



HAL
open science

End-to-end Multichannel Speaker-Attributed ASR: Speaker Guided Decoder and Input Feature Analysis

Can Cui, Imran Ahamad Sheikh, Mostafa Sadeghi, Emmanuel Vincent

► **To cite this version:**

Can Cui, Imran Ahamad Sheikh, Mostafa Sadeghi, Emmanuel Vincent. End-to-end Multichannel Speaker-Attributed ASR: Speaker Guided Decoder and Input Feature Analysis. Rencontre des Jeunes Chercheurs en Parole 2023 - 10E Edition, Nov 2023, Grenoble, France. hal-04321252

HAL Id: hal-04321252

<https://hal.science/hal-04321252>

Submitted on 4 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

End-to-end Multichannel Speaker-Attributed ASR: Speaker Guided Decoder and Input Feature Analysis

Can Cui, Imran Sheikh, Mostafa Sadeghi, Emmanuel Vincent



Task

- ▶ Transcribe a multichannel multi-speaker meeting

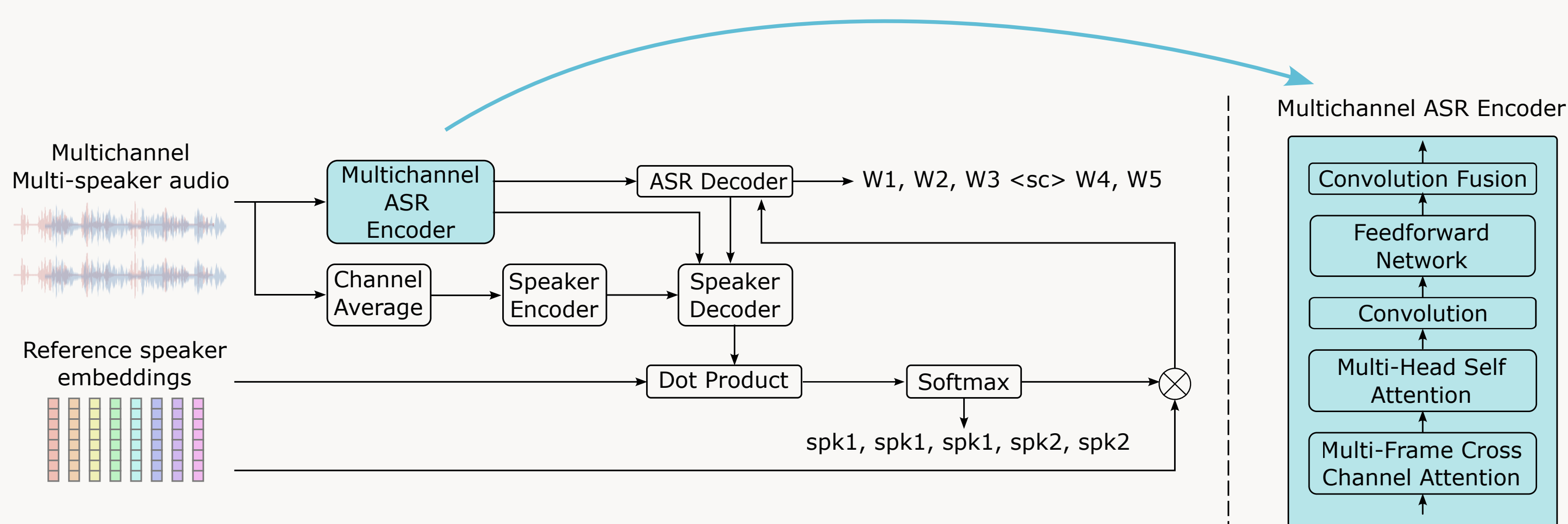
Key challenges

- ▶ Complex audio signals containing overlapping speech, noise, and reverberation
- ▶ Information sharing between ASR and speaker identification modules
- ▶ Design of optimal training losses, regularization, and training schedules of the E2E model

Main contributions

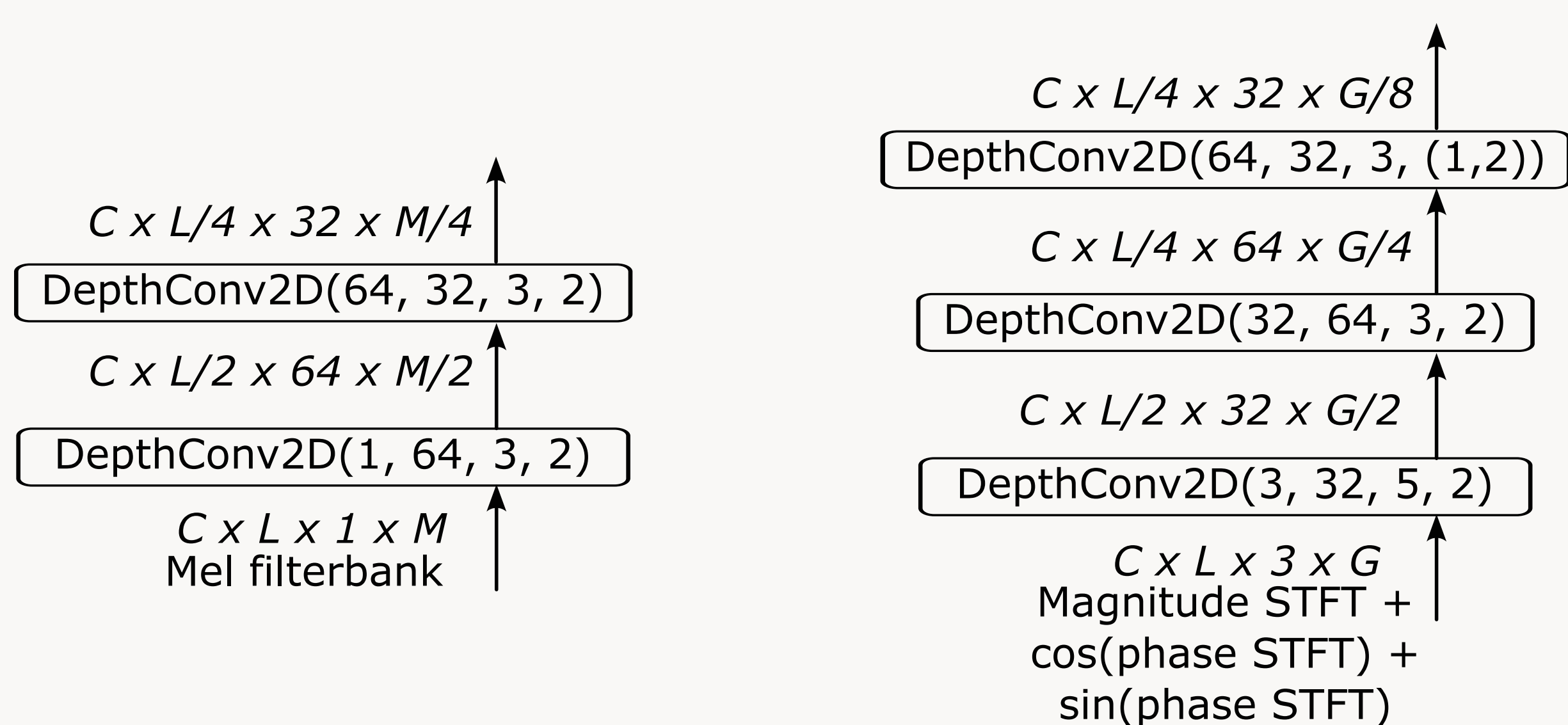
- ▶ E2E Multichannel Speaker-Attributed ASR (MC-SA-ASR) system
- ▶ Impact of multichannel input features

Proposed methods: MC-SA-ASR



- ▶ Proposed MC-SA-ASR combines Conformer-based multichannel Encoder [1] and Transformer-based speaker-wise Decoder [2]
- ▶ $W1, W2, W3 <sc> W4, W5$ where $<sc>$ indicates speaker change

Proposed methods: Input features



Experiments on LibriSpeech: Data preparation

- ▶ Data: LibriSpeech train-960, dev-clean, test-clean
- ▶ Preparation for multichannel multi-speaker setting on train-dev-test set: Room Impulse Responses simulated by the gpuRIR toolkit; Dynamic mixing for 2 and 3 speakers without duplication

References

- ▶ F. Yu, S. Zhang, P. Guo, Y. Liang, Z. Du, Y. Lin, and L. Xie, "MFCCA: Multi-frame cross-channel attention for multi-speaker ASR in multi-party meeting scenario," in *SLT*, 2023.
- ▶ N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-end speaker-attributed ASR with Transformer," in *INTERSPEECH 2021*, 2021.

Experiments on LibriSpeech: Results and discussion

Table: WER (%), sentence-level speaker error rate (SER) (%) on test set.

System	Input	# Ch	1-spk		2-spk mix		3-spk mix		1,2,3-spk mix		#Prm
			WER	SER	WER	SER	WER	SER	WER	SER	
Baseline											
SC-SA-ASR	Mel	1	7.15	1.64	14.68	3.23	20.39	5.59	16.81	3.95	61.1M
		2	8.52	-	16.76	-	22.87	-	18.21	-	
		4	8.11	-	16.43	-	21.76	-	17.84	-	
Proposed											
MC-SA-ASR	Mel	2	6.64	1.15	14.21	3.19	19.03	5.92	15.41	4.34	51.6M
		3	6.62	1.11	14.24	3.15	18.92	6.11	15.25	4.12	
		4	7.17	1.41	14.09	3.06	18.70	6.06	14.77	4.05	
		2	7.24	1.53	15.42	3.85	20.34	7.06	16.76	4.87	
		Mag+phase	3	6.86	1.78	13.69	3.72	18.14	6.86	15.04	
4	6.91	1.86	13.97	3.10	18.03	6.69	14.69	4.12			

Experiments on AMI: Data preparation

- ▶ Data: 100 hours of meeting data with 3–5 participants and 8–16 microphones
- ▶ Dividing each meeting (approximately 1 hour) into fixed-size chunks while avoiding speaker overlaps or cutting within a word
- ▶ The training and development sets consist of 5 s, 10 s, or 15 s chunks only. The test set comprises 5 s, 10 s, and 15 s chunks

Experiments on AMI: Results and discussion

Table: WER (%) and token-level SER (%) of models adapted to AMI.

System	1-spk		2-spk mix		3-spk mix		4-spk mix	
	WER	SER	WER	SER	WER	SER	WER	SER
SC-SA-ASR								
5 s	28.10	13.37	42.03	25.53	54.04	35.65	67.27	43.17
10 s	31.84	17.25	43.93	27.88	54.20	35.82	67.92	43.49
15 s	48.38	32.38	60.73	43.12	64.39	44.53	73.68	49.40
MC-SA-ASR (prop)								
5 s	27.68	13.43	39.54	24.83	52.76	35.70	64.72	41.24
10 s	31.81	17.52	43.77	27.18	53.92	34.99	66.54	42.80
15 s	48.69	31.79	60.34	42.48	63.66	44.19	72.69	48.36

Conclusions

- ▶ WER reduction up to 16% relative on simulated data
- ▶ Phase features perform better on chunks with a larger number of channels and/or with a larger number of speakers

Perspectives

- ▶ Improve the speaker count accuracy on real data
- ▶ Make effective usage of multichannel phase features for a better localization of speakers