Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization

[University of Texas at Austin, Carnegie Mellon University]

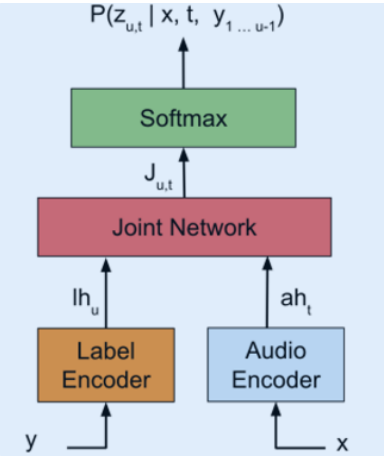
I. 논문 목적

- Whisper의 gradient-free zero-shot 과제 일반화 능력 최초 연구
- Whisper 모델에 대한 AVSR, CS-ASR, ST를 Zero-shot Prompting으로 성능 향상
- Whisper의 학습 측면 잠재 능력과 약점 연구

II. 기존 연구

- 대규모 선학습 모델 → 영어
  - ex) NLP’s LLM, CV’s Vision-and-Language Model
  - 처음 본 데이터나 과제를 처리하는데 domain-specific 데이터나 학습이 필요 없는 **zero-shot prompting**의 대두
    - Prompt: 원하는 출력값을 얻기 위해 AI가 수행해야 할 과제를 설명하는 NL text
    - PE(Prompt Engineering): 더 우수한 성능의 Prompt를 만들기 위한 과정
- Audio-Speech Processing에서의 PE는 최근에 각광
  - frozen GPT-2 가 음성 분류 위한 프롬프트 토큰 생성하게 wav2vec2 모델 FT
  - 음성 분류 & 생성을 위해 선학습된 음성단위 언어모델에 대한 그래디언트 기반 프롬프트 튜닝 연구
  - 음성 모델의 효율적 FT 위해 학습가능한 프롬프트와 어댑터 결합
  - 논문과 가장 유사: Transformer-Transducer 모델 학습, 처음 본 언어 쌍 음성 번역 모델의 FT위해 그래디언트 기반 적용 실험

        - Tranformer-Transducer: Transformer Encoder + RNN-T Loss → CTC 無



III. 제안 알고리즘(prompt)

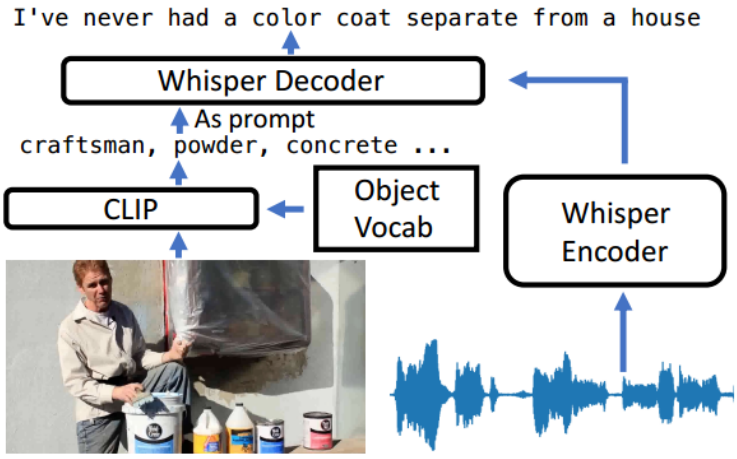
- whisper
    - 음성뿐만 아니라 음향, 다국어 번역 디코더 하나로 처리 web-scale speech model
        - Web-Scale ~= Big Data
    - 계열: 39M(tiny)~1.55B(large)개의 파라미터의 transformer기반 encoder-dedcoder 모델
    - 종류
        - 다국어
            - 데이터: 630k시간 웹스크래핑 데이터
            - 과제: 다국어 ASR, X→영 음성번역(ST), 언어 식별(LID), timestamp 예측(VAD)
        - 영어
            - 데이터: 438k(subset)
            - 과제: ASR, timestamp 예측(VAD)
    - 내부 구조
        - encoder: log Mel spectrogram로 특성 추출

        - decoder: 인코더에서 추출된 특성, 위치 임베딩, 프롬프트 토큰 시퀀스로 프롬프트에 따라 전사 or 번역 생성
        - default prompt: <|sop|>previous text<|sot|><|language|><|task|><|notimestamps|>
            - <||>: 특수 토큰, sop = startofprev, sot = startoftranscript
            - <|sop|> ~ previous text는 옵션
            - <|task|>: asr=transcribe or st=translate
            - <|notimestamps|>: timestamp 필요 無 → 논문에서 사용(필요x)
            - 다국어 모델 → 추론 시 언어 모르면 LID(99개언어) 이후 <|language|>값 채움
- 방법
    - 모델 가중치나 아키텍쳐 수정 無
    - default prompt 토큰의 일부만 수정

Task	Language(s)	Default prompt	Our proposed prompt	Improvement
AVSR	En	<|sot|><|en|><|asr|>	**<|sop|>CLIP retrie.**<default>	9%
CS-ASR	Zh+En	<|sot|><|zh|>or<|en|><|asr|>	<|sot|>**<|zh|><|en|>**<|asr|>	19%
ST	En→Ru	<|sot|><|ru|><|st|>	<|sot|><|ru|>**<|asr|>**	45%

IV. 실험 및 결과

1. **Audio-Visual Speech Recognition (AVSR)**
    - 일반적 AVSR: 얼굴(입술) 움직임 영상 & 음성의 멀티모달로 텍스트 생성
    - 과제에서의 AVSR: 음성과 의미적으로 관련되어 인식에 도움이 되는 영상 & 음성 멀티모달로 텍스트 생성



CLIP이 yogurt, heavy cream, mayonnaise와 같이 미관련 물체 출력

- 접근법: 음성 모델인 whisper를 AVSR에 이용하기 위해 Whisper에 시각 조건 프롬프트 제공
  - 시각 & 언어 모델인 CLIP + 외부 어휘 활용하여 먼저 시각 stream을 단어 토큰으로 변환
  - 세부 과정: 외부 어휘를 “This is a photo of a {}” 형식으로 문장 구성 → CLIP text encoder로 각 문장에 대한 임베딩 벡터 계산 → 각 영상에서 RGB 이미지 프레임 추출 → CLIP 이미지 인코더로 임베딩 → 이미지 임베딩과 텍스트 임베딩 간 유사성 계산 → 프롬프트로 최상위 K개 물체 선택해 comma-sep 리스트 생성 → 프롬프트의 previous text에 삽입

    ⇒ 사진에 포함된 물체를 previous text token으로
  - 영감: Socratic Model(대규모 선학습 모델이 복잡한 과제 수행을 위해 다른 모델과 talk하는 인터페이스)
- 데이터 및 세부사항
  - 시각-영상
    - VisSpeech: ⊂ HowTo100M(교육용 영상 데이터셋)
      - 음성만으로는 ASR의 성능 ↓ but 시각 stream & 음성이 의미적 연관 有
      - 테스트셋으로 제안됨, 508개 예시(↓)
    - How2(교육용 영상 데이터셋) 2000 랜덤 추출 + pub 소음
      - hyperparameter 튜닝 위해 사용
      - 소음 추가 이유: clean 음성에 편향 방지, 시각 양상이 음성인식 성능에 영향을 줘야 해서

- 외부 어휘 ← VLM(Visual-Language Model)
  - ViLD(VIsion and Language knowledge Distillation) 사용하여 장면에서 물체 묘사하여 언어모델에 맥락 전달
- 이미지-텍스트 라벨 셋: Tencent ML-Image(10k개의 일반 물체)
- K:= 5,0,−5dB(소음)에 상응하게 각 Whisper 모델마다 달리 설정

- 결과
  - 90 물체 → 성능 ↓ X / 30개 물체 사용했을 때도 CLIP 잘못 반환한 관련x 물체 有
  - How2 90% 발화 전사가 30개 단어 미만 → whisper가 잡음과 프롬프트 길이에 강건(robust)
  - 음성에 시각 정보 프롬프트 추가했을 때 영어 모델에선 모두 우수, 다국어 모델에서는 절반(작은모델)에서만 우수

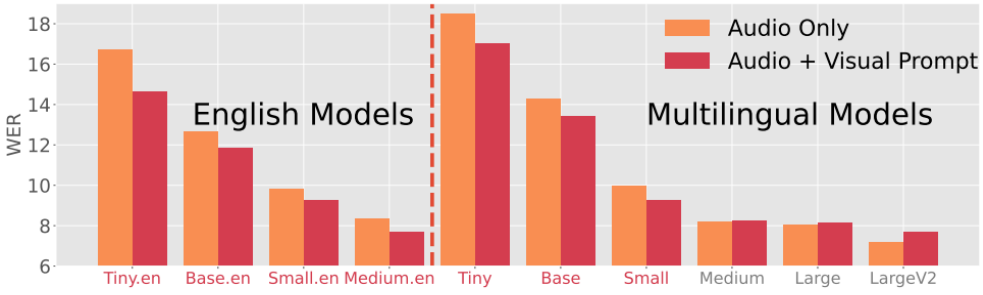


Figure 2: *The effectiveness of visual prompt on VisSpeech across different models.*

top3 K(물체 개수)가 실험에 사용됨

  - SotA AVSR & 음성whisper & 음성+ 시각 프롬프트 whisper 비교

Table 2: *Comparison of model performance on VisSpeech. With visual prompt, Medium.en outperforms Large.*

Model	Modality	WER
SotA [19]	A+V	11.28
Whisper Medium.en	A	8.35
Whisper Medium.en	A+V	7.60
Whisper Large	A	8.02
Whisper LargeV2	A	**7.16**

        - LargeV2 Whisper(다국어) > 시각 프롬프트 있는 Medium 영어 Whisper 모델 >> Large Whisper(다국어) > Medium 영어 Whisper > SotA AVSR
- 발견
    - 영어 Large 모델은 없기 때문에 모델사이즈나 다국어성이 시각프롬프트의 성능 향상에 방해되는지 알 수 X
    - medium은 영어/다국어모델 비교가 가능 → 시각 프롬프트의 성능저하가 다국어성 때문 → 단일어 데이터에 FT하여 다국어 모델 성능 향상 가능성

2. **Code-Switched Automatic Speech Recognition (CS-ASR)**
    - Code-Switched: 문장에 여러 언어 有, Code~=language
        - intra-sential: 정말 delicious!
        - inter-sential: 나는 오늘 10시에 출근했어. It was a tough day.
    - 접근법
        - whisper의 CS-ASR 능력 평가 위해 2개의 zh(Madarin)-En code-switched 말뭉치 사용
        - 기존에는 LID 결과 언어를 프롬프트에 사용하는 방식 → LID 능력에 의존, 억양 有 or 문장 내 CS 경우 성능 ↓

    - concat: 프롬프트에서 단일 언어 토큰(<|en|> or <|zh|>)을 두 언어 토큰(<|en|><|zh|> or <|zh|><|en|>으로 대체
- 데이터 및 세부사항
    - Mandarin-en CS 데이터
        - ASCEND(A Spontaneous Chinese-English Dataset for code-switching in multi-turn conversion): 다양한 중국어 방언 2개국어 화자
        - SEAME(South East Asia Mandarin-English): 싱가폴, 말레이시아 화자
    - hyperparameter
        - 프롬프트에서 두 언어 토큰 순서
            - large whisper는 <|zh|><|en|> 성능 ↑
        - LID confidence: 임계값 이상이면 1언어 토큰(이하면 concat(1언어, 2언어))
            - ASCEND에서는 0.9. 임계값이 최고 성능, SEAME에서는 1.0(항상 두 언어)
- 결과(Whisper Large)
    - MER(Mixed Error Rate): Mandarin: CER, English: WER ≠ Match Error Rate
        - CS MER: CS 발화에서의 MER
        - total MER: 전체 데이터셋에 대한 MER(CS 아닌 문장도 포함)

Dataset	Lang. prompt.	Zh CER	En WER	CS MER	Total MER
ASCEND	<|zh|>	**16.3**	93.1	33.1	32.6
	<|en|>	90.4	**31.5**	80.1	78.9
	default	17.0	31.8	26.6	22.1
	concat	16.6	31.8	**25.0**	**21.3**
SEAME	<|zh|>	26.3	97.4	43.3	46.7
	<|en|>	99.3	**33.8**	86.9	82.2
	default	27.1	85.5	43.2	45.3
	concat	**25.9**	44.7	**38.4**	**36.9**

- 비교를 위해 default, concat(제안), en, zh 모두 실험해서 concat이 최고성능
- SEAME total MER: concat(-19%) < default
- whipser 단일어 ASR 성능: ASCEND > SEAME
- default VS en의 en WER비교 → Whisper LID 성능 ASCEND > SEAME

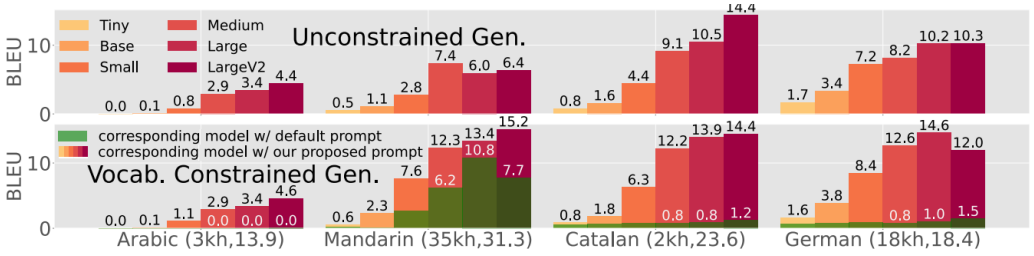
Dataset	Approach	Zh CER	En WER	CS MER	Total MER
ASCEND	Sup. SotA [27]	-	-	-	25.0
	Whisper+default	19.6	30.3	23.6	22.8
	Whisper+concat	16.8	30.8	22.0	**20.9**
SEAMEDEVMAN	Sup. SotA [28]	-	-	-	**16.6**
	Whisper+default	24.7	76.3	38.2	38.2
	Whisper+concat	23.6	45.8	33.4	32.7
SEAMEDEVSGE	Sup. SotA [28]	-	-	-	**23.3**
	Whisper+default	32.4	82.8	56.4	65.0
	Whisper+concat	31.0	46.7	49.6	47.6

- zero-shot whisper large VS 지도학습 SotA 테스트
- ASCEND: concat > default > SotA
- SEAME(말레이시아, 싱가폴): SotA > concat > default

- 발견
  - whisper의 LID & ASR 성능은 accent 따라 차이
  - CS 문장에 대해 단일어로 번역하는 오류 有 → 특히 en → zn의 결과 → 3과제에 대한 영감(Whisper는 en→X 학습X)

3. **Speech Translation (ST) ← 처음 본 언어 쌍, 영 ⇒ X**
   - 목표: SotA 이상의 성능 X
     - whisper의 모델 사이즈 & 데이터 크기 & 다국어성 & 병렬학습으로부터 기인된 zero-shot 번역(음성 이해) 능력 파악
   - 접근법
     - <|st|> task 토큰은 출력 어휘 제약주지 않으면 영어만 출력 ⇒ <|asr|> 토큰 & 목표언어토큰 사용
     - 비직관적이지만 default 보다 결과 우수, 일부 언어는 지도학습 성능

- 데이터 및 세부사항
  - Arabic, Mandarin, Catalan, German in CoVoST2
    - CoVoST2(COmmon VOice project database of multilingual and diversified Speech Translation corpus)
      - 21→En, En→15, 2,900 hr
    - → 리소스/위상학적으로 다양한 평가
  - En→Ru, En→De from MuST-C V1, En→Fr from LibriTrans
    - → (비)지도학습, 제로샷번역 과의 비교
  - 어휘 제약 시
    - Arabic, Mandarin, Russian → 스크립트에 있는 토큰만 포함하는 유니코드 범위의 어휘 사용
    - German, Catalan, French는 top K% 훈련셋 텍스트 어휘 사용 + K는 언어별로 CoVoST2에 튜닝
- 결과



  - BLEU(BiLigual Evaluation Understudy, y축): n-gram 일치를 사용하여 기계 번역 성능 & 품질 측정 지표
  - 어휘제약X(위)
    - default: 영어만 출력되기 때문에 성능 평가 불가
  - 어휘제약O(아래)
    - default : mandarin 제외하고 성능 매우 ↓
  - 종합: 평가 불가 제외하고 default < 제안 prompt, 모델크기 ↑ → 성능 ↑, 어휘 제약 성능 >> 제약 x

Category	Approach	En→De	En→Ru	En→Fr
Supervised	w2v2+mBART [30]	32.4	20.0	23.1
	E2E Transformer [35]	27.2	15.3	11.4
Unsupervised	Chung et al. [36]	-	-	12.2
	Cascaded [30]	22.0	10.0	15.4
	E2E (w2v2+mBART) [30]	23.8	9.8	15.3
Zero-shot	Escolano et al. [33]	6.8	-	10.9
	T-Modules* [34]	23.8	-	32.7
	Whisper w/ default prompt	0.4	8.8	0.8
	Whisper w/ our prompt	18.1	12.8	13.1

    - 제안 prompt 3 ST모두 성능 우수: En→Ru: 8.8(default) → 12.8(제안), 45% ↑)
    - 다른 접근법들: 기계 번역 시스템이나 다국어 문장 임베딩 모델 수정을 통해 얻은 성능(높을 수밖에)
- 추가 정보
    - En→X 학습 어려운 이유
        - <|st|>과제에서는 비영어 출력 학습 X
        - <|asr|> 과제에서는 입력 언어랑 다른 언어를 생성하는 학습 X
    - 간단한 프롬프트 수정만으로 영 ⇒ X 번역이 가능 → 서로 다른 언어 사이에서도 유사한 단어나 구가 잠재 공간 사이에서는 가까울 수 有

V. 실험 분석 및 결론

- 결과 요약
    - 3 과제에서 모두 제안한 프롬프트 >> 디폴트 프롬프트
        - 9, 19, 45 %로 성능 향상
    - 일부 데이터셋에서 SotA 지도학습 모델보다 성능 ↑
- 각 과제에서 발견한 whisper의 속성
    - AVSR: 시각 프롬프트의 길이와 잡음에 강건, 영어 모델과 다국어 모델에서 시각 프롬프트 효과 차이 有
    - CS-ASR: accent 따른 잠재 성능 차이

    - ST: asr 토큰이 st 토큰보다 st 성능 우수

→ Whisper의 강건성, 일반화 가능성, 비편향성 높이는 연구 필요

VI. 참고 문헌

- A. Zeng, M. Attarian, brian ichter, K. M. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, “Socratic models: Composing zero-shot multimodal reasoning with language,” in ICLR, 2023.
- D.-C. Lyu, T. P. Tan, C. E. Siong, and H. Li, “Seame: a mandarinenglish code-switching speech corpus in south-east asia,” in Interspeech, 2010.
- Dalmia, Siddharth, et al. "Transformer-transducers for code-switched speech recognition." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.
- H. Lovenia, S. Cahyawijaya, G. Winata, P. Xu, Y. Xu, Z. Liu, R. Frieske, T. Yu, W. Dai, E. J. Barezi, Q. Chen, X. Ma, B. Shi, and P. Fung, “ASCEND: A spontaneous Chinese-English dataset for code-switching in multi-turn conversation,” in LREC, 2022
- P.-A. Duquenne, H. Gong, B. Sagot, and H. Schwenk, “T-modules: Translation modules for zero-shot cross-modal machine translation,” in EMNLP, 2022.
- Salesky, Elizabeth, Julian Mäder, and Severin Klinger. "Assessing evaluation metrics for speech-to-speech translation." *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021.
- Wang, Changhan, Anne Wu, and Juan Pino. "Covost 2 and massively multilingual speech-to-text translation." *arXiv preprint arXiv:2007.10310* (2020).