Scale Weak Supervision (Whisper)**](https://kimjy99.github.io/%EB%85%BC%EB%AC%B8%EB%A6%AC%EB%B7%B0/whisper/)

arXiv 2022. [[Paper](https://arxiv.org/abs/2212.04356)] [[Page](https://openai.com/research/whisper)] [[Github](https://github.com/openai/whisper)]  
Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever Frank, Jesse Engel, Quoc V. Le, William Chan, Wei Han  
OpenAI  
6 Dec 2022

**Introduction**

음성 인식의 발전은 Wav2Vec 2.0와 같은 unsupervised(비지도) 사전 학습 기술의 개발에 의해 활성화되었다. 이러한 방법은 레이블이 지정되지 않은 오디오에서 직접 학습하기 때문에 레이블이 지정되지 않은 음성의 대규모 데이터셋을 생산적으로 사용할 수 있으며 supervisd 데이터셋의 1,000시간 정도보다 훨씬 많은 100만 시간의 학습 데이터로 빠르게 확장되었다. 표준 벤치마크에서 fine-tuning될 때 이 접근 방식은 특히 low-data 세팅에서 state-of-the-art를 개선했다.

이러한 사전 학습된 오디오 인코더는 음성의 고품질 표현을 학습하지만 순수하게 unsupervised이므로 이러한 표현을 사용 가능한 출력에 매핑하는 동등한 성능의 디코더가 부족하여 음성 인식과 같은 작업을 실제로 수행하기 위해 fine-tuning 단계가 필요하다. 이는 불행하게도 fine-tuning이 여전히 숙련된 실무자가 필요한 복잡한 프로세스일 수 있으므로 유용성과 영향을 제한한다. Fine-tuning이 필요한 것은 추가적인 위험이 있다. 머신 러닝 방법은 학습 데이터셋 내에서 동일한 데이터셋의 hold-out 데이터에 대한 성능을 향상시키는 패턴을 찾는 데 매우 능숙하다. 그러나 이러한 패턴 중 일부는 깨지기 쉽고 가짜이며 다른 데이터셋과 분포로 일반화되지 않는다. [Learning transferable visual models from natural language supervision 논문](https://arxiv.org/abs/2103.00020)은 ImageNet에서 컴퓨터 비전 모델을 fine-tuning할 때 classification 정확도가 9.2% 증가했다고 보고했다. 한 데이터셋에 대해 학습을 받았을 때 인간을 뛰어넘는 성능을 달성한 모델은 다른 데이터셋에서 평가할 때 여전히 많은 기본적인 오차를 만들 수 있다. 아마도 정확하게는 인간이 인식하지 못하는 데이터셋별 특이점을 이용하기 때문일 수 있다.

이것은 unsupervised 사전 학습이 오디오 인코더의 품질을 극적으로 향상시켰지만, 데이터셋별 fine-tuning 시 동등한 고품질 사전 학습된 디코더의 부족이 유용성과 견고성을 제한하는 결정적인 약점임을 시사한다. 음성 인식 시스템의 목표는 적용하고자 하는 모든 분포에 대한 디코더의 supervised fine-tuning 없이 광범위한 환경에서 즉시 안정적으로 작동하는 것이다.

여러 데이터셋 / 도메인에 걸쳐 supervisd 방식으로 사전 학습된 음성 인식 시스템은 단일 소스에서 학습된 모델보다 더 높은 견고성을 나타내고 hold-out 데이터셋에 훨씬 더 효과적으로 일반화된다. 이러한 작업은 가능한 한 많은 기존 고품질 음성 인식 데이터셋을 결합하여 이를 달성한다. 그러나 쉽게 사용할 수 있는 이 데이터의 양은 여전히 적다. [SpeechStew](https://arxiv.org/abs/2104.02133)는 총 5,140시간의 supervision에 해당하는 7개의 기존 데이터셋을 혼합하였다. 사소하지는 않지만 이전에 언급한 레이블이 지정되지 않은 음성 데이터 100만 시간에 비하면 여전히 작다.

기존 고품질 supervisd 데이터셋의 제한적인 크기를 인식하여 최근에는 음성 인식을 위한 더 큰 데이터셋을 만들었다. [GigaSpeech](https://arxiv.org/abs/2106.06909)와 [The People’s Speech](https://arxiv.org/abs/2111.09344)는 정교한 자동화 파이프라인을 사용하여 weakly-supervised 음성 인식을 1만 시간과 3만 시간의 noisy한 학습 데이터로 확장한다. 품질과 수량 사이의 이러한 trade-off는 종종 올바른 선택이다. 지금까지 음성 인식에 대해 충분히 연구되지 않았지만, 컴퓨터 비전의 최근 연구들은 ImageNet과 같은 크라우드소싱 데이터셋을 넘어 훨씬 더 크지만 weakly-supervisd 데이터셋으로 이동하면 모델의 견고성과 일반화가 크게 향상됨을 보여주었다.

그러나 이러한 새로운 데이터셋은 기존의 고품질 데이터셋의 합계보다 몇 배 더 클 뿐이며 이전의 unsupervised 작업보다 여전히 훨씬 작다. 본 논문에서는 weakly-supervised 음성 인식을 680,000시간의 레이블이 지정된 오디오 데이터로 확장하여 그 격차를 좁힌다. 본 논문의 접근 방식을 **Whisper**라고 부른다. 이 규모에서 학습된 모델이 기존 데이터셋 zero-shot으로 잘 전송되어 고품질 결과를 달성하기 위해 데이터셋별 fine-tuning이 필요하지 않음을 보여준다.

규모 외에도 weakly-supervised 사전 학습의 범위를 영어 전용 음성 인식을 넘어 다국어 및 멀티태스킹으로 확장하는 데 중점을 둔다. 680,000시간의 오디오 중 117,000시간은 96개의 다른 언어를 포함한다. 이 데이터셋에는 125,000시간의 X→→en 번역 데이터도 포함되어 있다. 저자들은 충분히 큰 모델의 경우 다국어 및 멀티태스킹 학습에 단점이 없으며 심지어 이점도 있음을 발견했다.

본 논문은 weakly-supervisd 사전 학습의 간단한 확장이 음성 인식에 대해 지금까지 과소 평가되었음을 시사한다. 최근 대규모 음성인식 작업의 주축이 되어온 self-supervision이나 self-training 기술 없이도 이러한 결과를 얻을 수 있다.

**Approach**

**1. Data Processing**

기계 학습 시스템을 학습하기 위해 인터넷의 웹 스케일 텍스트를 활용하는 최근 연구들의 추세에 따라 데이터 전처리에 대한 최소한의 접근 방식을 취한다. 음성 인식에 대한 많은 연구들과 달리, 본 논문은 발화와 전사(transcription) 사이를 매핑하는 방법을 배우기 위해 sequence-to-sequence 모델의 표현력에 의존하여 중요한 표준화 없이 녹취록의 텍스트를 예측하도록 Whisper 모델을 학습시킨다. 이렇게 하면 자연스러운 전사를 생성하기 위해 별도의 inverse text normalization step이 필요하지 않으므로 음성 인식 파이프라인이 간소화된다.

저자들은는 인터넷의 전사와 쌍을 이루는 오디오에서 데이터셋을 구성하였다. 그 결과 다양한 환경, 녹음 설정, 스피커 및 언어의 광범위한 오디오 분포를 다루는 매우 다양한 데이터셋이 생성되었다. 오디오 품질의 다양성은 모델을 견고하게 학습시키는 데 도움이 될 수 있지만 전사 품질의 다양성은 마찬가지로 유익하지 않다. 초기 데이터셋에서는 수준 이하의 전사가 많이 나타났다. 저자들은 전사 품질을 개선하기 위해 몇 가지 자동화된 필터링 방법을 개발했다.

인터넷에 있는 많은 전사들은 실제로 인간이 생성한 것이 아니라 기존 ASR 시스템의 결과물이다. 최근 연구에 따르면 사람과 기계가 혼합한 데이터셋에 대한 학습은 번역 시스템의 성능을 크게 저하시킬 수 있다. 저자들은 “transcript-ese” 학습을 피하기 위해 학습 데이터셋에서 기계가 생성한 전사를 감지하고 제거하는 많은 휴리스틱을 개발했다. 기존의 많은 ASR 시스템은 오디오 신호만으로는 예측하기 어려운 측면을 제거하거나 정규화하는 제한된 부분 집합만 출력한다 (ex. 복잡한 구두점(느낌표, 쉼표, 물음표), 단락과 같은 서식 공백, 문체). 모두 대문자 또는 모두 소문자로 된 전사는 사람이 생성한 것일 가능성이 매우 낮다. 많은 ASR 시스템에는 일정 수준의 inverse text normalization이 포함되어 있지만 단순하거나 규칙 기반인 경우가 많으며 처리되지 않은 다른 측면에서 여전히 감지할 수 있다.

저자들은 또한 VoxLingua107에 있는 데이터셋의 프로토타입 버전에서 학습된 프로토타입 모델을 fine-tuning하여 생성된 audio language detector를 사용하여 음성 언어가 전사의 언어와 일치하는지 확인하였다. 둘이 일치하지 않으면 데이터셋의 음성 인식 학습 예제로 (오디오, 전사) 쌍을 포함하지 않는다. 전사의 언어가 영어인 경우 예외를 만들고 대신 이러한 쌍을 X→→en 음성 번역 학습 예제로 데이터셋에 추가하였다. 전사 텍스트의 중복 제거를 사용하여 학습 데이터셋에서 중복 및 자동 생성 콘텐츠의 양을 줄였다.

오디오 파일을 30초 세그먼트 내에서 발생하는 전사와 쌍을 이루는 30초 세그먼트로 나눈다. 음성이 없는 세그먼트를 포함하여 모든 오디오에 대해 학습하고 이러한 세그먼트를 음성 활동 감지를 위한 학습 데이터로 사용한다.

저자들은 추가적인 filtering pass를 위해 초기 모델을 학습시킨 후 학습 데이터 소스에 대한 오차율에 대한 정보를 집계하고, 효율적으로 낮은 품질의 데이터를 식별하고 제거하기 위해 높은 오차율과 데이터 소스 크기의 조합으로 정렬하여 수동 검사하였다. 이 검사에서는 필터링 휴리스틱이 감지하지 못한 남아 있는 낮은 품질의 기계 생성 캡션뿐만 아니라 부분적으로만 전사되었거나 잘못 정렬된 많은 양의 전사를 보여주었다.

오염을 피하기 위해 학습 데이터 세트와 중복 위험이 더 높다고 생각한 평가 데이터셋, 즉 TED-LIUM 3 사이의 전사 레벨에서 중복 제거를 수행한.

**2. Model**

본 논문의 초점은 음성 인식을 위한 대규모 supervised 사전 학습의 능력을 연구하는 것이므로 모델 개선과 저자들의 발견을 혼동하지 않도록 기성 아키텍처를 사용한다. 저자들은 인코더-디코더 Transformer를 선택했으며, 안정적으로 확장할 수 있다는 것이 잘 검증되었기 때문이다. 모든 오디오는 16kHz로 다시 샘플링되고 80채널 log Mel spectrogram 표현은 stride가 10ms인 25ms window에서 계산된다. Feature 정규화를 위해 사전 학습 데이터셋에서 평균이 대략 0인 -1과 1 사이가 되도록 입력을 글로벌하게 조정한다. 인코더는 filter 크기가 3인 2개의 convolution layer와 GELU로 구성된 작은 stem으로 이 입력 표현을 처리한다. 여기서 두 번째 convolution layer의 stride는 2다. 그런 다음 sinusoidal position embedding이 stem의 출력에 추가된 후 인코더 Transformer block이 적용된다. Transformer는 pre-activation residual block을 사용하고 최종 layer normalization이 인코더 출력에 적용된다. 디코더는 학습된 position embedding과 입출력 토큰 표현을 사용한. 인코더와 디코더는 Transformer block의 크기와 개수가 동일하다. 아래 그림은 모델 아키텍처를 요약한 것이다.

텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명

GPT2에서 사용된 것과 동일한 바이트 레벨의 BPE text tokenizer를 영어 전용 모델에 사용한다. GPT-2 BPE vocabulary가 영어 전용이므로 다른 언어에 대한 단편화를 피하기 위해 다국어 모델에 대한 vocabulary를 다시 맞춘다 (크기는 동일하게 유지).

**3. Multitask Format**

주어진 오디오 snippet에서 어떤 단어가 말했는지 예측하는 것이 전체 음성 인식 문제의 핵심 부분이며 연구에서 광범위하게 연구되었지만 유일한 부분은 아니다. 모든 feature을 갖춘 음성 인식 시스템에는 음성 활동 감지, 화자 분할, inverse text normalization과 같은 많은 추가 구성 요소가 포함될 수 있다. 이러한 구성 요소는 종종 개별적으로 처리되므로 핵심 음성 인식 모델 주변의 시스템이 상대적으로 복잡해진다. 저자들은 이 복잡성을 줄이기 위해 하나의 모델이 핵심 인식 부분뿐만 아니라 전체 음성 처리 파이프라인을 수행하도록 하려고 했다. 여기서 중요한 고려 사항은 모델의 인터페이스이다. 동일한 입력 오디오 신호에서 수행할 수 있는 여러 가지 task가 있다. 녹음, 번역, 음성 활동 감지, 일치, 언어 식별이 몇 가지 예이다.

이러한 종류의 일대다 매핑이 하나의 모델에서 작동하려면 어떤 형태의 task specification이 필요하다. 간단한 형식을 사용하여 모든 작업과 컨디셔닝 정보를 디코더에 대한 일련의 입력 토큰으로 지정한다. 디코더는 오디오 조건부 언어 모델이기 때문에 모호한 오디오를 해결하기 위해 더 긴 범위의 텍스트 컨텍스트를 사용하는 방법을 배우기를 바라며 전사의 텍스트 기록을 조건으로 학습한다. 특히 현재 오디오 세그먼트 앞에 있는 전사 텍스트를 디코더의 컨텍스트에 추가할 가능성이 있다.

⟨|startoftranscript|⟩⟨|startoftranscript|⟩ 토큰으로 예측의 시작을 나타낸다. 먼저, 학습 셋의 각 언어(총 99개)에 대한 고유한 토큰으로 표현되는 말하는 언어를 예측한다. 이러한 언어 target은 앞서 언급한 VoxLingua107 모델에서 가져온 것이다. 오디오 세그먼트에 음성이 없는 경우 이를 나타내는 ⟨|nospeech|⟩⟨|nospeech|⟩ 토큰을 예측하도록 모델을 학습한다. 다음 토큰은 ⟨|transcribe|⟩⟨|transcribe|⟩ 또는 ⟨|translate|⟩⟨|translate|⟩ 토큰을 사용하여 task(전사 또는 번역)를 지정한다. 그런 다음 각 케이스에 대한 ⟨|notimestamps|⟩⟨|notimestamps|⟩ 토큰을 포함하여 타임스탬프를 예측할지 여부를 지정한다. 이 시점에서 task와 원하는 형식이 완전히 지정되고 출력이 시작된다.

타임스탬프 예측을 위해 현재 오디오 세그먼트와 관련된 시간을 예측하고, Whisper 모델의 기본 시간 해상도와 일치하는 가장 가까운 20ms로 모든 시간을 양자화하고, 이들 각각에 대한 vocabulary에 추가 토큰을 추가한다. 저자들은 예측을 캡션 토큰과 인터리브하였다. 시작 시간 토큰은 각 캡션의 텍스트 전에 예측되고 종료 시간 토큰은 그 후에 예측된다. 최종 전사 세그먼트가 현재 30초 오디오 청크에 부분적으로만 포함된 경우 타임스탬프 모드에 있을 때 세그먼트의 시작 시간 토큰만 예측하여 해당 시간과 일치하는 오디오 window에서 후속 디코딩을 수행해야 함을 나타낸다. 그렇지 않으면 세그먼트를 포함하지 않도록 오디오를 자른다. 마지막으로 ⟨|endoftranscript|⟩⟨|endoftranscript|⟩ 토큰을 추가한다. 이전 컨텍스트 텍스트에 대한 학습 loss만 가리고 다른 모든 토큰을 예측하도록 모델을 학습시킨다. 형식 및 학습 설정에 대한 개요는 위 그림에 요약되어 있다.

**4. Training Details**

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

저자들은 Whisper의 스케일링 속성을 연구하기 위해 다양한 크기의 모델들을 학습시켰다 (위 표 참조). Dynamic loss scaling과 activation checkpointing과 함께 FP16을 사용하여 가속기에서 데이터 병렬 처리로 학습한다. 모델은 AdamW과 gradient norm clipping으로 학습되었으며 처음 2048번의 업데이트에 대한 warm-up 후 선형적으로 learning rate가 0으로 감소한다. Batch size로 256이 사용되었으며, 모델은 데이터셋에 대해 2~3회 통과하는 220번의 업데이트에 대해 학습된다. 몇 epoch 동안만 학습하기 때문에 overfitting은 큰 문제가 아니며 data augmentation 또는 정규화를 사용하지 않고 대신 일반화 및 견고성을 장려하기 위해 대규모 데이터 세트에 포함된 다양성에 의존한다.

초기 개발 및 평가 과정에서 저자들은 Whisper 모델이 화자의 이름에 대해 그럴듯하지만 거의 항상 잘못된 추측을 전사하는 경향이 있음을 관찰했다. 이는 사전 학습 데이터셋의 많은 전사에 말하는 사람의 이름이 포함되어 있어 모델이 예측을 시도하도록 권장하기 때문에 발생하지만 이 정보는 오디오 컨텍스트의 최근 30초에서만 거의 유추할 수 없다. 이를 방지하기 위해 화자 주석을 포함하지 않는 전사의 부분 집합에서 간략하게 Whisper 모델을 fine-tuning한다.

**Experiments**

**1. Zero-shot Evaluation**

Whisper의 목표는 데이터셋별 fine-tuning 없이 안정적으로 작동하는 강력한 단일 음성 처리 시스템을 개발하는 것이다. 이 능력을 연구하기 위해 다양한 기존 음성 처리 데이터셋을 재사용하여 Whisper가 도메인, task, 언어 전반에 걸쳐 잘 일반화할 수 있는지 확인한다. 각 데이터셋에 대한 학습 데이터를 사용하지 않고 zero-shot 세팅에서 Whisper를 평가하여 광범위한 일반화를 측정한다.

**2. Evaluation Metrics**

음성 인식 연구는 일반적으로 WER(단어 오차율)을 기반으로 시스템을 평가하고 비교한다. 그러나 문자열 편집 거리를 기반으로 하는 WER은 전사 스타일의 무해한 차이를 포함하여 모델의 출력과 레퍼런스 전사 사이의 모든 차이에 페널티를 준다. 결과적으로 시스템이 출력한 전사를 모두 인간이 올바르다고 판단하더라도 사소한 형식 차이로 인해 여전히 큰 WER을 가질 수 있다. 이것은 특정 데이터셋의 전사 형식의 예를 관찰하지 않는 Whisper와 같은 zero-shot 모델의 경우 특히 심각하다.

이것은 새로운 관찰이 아니다. 인간의 판단과 더 잘 연관되는 평가 지표의 개발은 활발한 연구 분야이며, 몇 가지 유망한 방법이 있지만 음성 인식에 널리 채택된 것은 아직 없다. 저자들은 non-semantic 차이에 대한 페널티를 최소화하기 위해 WER 계산 전에 텍스트를 광범위하게 표준화하여 이 문제를 해결하기로 했다. 본 논문에 사용된 text normalizer는 naive한 WER이 무해한 차이에 대해 Whisper 모델에 페널티를 부여한 일반적인 패턴을 식별하기 위해 반복적인 수동 검사를 통해 개발되었다. 여러 데이터셋의 경우 WER이 최대 50%까지 감소하는 것을 관찰했다. 일반적으로 데이터셋의 레퍼런스 전사가 공백이 있는 단어와 약어를 구분하는 등의 특이점으로 인해 발생한다. 이 개발 절차가 Whisper 모델의 전사 스타일에 overfitting될 위험이 있음을 주의해야 한다.

**3. English Speech Recognition**

2015년 Deep Speech 2는 LibriSpeech test-clean split을 전사할 때 인간 수준의 성능과 일치하는 음성 인식 시스템을 보고했다. 분석의 일부로 Deep Speech 2의 저자들은 일반 음성 시스템이 추가 도메인 적응 없이 깨끗한 읽기 음성을 추가로 개선할 여지가 거의 없다고 생각한다고 결론을 내렸다. 그러나 7년 후 LibriSpeech test-clean의 SOTA WER은 5.3%에서 1.4%로 73% 더 떨어졌으며, 보고된 인간 수준 오차율인 5.8%보다 훨씬 낮다. In-distribution의 hold-out 데이터에 대한 이러한 큰 규모의 성능 향상에도 불구하고 LibriSpeech에서 학습된 음성 인식 모델은 다른 세팅에서 사용될 때 사람의 오차율보다 훨씬 높다. 즉, in-distribution에서는 초인적인 성능을 보이고 out-of-distribution에서는 인간 이하의 성능을 보인다.

저자들은 이러한 격차의 상당 부분이 테스트셋에서 인간과 기계 성능으로 측정되는 서로 다른 능력을 통합하기 때문이라고 생각한다. 즉, 인간과 기계가 같은 시험을 치르고 있지만 서로 다른 능력이 시험된다는 것이다. 차이점은 테스트가 아니라 테스트 방법에서 발생한다. 인간은 종종 연구 중인 특정 데이터 분포에 대한 supervision이 거의 또는 전혀 없는 task를 수행하도록 요청받는다. 따라서 인간의 성능은 out-of-distribution 일반화의 척도이다. 그러나 기계 학습 모델은 일반적으로 평가 분포에서 많은 양의 supervision에 대한 학습 후에 평가되므로, 기계 성능은 in-distribution 일반화의 척도가 된다. 인간과 기계 모두 동일한 테스트 데이터로 평가를 받는 반면, 학습 데이터의 차이로 인해 두 가지 완전히 다른 능력이 측정된다.

광범위하고 다양한 오디오 분포에 대해 학습되고 zero-shot 세팅에서 평가되는 Whisper 모델은 잠재적으로 기존 시스템보다 훨씬 더 인간의 행동과 일치할 수 있다. 이것이 사실인지 또는 기계와 인간의 성능 차이가 아직 이해되지 않은 요인으로 인한 것인지 연구하기 위해 Whisper 모델을 인간의 성능과 표준 fine-tuning 모델과 비교하고 어느 것이 더 밀접하게 일치하는지 확인할 수 있다.

이 차이를 정량화하기 위해 많은 분포/데이터셋의 평균 성능인 **overall robustness**과 **effective robustness**을 모두 조사한다. 일반적으로 in-distribution인 레퍼런스 데이터셋과 하나 이상의 out-of-distribution 데이터셋 간의 예상 성능 차이를 측정한다. Effective robustness이 높은 모델은 레퍼런스 데이터셋에 대한 성능의 함수로 out-of-distribution 데이터셋에서 예상보다 더 잘 수행되며 모든 데이터셋에서 동일한 성능에 접근한다. LibriSpeech는 최신 음성 인식 연구의 중심 역할과 많은 릴리스 모델의 가용성으로 인해 레퍼런스 데이터셋으로 사용한다. 저자들은 12개의 다른 학술 음성 인식 데이터셋을 사용하여 out-of-distribution 행동을 연구하였다.

다음은 LibriSpeech에서의 WER과 3가지 out-of-distribution 데이터셋(Common Voice, CHiME-6, TED-LIUM)의 평균 WER을 plot한 그래프이다.

텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

다음은 다양한 데이터셋에서 effective robustness를 비교한 표이다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

위 그래프와 표에서 볼 수 있듯이 zero-shot Whisper 모델은 supervised LibriSpeech model과 매우 다른 robustness 속성을 가지며 다른 데이터셋에서 모든 벤치마크된 LibriSpeech model을 큰 차이로 능가한다.

이 발견은 모델의 zero-shot과 out-of-distribution 평가를 강조할 것을 제안한다. 특히 인간 성능과 비교하려고 시도할 때 오해의 소지가 있는 비교로 인해 기계 학습 시스템의 능력을 과장하는 것을 방지할 수 있다.

**4. Multi-lingual Speech Recognition**

다음은 Multilingual LibriSpeech (MLS)와 VoxPopuli에서 다국어 음성 인식 성능을 비교한 표이다.

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

이 두 벤치마크는 15개의 고유한 언어만 포함하기 때문에 다소 좁다. 거의 모든 언어가 인도-유럽 어족에 속하고 그 중 많은 언어가 리소스가 많은 언어이다. 이러한 벤치마크는 75개 언어의 음성 인식을 위한 학습 데이터를 포함하는 Whisper 모델의 다국어 기능을 연구할 수 있는 범위와 공간이 제한되어 있다.

저자들은 Whisper의 성능을 보다 광범위하게 연구하기 위해 Fleurs 데이터셋에 대한 성능도 측정하였다. 특히 저자들은 주어진 언어에 대해 가지고 있는 학습 데이터의 양과 해당 언어에 대한 downstream zero-shot 성능 사이의 관계를 연구하는 데 관심이 있었다. 아래 그래프에서 이 관계를 시각화한다.

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

WER의 로그와 언어당 학습 데이터 양의 로그 사이에 0.83의 강한 제곱 상관 계수가 있다. 이러한 로그-로그 값에 대한 선형 맞춤에 대한 회귀 계수를 확인하면 학습 데이터가 16배 증가할 때마다 WER이 반으로 줄어드는 추정치가 나온다. 또한 이 추세에 따라 예상보다 못한 성능이 낮은 언어 중 많은 수가 고유한 스크립트를 가지고 있고 인도-유럽 언어와 관련이 먼 언어이다 (ex. 히브리어(HE), 텔루구어(TE), 중국어(ZH), 한국어(KO)). 이러한 차이는 언어적 거리로 인한 전송 부족, 바이트 레벨 BPE tokenizer가 이러한 언어와 일치하지 않거나 데이터 품질의 차이로 인해 발생할 수 있다.

**5. Translation**

다음은 X→→en 음성 번역 성능을 나타낸 표이다.

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

다음은 언어당 번역 학습 데이터의 양과 Fleurs에서의 zero-shot 번역 성능의 상관 관계를 시각화한 그래프이다.

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

학습 데이터가 증가함에 따라 명확한 개선 추세가 있지만 제곱 상관 계수는 음성 인식에서 관찰된 0.83보다 훨씬 낮고 0.24에 불과하다. 부분적으로는 오디오 언어 식별 오류로 인해 더 noise가 많은 학습 데이터가 원인인 것으로 의심된다.

예를 들어, 웨일스어(CY)는 9,000시간의 번역 데이터를 가지고 있음에도 불구하고 BLEU가 13으로 예상보다 훨씬 더 나쁜 성능을 보인다. 이 많은 양의 웨일스어 번역 데이터는 전체 번역 데이터에서 4위를 차지한다. 검사 결과 대부분의 웨일스어 번역 데이터가 실제로는 영어 자막이 있는 영어 오디오이며 언어 식별 시스템에 의해 영어 오디오가 웨일스어로 잘못 분류되어 데이터셋 생성 규칙에 따라 전사 데이터가 아닌 번역 학습 데이터로 포함되었다.

**6. Language Identification**

다음은 Fleurs에서의 언어 식별 성능을 나타낸 표이다.

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

Whisper의 zero-shot 성능은 다른 supervised SOTA보다 떨어진다. 이는 Fleurs가 다루는 102개의 언어 중 20개가 Whisper 데이터셋에 없기 때문이다. 겹치는 82개의 언어에 대해서만 평가하면 성능이 80.3%까지 나온다고 한다.

**7. Robustness to Additive Noise**

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

위 그래프는 additive noise가 더욱 집중됨에 따라 ASR 성능이 저하되는 것을 보여준다. 낮은 noise(40dB SNR)에서 zero-shot 성능을 능가하는 모델이 많이 있다. 이러한 모델이 주로 LibriSpeech에서 학습된다는 점을 감안하면 놀라운 일이 아니다. 그러나 모든 모델은 noise가 더 강해짐에 따라 빠르게 저하되어 10dB 미만의 SNR의 additive pub noise에서 Whisper 모델보다 성능이 떨어진다. 이는 특히 pub noise와 같은 보다 자연스러운 분포 변화에서 noise에 대한 Whisper의 견고성을 보여준다.

**8. Long-form Transcription**

다음은 긴 형태의 전사를 가지는 7개의 데이터셋에 대해 Whisper가 다른 SOTA ASR 시스템들과 경쟁력이 있음을 보여준다.

도표, 스크린샷, 텍스트, 그래프이(가) 표시된 사진

자동 생성된 설명

**9. Comparison with Human Performance**

다음은 Kincaid46 데이터셋에서의 전사 능력을 인간 전문가와 함께 비교한 그래프이다.

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

**Analysis and Ablations**

**1. Model Scaling**

다음은 다양한 task에서 모델 크기에 따른 WER을 측정한 그래프이다.

라인, 그래프, 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

**2. Dataset Scaling**

다음은 데이터셋 크기에 따른 성능을 나타낸 표이다.

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

**3. Multitask and Multilingual Transfer**

다음은 영어 전용 모델과 다국어 멀티태스킹 모델을 학습량에 따른 평균 WER로 비교한 그래프이다.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

적당한 양의 컴퓨팅으로 학습된 작은 모델의 경우 실제로 task와 언어 간에 부정적인 transfer가 있음을 보여준다. 그러나 다국어 멀티태스킹 모델은 더 잘 확장되며 가장 큰 실험의 경우 다른 task에서 긍정적인 transfer를 보여주는 영어 전용 모델보다 성능이 뛰어나다.

**4. Text Normalization**

무해한 단어 오차를 줄이기 위해 Whisper와 공동으로 텍스트 정규화를 개발했기 때문에 정규화가 전사의 일반적인 변형을 해결하기보다 Whisper의 특성을 수정하는 데 overfitting할 위험이 있다. 이를 확인하기 위해 저자들은 본 논문의 normalizer를 사용한 Whisper의 성능과 FairSpeech 프로젝트에서 독립적으로 개발한 normalizer를 비교했다.

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

위 그림은 차이점을 시각화한다. 대부분의 데이터셋에서 두 개의 normalizer는 Whisper 모델과 비교된 오픈 소스 모델 간의 WER 감소에 큰 차이 없이 유사하게 수행되는 반면, 일부 데이터셋(WSJ, CallHome, Switchboard)에서는 본 논문의 normalizer가 Whisper 모델의 WER을 훨씬 더 많이 줄인다. 감소의 차이는 ground truth에서 사용하는 다양한 형식과 두 normalizer가 어떻게 패널티를 부과하는지 추적할 수 있다.

**5. Strategies for Reliable Long-form Transcription**

Whisper를 사용하여 긴 형식의 오디오를 전사하는 것은 모델의 30초 오디오 context window를 이동할 양을 결정하기 위해 타임스탬프 토큰의 정확한 예측에 의존하며, 한 window에서 부정확한 기록은 나중 window의 기록에 부정적인 영향을 미칠 수 있다. 저자들은 긴 형식 전사의 실패 사례를 피하는 데 도움이 되는 다양한 휴리스틱들을 개발했다.

먼저, greedy decoding에서 더 자주 발생하는 반복 looping을 줄이기 위해 로그 확률을 score function으로 사용하여 5개의 beam으로 beam search를 사용한다. Temperature 0부터 시작한다. 즉, 항상 확률이 가장 높은 토큰을 선택하고 생성된 토큰에 대한 평균 로그 확률이 -1보다 낮거나 생성된 텍스트의 gzip 압축률이 2.4보다 높을 때 temperature를 0.2씩 1.0까지 높인다. 적용된 temperature가 0.5 미만일 때 이전 window에서 전사된 텍스트를 이전 텍스트 조건으로 제공하면 성능이 더욱 향상된다. 저자들은 ⟨|nospeech|⟩⟨|nospeech|⟩ 토큰의 확률만으로는 음성이 없는 세그먼트를 구별하기에 충분하지 않지만 음성 없음 확률 임계값 0.6과 평균 로그 확률 임계값 -1을 결합하면 음성 활동 감지를 더 믿을 수 있게 만든다는 것을 발견했다. 마지막으로 모델이 입력의 처음 몇 단어를 무시하는 failure mode를 피하기 위해 초기 타임스탬프 토큰을 0.0~1.0초 사이로 제한했다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

위 표는 위의 각 개입을 추가하면 전반적으로 WER이 점진적으로 감소하지만 데이터셋 전체에 걸쳐 균등하지는 않음을 보여준다.