



AISHELL-5: The First Open-Source In-Car Multi-Channel Multi-Speaker Speech Dataset for Automatic Speech Diarization and Recognition

Yuhang Dai¹, He Wang¹, Xingchen Li¹, Zihan Zhang¹, Shuiyuan Wang¹, Lei Xie^{1*}, Xin Xu²,
Hongxiao Guo², Shaoji Zhang², Hui Bu², Wei Chen³

¹Audio, Speech and Language Processing Group(ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, Xi'an, China

²Beijing AISHELL Technology Co., Ltd., Beijing, China

³Li Auto Inc., Beijing, China

yhdai@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

This paper delineates AISHELL-5, the first open-source in-car multi-channel multi-speaker Mandarin automatic speech recognition (ASR) dataset. AISHELL-5 includes two parts: (1) over 100 hours of multi-channel speech data recorded in an electric vehicle across more than 60 real driving scenarios. This audio data consists of four far-field speech signals captured by microphones located on each car door, as well as near-field signals obtained from high-fidelity headset microphones worn by each speaker. (2) a collection of 40 hours of real-world environmental noise recordings, which supports the in-car speech data simulation. Moreover, we also provide an open-access, reproducible baseline system based on this dataset. This system features a speech frontend model that employs speech source separation to extract each speaker's clean speech from the far-field signals, along with a speech recognition module that accurately transcribes the content of each individual speaker. Experimental results demonstrate the challenges faced by various mainstream ASR models when evaluated on the AISHELL-5. We firmly believe the AISHELL-5 dataset will significantly advance the research on ASR systems under complex driving scenarios by establishing the first publicly available in-car ASR benchmark. **Index Terms:** AISHELL-5, in-car speech processing, speech frontend, speech recognition.

1. Introduction

Unlike common automatic speech recognition (ASR) applications in the home or smart assistant scenarios, in-car ASR systems encounter a range of unique challenges. These systems must contend with various internal and external noise sources, including wind, engine sound, tire noise, car stereos, and nearby vehicles, all of which contribute to a highly complex acoustic environment. Moreover, conversations between the driver and passengers frequently introduce overlapping speech, further complicating the recognition progress. Fundamentally, building a robust in-car speech recognition system hinges on addressing two key issues: the complex acoustic environment, and the frequent occurrence of overlapping speech. While innovative designs for speech frontend and recognition model architectures are crucial, there is an increasing demand among researchers for specialized open-source datasets that can facilitate in-car data augmentation and model training.

First, the issue of speaker overlap has attracted a lot of attention from researchers, leading to the development of several open-source datasets specifically designed to cover this situation. AISHELL-4 [1] and AliMeeting [2] are both multi-speaker Mandarin ASR datasets recorded by multi-channel mi-

crophone arrays in indoor meeting scenarios, of which the speech captures lots of key characteristics of free-talk conversation, such as pauses and speaker overlaps. To date, these two datasets have played a crucial role in evaluating ASR performance in multi-speaker Mandarin meeting scenarios, effectively testing the robustness of models against far-field and overlapping speech. LibriMix [3] is an open-source English dataset designed for speech separation and multi-speaker ASR in noisy environments. It derives from the freely available LibriSpeech [4] dataset and WHAM! [5] noise dataset, where LibriSpeech provides clean speech signals and WHAM! provides noise samples to create noisy mixtures. However, the data in the LibriMix dataset, which is constructed by directly splicing and overlapping existing speech samples, cannot accurately reflect real-world scenarios. Except for the open-source datasets, some researches [6, 7] bring up various data simulation methods, which alleviate the issue of low model accuracy caused by data shortage to a certain extent.

Second, regarding the complex in-car acoustic environment, only a few open-source datasets can cover it. MDT-ASR-C001¹ is recorded in real in-car environments, reflecting real acoustic conditions and background noise. However, it comprises only 6 hours of data, which is insufficient to cover the complexities of real-world in-car driving scenarios. The Intelligent Cockpit Speech Recognition Challenge (ICSRC) 2022 [8] was successfully held and introduced a 20-hour single-channel in-car ASR test set, recorded by a high-fidelity microphone in a hybrid electric vehicle, focusing on the speech command recognition within smart cockpits. Although the ICSRC dataset is indeed pioneering in the field of in-car ASR, its limited data size and focus on vehicle control commands restrict its applicability for enhancing ASR systems designed for multi-speaker dialogues in diverse driving conditions. In 2024, the In-Car Multi-Channel Automatic Speech Recognition (ICMC-ASR) challenge [9] was launched, attracting nearly 100 participating teams and focusing on advanced in-car speech processing under complex driving scenarios. The challenge consists of two tracks: Track I, designated as the ASR track, provides the ground-truth timestamps indicating when each speaker speaks. Its evaluation set (**Eval1**) is measured by character error rate (CER) as the evaluation metric; Track II is for automatic speech diarization and recognition (ASDR), with no timestamps provided in the evaluation set (**Eval2**), which requires a system to do speaker diarization first to get predicted timestamps of every speaker, and then transcribe their speech separately, with concatenated minimum permutation character error rate (cpCER) as the evaluation metric.

*Corresponding author.

¹<https://www.magicdatatech.cn/datasets/asr/mdt-asr-c001-mandarin-chinese-speech-recognition-corpus>

Table 1: An illustration of the diverse sub-scenes in AISHELL-5, including scene numbers (day and night), window status, car driving status, air-conditioning status, and car stereo status. Apart from the sub-scene shown in the table, number 1 represents the day scene, while number 2 denotes the night scene. For example, A1 indicates a stopped car with both the air-conditioning and car stereo turned off during the day scene.

N: Window open				M: All windows are closed			
Window state				Car state			
Index	Driver's side window	Sunroof		Index	Drive state	AC	Car Stereo
N1	Open 1/3	Closed		A	Stopped	Off	Off
N2	Closed	Open 1/2		B	Stopped	Medium	Off
N3	Open 1/2	Open 1/2		C	Stopped	High	Medium
Car state				D	0-40 km/h	Off	Off
Index	Drive state	AC	Car Stereo	E	0-40 km/h	Medium	Off
A	Stopped	Off	Off	F	0-40 km/h	High	Medium
B	Stopped	Medium	Off	G	40-80 km/h	Off	Off
C	Stopped	High	Medium	H	40-80 km/h	Medium	Off
D	0-60 km/h	Off	Off	I	40-80 km/h	High	Medium
E	0-60 km/h	Medium	Off	J	80-120 km/h	Off	Off
F	0-60 km/h	High	Medium	K	80-120 km/h	Medium	Off
				L	80-120 km/h	High	Medium

As mentioned above, to enhance the accuracy and robustness of in-car ASR systems, there is an urgent need for a sizable open-source dataset that can simultaneously encompass speaker overlapping situations and different acoustic environments under various driving scenarios. To further promote research on in-car speech processing, we fixed all the data-related issues of the ICMC-ASR challenge dataset, including audio truncation and mismatched transcription, and now officially open-source it, called AISHELL-5². It features multi-channel, multi-speaker free-talking, as close as possible to real-world scenarios, with an over 100-hour scale. In particular, it is recorded under 60 driving scenarios by varying lots of factors that may alter the in-car acoustic environment, including the driving speed, car window, car stereo, air-conditioning, driving day or night, and so on. Moreover, 40 hours of real-recorded multi-channel noise signals are also open-sourced for promoting in-car speech data simulation research. Based on the AISHELL-5 dataset, we also provide a baseline system that includes a frontend incorporating speech source separation and an ASR module. Furthermore, we adopt the track settings and the evaluation sets from the ICMC-ASR challenge. We believe that AISHELL-5, as a real-recorded and high-quality in-car speech dataset, can offer a certain amount of support for this increasingly important ASR application scenario, and contribute to advancing human-vehicle interaction towards greater accuracy and convenience.

2. Dataset

The AISHELL-5 dataset is recorded inside a hybrid electric car, with a far-field microphone placed above the door handles of all four doors to capture far-field audio from different areas of the car. Additionally, each speaker wears a high-fidelity microphone to collect near-field audio for data annotation. A total of 260 participants are involved in the recording with no notable accents. During the recording, 2-4 speakers are randomly seated in the four positions inside the car and engaged in free

conversations without content restrictions to ensure the naturalness and authenticity of the audio data. The average duration of each session is 10 minutes. The scripts for all our speech data are prepared in TextGrid format. Each session's TextGrid contains information such as the session duration, speaker details (number of speakers, speaker IDs, gender, etc.), timestamps for each audio segment, and the transcribed text.

In normal driving scenarios, the car typically contains various noises from both inside and outside. External noises include environmental sounds, wind noise, tire noise, etc., while internal noises come from sources like music players and air conditioning. These noises significantly impact the accuracy of in-car speech recognition systems. To comprehensively cover the various noise types encountered in real-world in-car scenarios, we carefully design the recording scenes. For environmental noise, recordings are made with different driving segments (urban streets and highways) during both daytime and nighttime. About the wind-induced noise and tire noise, we control the degree to which the car windows are open (fully closed, half open, and one-third open) and the car's speed (stationary, low-speed, medium-speed, and high-speed). For noise inside the car, we set the music player and air conditioning to different levels to cover a variety of in-car conditions. These different sub-scenes are numbered, and all sub-scenes are combined in various ways to form the final recording scenarios, resulting in over 60 recording scenarios in total. Specific settings for the sub-scenes are shown in Table 1.

Overall, the AISHELL-5 dataset contains more than 100 hours of speech data, divided into 94 hours of training data (Train), 3.3 hours of validation data (Dev), and two test sets (Eval1 and Eval2), with durations of 3.3 and 3.58 hours. Each dataset includes far-field audio from 4 channels, with only the training set containing near-field audio. Additionally, to promote research on speech simulation techniques, we also provide a large-scale noise dataset (Noise), which has the same recording settings as the far-field data but without any speaker speech, lasting approximately 40 hours. Detailed information about the

²https://www.aishelltech.com/aishell_5

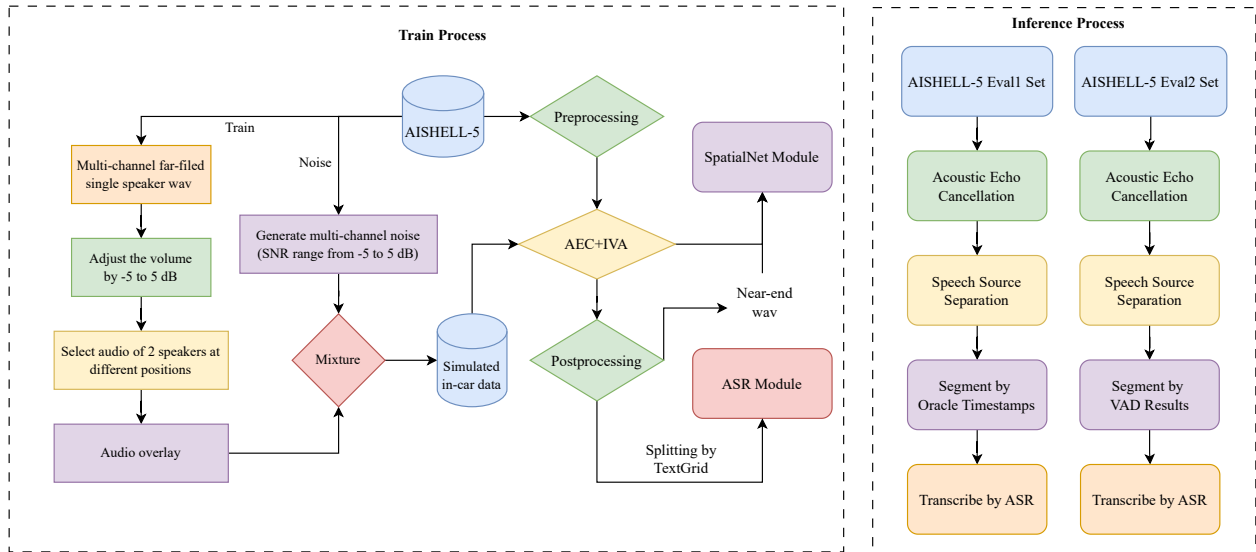


Figure 1: Structure of our baseline system, including train process and inference process

subsets of AISHELL-5 is provided in Table2.

Table 2: Statistics about AISHELL-5 datasets, including the duration of segmented near-field audio (Duration), number of sessions (Session), number of speakers (Speaker) and including near-field audio or not (Near-field).

Dataset	Duration (h)	Session	Speaker	Near-field
Train	94.75	568	147	✓
Dev	3.33	18	6	✗
Eval1	3.30	18	6	✗
Eval2	3.58	18	6	✗
Noise	40.29	60	-	-

3. Baseline

We develop a multi-channel in-car speech transcription system³ based on the ICMC-ASR baseline⁴. The baseline system consists of two primary sub-modules: speech frontend processing and automatic speech recognition (ASR). We provide the data preprocessing pipeline in the baseline system, and after processing the data, we train each sub-module independently.

During evaluation, Eval1 is directly processed by the ASR module. However, Eval2 undergoes a preprocessing step through the speech frontend module for echo cancellation and noise reduction before being fed into the ASR system. This process generates audio data for each speaker, which is subsequently segmented based on voice activity detection (VAD). Finally, the processed audio is passed through the ASR module to obtain the transcription. The training and inference process of our baseline is shown in Figure 1.

³<https://github.com/DaiYvhang/AISHELL-5>

⁴<https://github.com/MrSupW/ICMC-ASR.Baseline>

3.1. In-car Speech Frontend Processing

In our baseline system, we adopted a multi-channel acoustic echo cancellation (AEC) and an Independent Vector Analysis (IVA) [14] blind source separation algorithm. AEC is used to remove echo from the recorded microphone signal and IVA is used to separate speakers from different locations. For AEC, we use an adaptive filter to estimate and remove the echo from the microphone signal. Assuming that we have M microphones, the signal mixture can be represented as Eq. 1.

$$\mathbf{y}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t). \quad (1)$$

Here, t is the time index, and $\mathbf{y}(t)$ is the vector of the mixed signals received by the M microphones. \mathbf{A} is an $M \times N$ mixing matrix, where N is the number of speakers. $\mathbf{s}(t)$ represents the source signals of N speakers, and $\mathbf{n}(t)$ is the noise vector. In AISHELL-5, $M = 4$ and $N = 2$. The goal of IVA is to separate the independent audio $\mathbf{s}(t)$ for each seat from the mixed audio $\mathbf{s}(t)$. IVA is a maximum likelihood estimation (MLE) based blind source separation algorithm that separates the sources by minimizing the statistical dependence among the source signals. In IVA, assuming that the source signals are independent, the separation process can be achieved by maximizing the following objective function:

$$\mathcal{L} = \sum_t \log(\det(\mathbf{A}(\mathbf{y}(t))))). \quad (2)$$

This objective function maximizes the statistical independence of the source signals to find the demixing matrix \mathbf{A} , thus achieving the separation of the mixed signals.

In addition, our baseline system also integrates an end-to-end dereverberation, denoising, and separation approach based on Spatialnet, implemented using the NBSS⁵. Spatialnet makes extensive use of spatial information to perform multi-channel joint dereverberation, denoising, and separation, making it an ideal solution for in-car speech frontend processing. We use the same configuration as the open-source NBSS.

⁵<https://github.com/Audio-WestlakeU/NBSS>

Table 3: The results of our system on Eval1(CER(%)) and Eval2(cpCER(%)), using AEC + IVA and Spatialnet frontend with different ASR Models. The training data of AISHELL-5 includes far-field and near-field data with a total amount of 190 hours. (* indicates the model is fine-tuned with the AISHELL-5 training Set.)

Model Type	Model	Training Data	Train/finetune Epochs	Model Size	Eval1	Eval2	
						AEC + IVA	Spatialnet
ASR Models	Transformer	190 hours	100	29.89 M	31.75	77.32	58.23
	Conformer		100	45.73 M	26.89	69.55	53.78
	E-Branchformer		100	47.13 M	26.05	71.04	51.52
	Zipformer-Small		100	30.22 M	31.22	74.86	54.34
Open-Source Models	Paraformer [10]	60,000 hours	-	220 M	20.16	74.04	48.67
	Paraformer-Finetuned*	190 hours	10	220 M	16.65	66.68	47.18
	Whisper-Small [11]	680,000 hours	-	244 M	50.69	79.49	65.72
	SenseVoice-Small [12]	Over 400,000 hours	-	234 M	24.63	75.58	50.64
	Qwen2-Audio [13]	520,000 hours	-	7B	29.92	76.24	54.48

The model is trained using the complex mean squared error (cc.mse) loss, with a loss scale factor of 100. To stabilize gradient computation, we use gradient clipping with a norm-based clipping algorithm and a predefined clipping threshold. The Adam optimizer is employed with a learning rate of 1e-3, and the learning rate scheduler follows an exponential decay strategy.

The training objective is to minimize the loss on the validation set. We monitor the model’s performance using evaluation metrics such as Signal-to-Distortion Ratio (SDR), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), and Perceptual Evaluation of Speech Quality (PESQ). The window length and hop size for STFT are 256 and 128, respectively. The real and imaginary parts of the 4-channel input audio are concatenated to form an 8-channel input feature. The near-end speech of the speaker at each seat is used as the training target. Additionally, 40 hours of real-recorded noise is mixed with the training data as an additional noise source.

Finally, we use a model trained for 100 epochs to perform frontend processing on Eval2.

3.2. Automatic Speech Recognition

To compare the performance of different ASR models on the test set, we evaluate several models, including Transformer [15], Conformer [16], E-Branchformer [17], and Zipformer [18], as baselines. The ASR models are trained using Wenet [19], with Zipformer training and evaluation based on Icefall, though we only share the decoding results in the baseline system without providing the related code.

We use the AISHELL-5 training data for ASR training. For near-field data, we split the audio segments based on timestamp information. For far-field data, we first apply AEC and IVA for echo cancellation and speaker separation, then use VAD to segment the original long audio into shorter segments. This results in single-channel, single-speaker audio suitable for training. The final training data consists of near-field and high-quality far-field single-channel, single-speaker audio, totaling approximately 190 hours. During Zipformer training, the data is formatted to be compatible with the Icefall⁶.

We use 80-dimensional fbank features as input, with utterance-level mean-variance normalization. The Adam optimizer is used, with a maximum of 100 training epochs. A

warm-up and decay learning rate scheduler is employed, with a peak learning rate of 0.002, warm-up steps set to 25,000, and a batch size of 18. Finally, we average the last 30 epochs of all trained models and perform ASR inference on the averaged model, using attention rescoring for decoding with a beam size of 10.

3.3. Evaluation and results

We present the baseline system results on the test set, as shown in Table 3. The table displays results on the Eval1 task and cpCER results for the same ASR models on Eval2 after applying two different front-end processing methods. On Eval1, E-Branchformer achieved the best performance at 26.05% with the same number of training epochs. On Eval2, ASR recognition results after Spatialnet processing showed a significant improvement, with about a 20% gain, and E-Branchformer achieved 51.52% cpCER. We also report results from some open-source models on the test set. Paraformer performed well among the open-source models, reaching 20.16% on Eval1 and 48.67% on Eval2 with Spatialnet as the front-end. After fine-tuning for 20 epochs with the same baseline training data, Paraformer improved these results to 16.65% and 47.18% with Spatialnet front-end processing.

4. Conclusions

This paper introduces AISHELL-5, currently the largest open-source multi-channel, multi-speaker free-talk in-car speech dataset, tailored for two key issues in current in-car speech processing techniques: the complex acoustic environment within the car and the frequent occurrence of overlapping speech from driver and passengers. AISHELL-5 is suitable for speech separation, speech enhancement, noise reduction, and automatic speech recognition (ASR) tasks targeting in-car scenarios. All speech data is recorded from the real acoustic environment inside an electric vehicle, covering 60 different driving scenarios. It also includes 40-hour real-recorded noise data, which supports the in-car speech data simulation. Moreover, we provide an open-access, reproducible baseline system based on this dataset. We firmly believe that AISHELL-5, as the first open-source in-car multi-channel speech dataset, can offer a certain amount of support for this increasingly important ASR application scenario, advancing human-vehicle interaction towards greater accuracy and convenience.

⁶<https://github.com/k2-fsa/icefall>

5. References

- [1] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu *et al.*, “AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario,” *arXiv preprint arXiv:2104.03603*, 2021.
- [2] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, “M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [3] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, “Librimix: An open-source dataset for generalizable speech separation,” *arXiv preprint arXiv:2005.11262*, 2020.
- [4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [5] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, “Wham!: Extending speech separation to noisy environments,” *arXiv preprint arXiv:1907.01160*, 2019.
- [6] M. Ravanelli, P. Svaizer, and M. Omologo, “Realistic multi-microphone data simulation for distant speech recognition,” *arXiv preprint arXiv:1711.09470*, 2017.
- [7] C.-C. Wang, L.-W. Chen, H.-S. Lee, B. Chen, and H.-M. Wang, “Effective noise-aware data simulation for domain-adaptive speech enhancement leveraging dynamic stochastic perturbation,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024.
- [8] A. Zhang, F. Yu, K. Huang, L. Xie, L. Wang, E. S. Chng, H. Bu, B. Zhang, W. Chen, and X. Xu, “The iscslp 2022 intelligent cockpit speech recognition challenge (icsrc): Dataset, tracks, baseline and results,” in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2022, pp. 507–511.
- [9] H. Wang, P. Guo, Y. Li, A. Zhang, J. Sun, L. Xie, W. Chen, P. Zhou, H. Bu, X. Xu, B. Zhang, Z. Chen, J. Wu, L. Wang, E. S. Chng, and S. Li, “ICMC-ASR: the ICASSP 2024 in-car multi-channel automatic speech recognition challenge,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 - Workshops, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024.
- [10] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, “Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition,” in *INTERSPEECH*, 2022.
- [11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.
- [12] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu, S. Ji, Y. Li, Z. Li, H. Lu, H. Luo, X. Lv, B. Ma, Z. Ma, C. Ni, C. Song, J. Shi, X. Shi, H. Wang, W. Wang, Y. Wang, Z. Xiao, Z. Yan, Y. Yang, B. Zhang, Q. Zhang, S. Zhang, N. Zhao, and S. Zheng, “Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms,” 2024.
- [13] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, “Qwen2-audio technical report,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.10759>
- [14] T. Kim, T. Eltoft, and T. Lee, “Independent vector analysis: An extension of ICA to multivariate components,” in *Independent Component Analysis and Blind Signal Separation, 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings*, J. P. Rosca, D. Erdogmus, J. C. Principe, and S. Haykin, Eds. Springer, 2006.
- [15] L. Dong, S. Xu, and B. Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018.
- [16] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020.
- [17] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, “E-branchformer: Branchformer with enhanced merging for speech recognition,” in *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*. IEEE, 2022, pp. 84–91. [Online]. Available: <https://doi.org/10.1109/SLT54892.2023.10022656>
- [18] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, “Zipformer: A faster and better encoder for automatic speech recognition,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [19] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, “Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” in *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlíček, Eds. ISCA, 2021.