Joint Speaker Diarization and Automatic Speech Recognition

Srikanth Raj[1], Shakti Rath Prasad[1], Prachi Singh[1], Sriram Ganapathy[1], Kautubh Kulkarni
Shakti Srivastava[2], Michael Free[2]
[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Department of Electrical Engineering, Indian Institute of Science, Bengaluru
[2]British Telecom Research Center Bengaluru/UK.

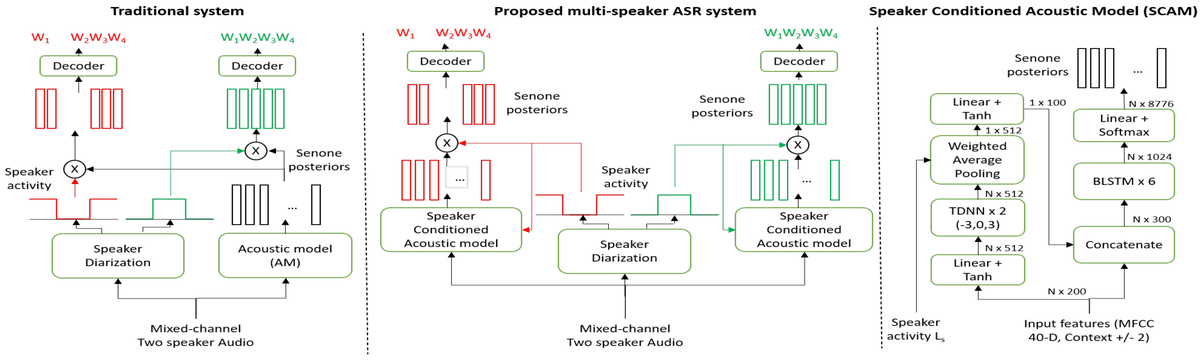
1 Introduction

- Transcribe speech to text: **Automatic speech recognition** (ASR)
- Segment the speech at the speaker level: **Speaker diarization**
- Speech transcription indexed by speaker: **Joint ASR and Diarization**

2 Applications and Challenges

- Call center conversations
- Meeting transcription
- Captions for TV and movies
- Multiple speakers, Unknown number of speakers
- Noise, reverberation, interfering speakers
- Conversational speech (different from command-style speech)

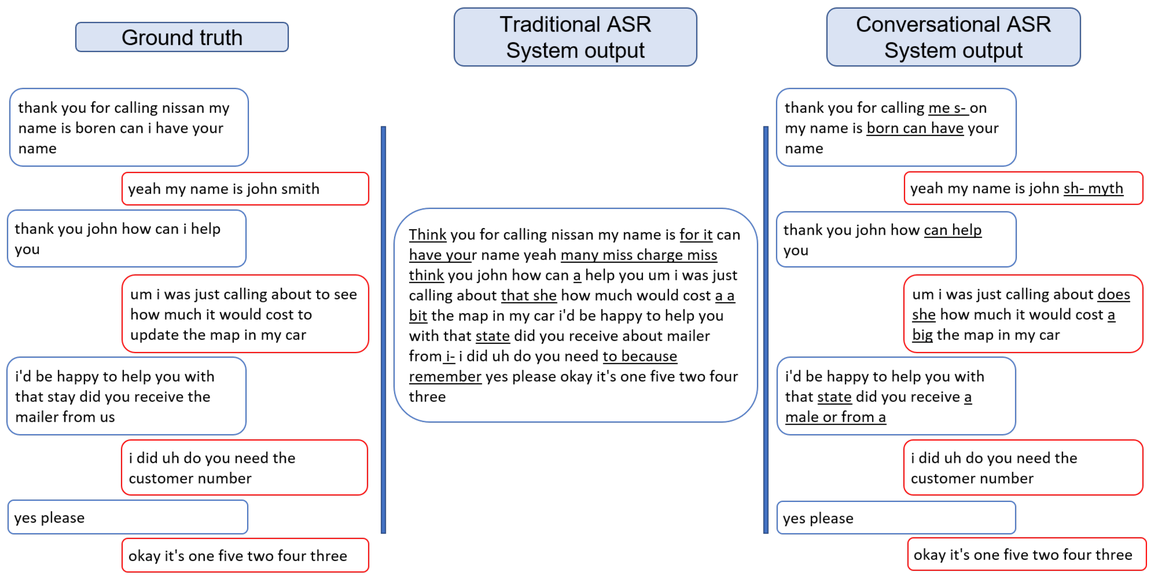
3 Joint ASR and Diarization



4 System description

- Switchboard – 300h. Telephone conversations between two parties.
- BLSTM based architecture is used for the acoustic model.
- Language model trained using the Switchboard and Fisher English conversations transcripts
- End-to-End speaker diarization module with 4 self-attention layers and LSTM encoder-decoder pair to generate speaker attractors
- Evaluation uses Hub5 evaluation set: 20 conversations from CallHome and Switchboard. Performance quantified using the speaker word error rate (SWER).

5 Example transcription



6 Speaker specific Word error rate (%)

System	GTS	EEND
Single-channel		
BLSTM-iso	21.4	24.0
Mixed-channel		
BLSTM-iso	37.7	37.2
ConvTasNet	25.3	27.4
SCAM	27.5	26.9
SCAM*	26.0	26.1
Joint training	23.4	24.1

System	Non-overlap	Overlap
BLSTM-iso	27.1	41.5
SCAM	31.4	26.1
SCAM*	29.8	24.7

References

[1] Dong Yu and Li Deng, "Automatic Speech Recognition: A Deep Learning Approach," Springer 2015
[2] Laurent El Shafey, Hagen Soltau, and Izhak Shafran, "Joint Speech Recognition and Speaker Diarization via Sequence Transduction," in Proc. Interspeech, 2019, pp. 396–400
[3] Shota Horiguchi et al., "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in Proc. Interspeech, 2020, pp. 269–273.