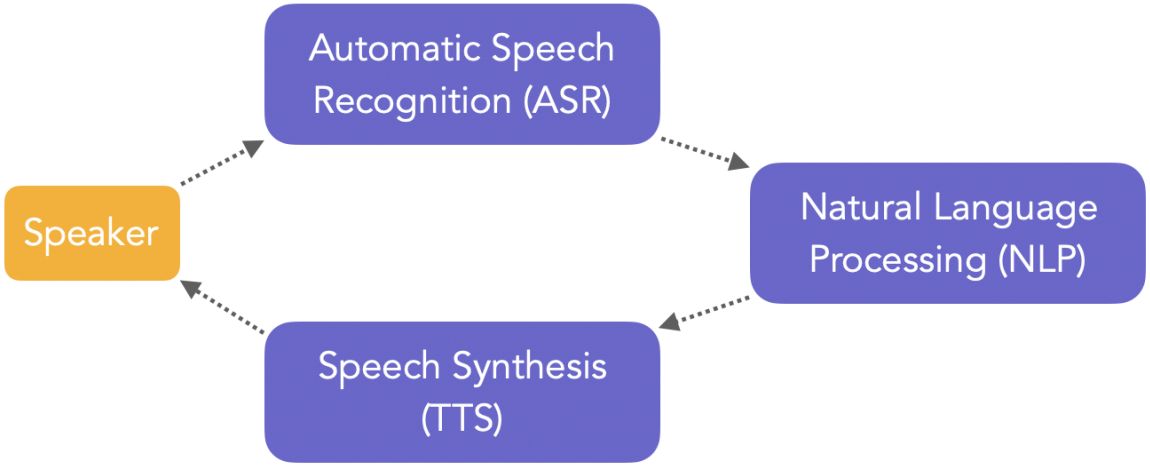
Adversarial teacher-student learning for unsupervised domain adaptation

Meng, Z., Li, J., Gong, Y., & Juang, B. H. (2018, April). *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5949-5953). IEEE.

Microsoft AI and Research, Redmond, WA, USA
Georgia Institute of Technology, Atlanta, GA, USA

Introduction

- **Automatic Speech Recognition (ASR)은 딥러닝 발전과 함께 성능이 향상 됨**
- **그러나 well-trained 음향 모델이 새로운 도메인에 적용될 때 ASR은 여전히 큰 성능 저하를 겪고 있음**



Problem statement

- ASR Unsupervised domain adaptation 시 발생하는 Condition variabilities로 인한 성능 저하 문제를 해결하고자 함

Key idea

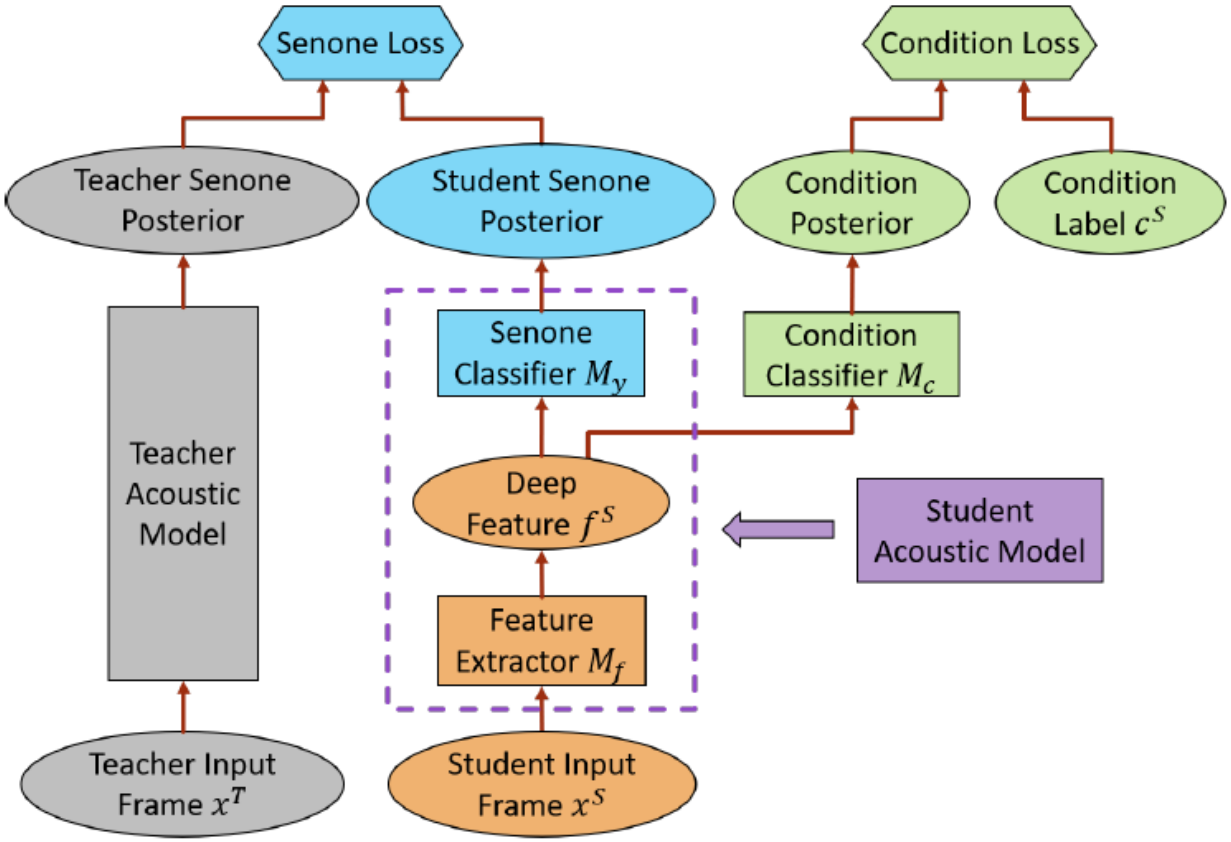
- **Adversarial teacher-student learning 제안**

  - 음성 신호 내의 condition (speaker and environment) 변동성을 효과적으로 억제

  - Robust unsupervised adaption

Method

- **The framework of adversarial teacher-student learning for unsupervised adaptation of the acoustic model**



Method

- Teacher model : 레이블이 있는 source 도메인에서 사전 훈련 됨. Training 중 student model에 가이드 제공
- Student model : target 도메인에서 훈련됨. Target 도메인에 대해 Teacher model의 출력을 모방
- Student acoustic model, condition classifier를 함께 최적화 시켜 teacher model과 student model 출력 분포 간 Kullback-Leibler 발산 최소화
- Adversarial Training : student model은 source 도메인과 target 도메인 차이를 줄이기 위해 adversarial 방식으로 training 됨. 도메인 판별기는 소스 및 타겟 도메인의 feature를 서로 구별할 수 없게 만드는 것을 목표함

Method

- Long Short-Term Memory (LSTM) 모델 사용

- Teacher LSTM model : 소스 도메인에서 먼저 training. 데이터의 특정 패턴과 정보를 학습하며 사후 확률 생성

- Student LSTM model : Teacher model의 사후 확률을 가이드로 사용하여 타겟 도메인의 데이터에서 training. Student model은 Teacher model의 지식을 가져와 타겟 도메인에서의 변동성과 노이즈에 대해 robust함

Dataset

- CHiME-3 (The 3rd CHiME Speech Separation and Recognition Challenge)
- 음성 처리 및 음성 인식 연구 분야에서 사용되는 벤치마크 데이터셋

- 음성 데이터: 음성 데이터는 여러 환경에서 녹음된 실제 대화 내용으로 구성되며, 다중 화자가 함께 발화하는 경우 포함
- 노이즈 데이터: 노이즈 데이터는 각 환경 조건에서 녹음된 노이즈 신호. 음성 데이터 + 노이즈가 결합된 형태의 데이터

Result

- Unadapted System과 T/S System에서의 WER(%) 성능 비교

System	Adaptation Data	BUS	CAF	PED	STR	Avg.
Unadapted	-	27.93	24.93	18.53	21.38	23.16
T/S	clean-noisy	16.00	15.24	11.27	13.07	13.88
	clean-noisy, clean-clean	15.96	14.32	11.00	13.04	13.56

Word Error Rate(WER) : 음성 인식 성능을 나타내는 지표. 인식된 결과와 실제 정답 간 차이 측정

Result

- Advesarial teach-student WAR(%) 성능 비교

System	Conditions	BUS	CAF	PED	STR	Avg.
Adversarial T/S	2 environments	15.24	13.95	10.71	12.76	13.15
	6 environments	15.58	13.23	10.65	13.10	13.12
	87 speakers	14.97	13.63	10.84	12.24	12.90
	87 speakers, 6 environments	15.38	13.08	10.47	12.45	**12.83**

Word Error Rate(WER) : 음성 인식 성능을 나타내는 지표. 인식된 결과와 실제 정답 간 차이 측정

Conclusion

- Unsupervised domain adaptation 위해 Adversarial T/S learning 제안함
- 음성 신호에서 조건 변동성을 억제하고 견고한 적응을 달성하기 위해 student acoustic model과 condition classifier을 함께 최적화 시킴
- Adversarial T/S learning 은 상대적으로 작은 CHiME-3 데이터셋으로 효과성이 검증됨

Q & A