Factual consistency oriented speech recognition [Microsoft]

- 논문목적
  - ASR에서 발생하는 hallucination을 줄이기 위해 factual consistency score을 최대화 하는 프레임 워크 제안.
    - hallucination: 언어 모델이 인간은 감지할 수 없는 부정확한 결과물을 생성하는 현상.
      - 사용자는 질문에대한 정확한 답변 기대하지만, AI 알고리즘은 훈련 데이터와 무관한 출력물을 생성.
      - 요인으로는 오버피팅, 데이터 편향, 모델의 복잡성등이 있음.
      - ex) bard: 최초의 외계행성 촬영은 james webb이 했음. (허블 우주망원경)
      - ex) 이미지처리에서 민무늬 벽인데 줄무늬로 표현함.
  - Factual consistency
    - ASR은 기술의 성장과 꾸준한 데이터의 품질향상으로 인해 과거에 비해 정확도가 향상 되었으나, 여전히 인식 오류에 취약.
    - 실제 전사: “I don’t know.” → N-best 결과:”I know.”, “I dunno.”

- WER관점에서는 “I Know.” 가 낫지만, Factual consistecy maximization 에서는 “I dunno.”가 낫다.
- 한국어 예) “오늘은 날씨가 맑아요.”가 실제 전사 → N-best 결과: “오늘은 날씨가 흐려요.”, “오늘은 날씨가 말가요.” → 후자가 전자보다 실제 전사와 일관성있음.

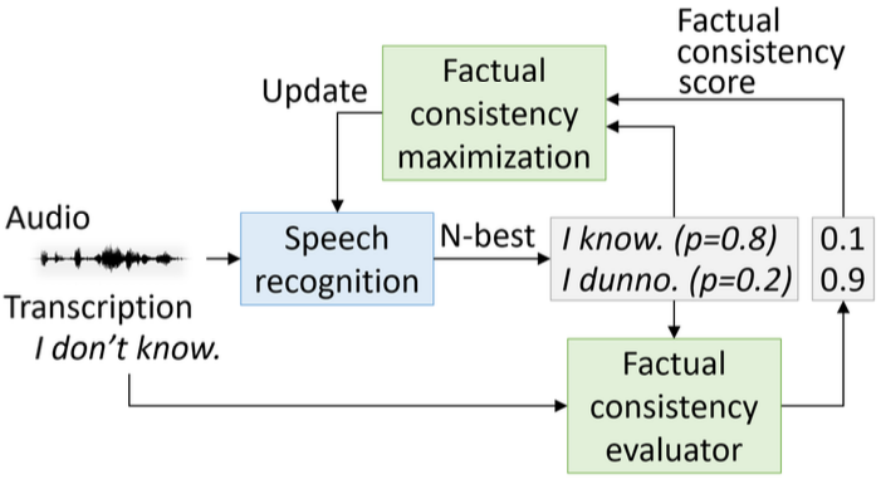

Figure 1: *Factual consistency maximization of ASR.*

- 기존연구
  - ASR 모델은 언어에 따라 WER로 평가되지만, NLP에서는 다양한 평가지표가 제안됨.
  - 최근에는 BERTScore, MoverScore같은 사전 훈련된 언어 모델을 기반으로 Factual consistency를 측정함.
  - 이번 연구에서는 UniEval을 사용하며, 요약평가에 유리한 UniEval-sum과 BERT 기반의 평가도구인 FactCC도 사용.

- 제안 알고리즘
  - 어텐션 인코더 디코더 기반의 모델 사용.
  - 빔 서치함수 사용하여 디코딩, N-best 결과물 생성.
  - 결과물의 사후 확률 구한뒤 Factual consistency score 구함.

    - 훈련 목적함수를 사용하여 역전파 하여 기울기 값을 그라디언트 상승 알고리즘을 기반으로 값 업데이트.

- 실험
    - 실험에서 사용하는 데이터는 AMI meeting corpus, VoxPopili corpus.
        - AMI meeting corpus: 100시간 가량의 회의 녹음 데이터.
        - VoxPopuli corpus: 23개국 언어 100K 시간의 라벨링 안된 데이터
          + 16개국 언어 1.8K 시간 전사 데이터(실험에서는 영어로 전사된 데이터 사용.)
        - 초기 모델은 whisper base 모델 사용, CE-loss 훈련, F.C.M 기반 모델 업데이트.
        - Factual consistency 평가 도구는 Unieval-fact 와 factCC 사용.
        - 요약평가에서는 데이터세트가 작아 AMI 데이터를 60초 길이로 분할한뒤 요약 텍스트를 생성 한뒤 진행.
        - 요약평가도구는 Unieval-sum 사용.
    - 결과
        - WER과 Factual consistency score에서 향상을 보임.
        - WER을 유지한채로 Factual consistency를 향상.
        - Factual consistency를 최대화 함이 실제 전사와 유사할 정도로 Factual consistency score를 향상.

Table 1: *WER (%), UniEval-fact consistency score (UE) and FactCC consistency score (FCC) for AMI-IHM development and evaluation sets. FCM: factual consistency maximization.*

ASR model	AMI-IHM dev			AMI-IHM eval		
	WER (↓)	UE (↑)	FCC (↑)	WER (↓)	UE (↑)	FCC (↑)
Whisper Base	19.7	0.727	0.925	20.3	0.719	0.913
↪ CE-loss	11.5	0.785	0.930	12.6	0.787	0.932
↪ FCM	**11.4**	**0.799**	**0.942**	**12.5**	**0.801**	**0.940**

Table 2: *WER (%), UniEval-fact consistency score (UE) and FactCC consistency score (FCC) for VoxPopuli development and test sets. FCM: factual consistency maximization.*

ASR model	VoxPopuli dev*			VoxPopuli test*		
	WER (↓)	UE (↑)	FCC (↑)	WER (↓)	UE (↑)	FCC (↑)
Whisper Base	9.4	0.755	0.896	9.4	0.747	0.887
↪ CE-loss	**8.2**	0.766	0.901	**8.3**	0.757	0.886
↪ FCM	8.4	**0.795**	**0.920**	8.5	**0.791**	**0.911**

* Utterances without case and punctuated transcriptions were excluded.

Table 3: *Summarization consistency score based on UniEval-sum for AMI development and evaluation sets. FC: factual consistency.*

ASR model	Summarization consistency (↑)	
	AMI dev	AMI eval
Whisper Base	0.697	0.739
↪ CE-loss fine-tuning	0.698	0.745
↪ FC maximization	**0.706**	**0.748**
Ground-truth transcription	0.709	0.753

ASR performance on the summarization.

- 결론
  - ASR 모델을 최적화하여 Factual consistency를 극대화 하는 새로운 프레임 워크 제안.
  - AMI와 Vox corpus를 사용한 실험에서 ASR 모델이 WER을 유지한채로 높은 Factual consistency score 을 갖는 결과 도출.
  - 제안된 프레임워크로 훈련된 모델을 사용시에 음성 요약 품질이 향상되어 NLP작업에서 유용함을 입증.