



# Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational(RAMC) Speech Dataset

Zehui Yang<sup>1,2,\*</sup>, Yifan Chen<sup>1,2,\*</sup>, Lei Luo<sup>3,†</sup>, Runyan Yang<sup>1,2</sup>, Lingxuan Ye<sup>1,2</sup>, Gaofeng Cheng<sup>1</sup>, Ji Xu<sup>1</sup>, Yaohui Jin<sup>4</sup>, Qingqing Zhang<sup>3</sup>, Pengyuan Zhang<sup>1,2</sup>, Lei Xie<sup>5</sup>, Yonghong Yan<sup>1,2</sup>

<sup>1</sup>Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, China

<sup>2</sup>University of Chinese Academy of Sciences, China

<sup>3</sup>Magic Data Technology Co., Ltd., China

<sup>4</sup>MoE Key Lab of Artificial Intelligence and AI Institute, Shanghai Jiao Tong University, China

<sup>5</sup>Northwestern Polytechnical University, China

chenggaofeng@hcccl.ioa.ac.cn, luolei@magicdatatech.com, zhangqingqing@magicdatatech.com

## Abstract

This paper introduces a high-quality rich annotated Mandarin conversational (RAMC) speech dataset called MagicData-RAMC. The MagicData-RAMC corpus contains 180 hours of conversational speech data recorded from native speakers of Mandarin Chinese over mobile phones with a sampling rate of 16 kHz. The dialogs in MagicData-RAMC are classified into 15 diversified domains and tagged with topic labels, ranging from science and technology to ordinary life. Accurate transcription and precise speaker voice activity timestamps are manually labeled for each sample. Speakers' detailed information is also provided. As a Mandarin speech dataset designed for dialog scenarios with high quality and rich annotations, MagicData-RAMC enriches the data diversity in the Mandarin speech community and allows extensive research on a series of speech-related tasks, including automatic speech recognition, speaker diarization, topic detection, keyword search, text-to-speech, etc. We also conduct several relevant tasks and provide experimental results to help evaluate the dataset.

**Index Terms:** Mandarin corpus, dialog scenario, speech recognition, speaker diarization, keyword search

## 1. Introduction

As an essential resource for the speech research community, speech data is especially required for spoken language processing technologies based on statistical models [1, 2, 3, 4, 5, 6] which are data-driven. Over the past few decades, a growing number of well-curated and freely available speech corpora with various accents, speaking styles, recording environments, channels, scales, etc., have been developed and released, trying to cover the infinite diversity of human speech.

There have been many datasets created for general speech recognition tasks. The Wall Street Journal corpus [7] is a classic dataset including 80 hours of narrated news articles. LibriSpeech [8], one of the most widely used English corpora, is a reading speech dataset based on LibriVox's audiobooks with steady speed and consistent tone. TED-LIUM 3 [9] consists of 452 hours of audio collected from TED talks. THCHS30 [10] is a toy Chinese speech database accessible to academic users. AISHELL-1 [11] is a commonly used Mandarin dataset with speeches in reading style. These datasets are relatively formal in

language and recorded in a single-speaker manner; hence lack natural speech characteristics and interaction between speakers. WenetSpeech [12] is a recently released Mandarin corpus containing more than 10000 hours of speech collected from YouTube and podcasts, adopted with optical character recognition (OCR) and automatic speech recognition (ASR) methods to generate the transcriptions, respectively. GigaSpeech [13] is a large-scale English corpus with audio collected from audiobooks, podcasts, and YouTube. Its transcripts are downloaded from various sources and applied with standard text normalization, and the word error rate of the training subset is around 4%.

To expand application areas and conduct more tasks such as speaker diarization, some corpora are collected among multi-speakers in more diverse scenarios. Meeting speech datasets, such as ISCI [14], CHIL [15] and AMI [16], are composed of audio recorded during the conferences held in academic laboratories. AISHELL-4 [17] is a 120-hour Mandarin speech dataset collected in conference scenarios. The number of attendees in each session is between four and eight. The indoor conversation corpus CHiME-6 [18] consists of unsegmented recordings of twenty separate dinner-party conversations among four friends captured by Kinect devices. It has a more casual and natural speaking style yet relatively low recording quality. For dialog scenario, Switchboard [19] and Fisher [20] are classic datasets of English telephony conversations with similar settings and different scales. HKUST is a Mandarin conversational corpus [21] made up of spontaneous telephone conversations. Audios in these three telephony conversational datasets are recorded with a sampling rate of 8 kHz, which is incompatible with the demand of some speech processing systems nowadays.

Although various corpora, including the ones mentioned above, have been introduced, most of them are not specifically designed for dialog scenarios. Meanwhile, the utilization of the few existing dialog corpora is limited by collection setups such as sampling rate. To the best of our knowledge, for Mandarin Chinese, there is no public-available dialog speech dataset adequate for the current requirement of high quality. With the boom in popularity of voice-driven interfaces to devices recently, some works [22, 23] concerned with communication scenes have been conducted. However, exploring speech processing techniques in dialog scenarios is still challenging. Deficient and mismatched data is one of the limitations that restrict the investigation of communication scenarios due to the attributes of dialog speech.

\* Equal contribution.

† Corresponding author.

In order to enrich the diversity of the Mandarin database and alleviate the scarcity of data specialized in dialog scenarios, we develop a high-quality rich annotated Mandarin conversational corpus in this work and refer to it as MagicData-RAMC<sup>1</sup>. It contains 180 hours of dialog speech recorded over mobile phones with a sampling rate of 16 kHz. Accurate transcriptions are manually labeled and proofed for each sample. Precise voice activity timestamps of each speaker, each sample’s topic label, and speakers’ demographic information are also provided, allowing further research on different tasks. The dialogs in MagicData-RAMC aim to reflect our realistic communication in the real world and cover various topics about our daily lives. On the one hand, there are rich and complicated speech characteristics in its samples, including colloquial expressions, hesitations, repetitions, nonsense syllables, and other speech disfluencies. On the other hand, during long dialogs in the dataset, people respond to each other in a flexible manner and continue the dialog with coherent questions and opinions instead of stiffly answering each other’s questions. Therefore, history and current utterances are closely related, and a consistent topic runs through the conversation by the contextual flow. We believe that MagicData-RAMC can facilitate the research on a series of tasks relevant to multi-turn dialogs, and we conduct several of them to evaluate the dataset.

The rest of this paper is organized as follows. We introduce the construction process of MagicData-RAMC and present its structure and details in Section 2. Then we describe the baseline systems built on popular toolkits for speech recognition, speaker diarization (SD), and keyword search (KWS) tasks and provide experiment setup and results in Section 3. Finally, we conclude the entire paper in Section 4.

## 2. Dataset Description

MagicData-RAMC comprises dialog speech data, corresponding transcriptions, voice activity timestamps, and speakers’ demographic information. It contains 351 multi-turn Mandarin Chinese dialogs, which amount to about 180 hours. The speech data is carefully annotated and manually proofed.

### 2.1. Collection setup

The dataset is collected indoors. The domestic environments are small rooms under 20 square meters in area, and the reverberation time (RT60) is less than 0.4 seconds. The environments are relatively quiet during recording, with ambient noise level lower than 40 dB.

The audios are recorded via an application developed by Magic Data Technology Co., Ltd.<sup>2</sup> over mainstream smartphones, including Android phones and iPhones. The ratio of Android phones to iPhones is around 1 : 1. All recording devices work at 16 kHz, 16-bit to guarantee high recording quality.

All speech data are manually labeled using the platform built by Magic Data Technology Co., Ltd. Sound segments without semantic information during the conversations, including laughter, music, and other noise events, are annotated with specific symbols. Phenomena common in spontaneous communications, such as colloquial expressions, partial words, repetitions, and other speech disfluencies, are recorded and fully tran-

<sup>1</sup><https://www.magicdatatech.com/datasets/mdt2021s003-1647827542>. The partition of the dataset and the keyword list used in the subsequent KWS experiment are provided together in the URL.

<sup>2</sup><https://www.magicdatatech.com>

scribed. Each recording is labeled by one annotator. Punctuation is carefully checked to ensure accuracy. We also segment the dialog speech and provide precise voice activity timestamps of each speaker. The transcriptions are proofed by professional inspectors to ensure the labeling and segmentation quality.

### 2.2. Speaker information

There are a total of 663 speakers involved in the recording, of which 295 are female and 368 are male. Each segment is labeled with the corresponding speaker-id. All participants are native and fluent Mandarin Chinese speakers with slight variations of accent and participants in each group are acquaintances. The accent region is roughly balanced, with 334 Northern Chinese speakers and 329 Southern Chinese speakers. The detailed distribution based on the birthplaces of the speakers is shown in Table 1. Besides, each speaker participates in up to three conversations.

Table 1: Region distribution of speakers

| Region         | Province  | #Speaker | Total |
|----------------|-----------|----------|-------|
| Northern China | Shandong  | 94       | 334   |
|                | Shanxi    | 228      |       |
|                | Beijing   | 12       |       |
| Southern China | Guangdong | 12       | 329   |
|                | Hunan     | 249      |       |
|                | Jiangsu   | 10       |       |
|                | Sichuan   | 58       |       |

### 2.3. Dataset Partition

We divide the corpus into 150 hours training set, 10 hours development set, and 20 hours test set, containing 289, 19, and 43 conversations, respectively. The partition of the speech data is provided in TSV format. There are 556, 38, and 86 speakers split into three subsets. The gender and region distribution is roughly proportional to the entire dataset. Table 2 provides a summary of the partition of the corpus.

Table 2: Corpus partition

|              | Training | Development | Test  |
|--------------|----------|-------------|-------|
| Duration (h) | 149.65   | 9.89        | 20.64 |
| #Sample      | 289      | 19          | 43    |
| #Speaker     | 556      | 38          | 86    |
| #Male        | 307      | 23          | 49    |
| #Female      | 249      | 15          | 37    |
| #Northern    | 271      | 20          | 52    |
| #Southern    | 285      | 18          | 34    |

### 2.4. Utterance statistics

The whole dataset is composed of 351 conversations with 219325 segments in total. Each conversation is of 30.80 minutes duration and segmented to about 625 speech segments on average. The start and end times of all segments are specified to within a few milliseconds. We also count the length of segments and the number of tokens per segment to give a simple and intuitive view of the corpus. The statistics are presented in Table 3.

Table 3: *Statistics of speech information*

| Statistical Criterion | Max   | Min   | Average |
|-----------------------|-------|-------|---------|
| Sample Duration (min) | 33.02 | 14.06 | 30.80   |
| #Segments Per Sample  | 1215  | 231   | 624.86  |
| Segment Duration (s)  | 14.91 | 0.09  | 2.54    |
| #Tokens Per Segment   | 89    | 1     | 13.58   |
| #Segments Per Speaker | 1155  | 46    | 304.55  |

The dialogs in MagicData-RAMC aim to reflect the natural communication way in the real world and cover a broad range of topics closely relevant to our daily life. During the multi-turn conversations in the dataset, people respond flexibly to each other and continue the dialog with relevant questions and items instead of replying and waiting for the following questions rigidly. Therefore, every sample is a coherent and compact conversation centered around one theme, with history utterances and current utterance closely related. Higher-level information is maintained by contextual dialog flow across multiple sentences.

Topics are freely chosen by the participants and we classify the dialogs in MagicData-RAMC into 15 diversified domains, ranging from science and technology to ordinary life. The diversity of topics and the consistency in one dialog are beneficial to the development of open-domain spoken dialog systems. We summarize the statistics of the categories in Table 4.

Table 4: *The distribution over topics*

| Topic                    | #Sample | Duration (h) |
|--------------------------|---------|--------------|
| Humanities               | 22      | 11.46        |
| Entertainment            | 1       | 0.52         |
| Sports                   | 32      | 16.62        |
| Military                 | 4       | 1.98         |
| Finance & Economy        | 5       | 2.49         |
| Religion                 | 1       | 0.52         |
| Family Life              | 6       | 1.48         |
| Politics & Law           | 4       | 2.07         |
| Education & Health       | 53      | 27.30        |
| Digital Devices          | 39      | 20.14        |
| Climate & Environment    | 13      | 6.62         |
| Science & Technology     | 11      | 5.66         |
| Professional Development | 35      | 18.26        |
| Art                      | 84      | 43.87        |
| Ordinary Life            | 41      | 21.19        |

## 2.5. Comparison with other datasets

The main properties of several public-available manually labeled speech datasets are summarized in Table 5 for a brief comparison. Compared to other corpora, MagicData-RAMC is the most suitable conversational speech dataset for research on Mandarin dialog speech.

The quality of MagicData-RAMC is comparable to open Mandarin speech datasets collected in a single-speaker manner. At the same time, MagicData-RAMC is closer to applications in the real world due to its natural spontaneous speaking style and realistic speech characteristics as a corpus specifically designed for dialog scenarios. MagicData-RAMC is also different from conference datasets in the number of attendees, though they are all developed in speech interactive scenarios. Compared to classic telephony conversational datasets with a sampling rate of 8

KHz, the audios recorded via mobile phones and sampled at 16 kHz in MagicData-RAMC are more compatible with most of the current speech processing systems’ higher requirements for recording quality. What’s more, rich annotations and detailed information in MagicData-RAMC allow further exploration of various speech processing tasks.

## 3. Experiments

In this section, we build baseline systems for ASR, SD, and KWS tasks and present the experimental setup and results, respectively, to evaluate this dataset.

### 3.1. Automatic Speech Recognition

We use a Conformer-based end-to-end (E2E) model implemented by ESPnet2 toolkit [24] to conduct speech recognition. The Conformer model is composed of a Conformer encoder proposed in [25] and a Transformer decoder. We adopt 14 Conformer blocks in the encoder and 6 Transformer blocks in the decoder. Connectionist temporal classification (CTC) [26] is employed on top of the encoder as an auxiliary task to perform joint CTC-attention (CTC/Att) training and decoding [27] within the multi-task learning framework. During the beam search decoding, we set the beam size to 10.

We compute 80-dimensional logarithmic filterbank features from the raw waveform and utilize utterance-wise mean variance normalization. The frame length is 25 ms with a stride of 10 ms. SpecAugment [28] is applied with 2 frequency masks and 2 time masks for data augmentation. The maximum widths of each frequency mask and time mask are  $F = 30$  and  $T = 40$  respectively. The input sequence is sub-sampled through a 2D convolutional layer by a factor of 4. The inner dimension of position-wise feed-forward networks in both encoder and decoder is 2048. We apply dropout in the encoder layer with a rate of 0.1 and set the label smoothing weight to 0.1 for regularization. The multi-head attention layer contains 8 heads with 256 attention transformation dimensions. CTC weight used for multi-task training is 0.3. We train the model using the Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$ ,  $\epsilon = 10^{-9}$ , gradient clipping norm 5 and no weight decay. Noam learning rate scheduler is set to 25000 warm-up steps. The maximum trainable epoch is 100. The final model is averaged by the last 5 checkpoints.

The training set is made up of two parts: the 150 hours training set of MagicData-RAMC and the 755 hours MAGIC-DATA Mandarin Chinese Read Speech Corpus (MAGICDATA-READ<sup>3</sup>). The two sets are combined to compose over 900 hours of data for training. The experimental result is shown in terms of character error rate (CER) in Table 6.

### 3.2. Speaker Diarization

For SD task, our baseline system consists of three components: speaker activity detection (SAD), speaker embedding extractor and clustering.

Following Variational Bayes HMM x-vectors (VBx) [29] experiment setting, Kaldi-based SAD module [30] is used for detecting speech activity. We adopt ResNet [31] trained on VoxCeleb Dataset [32] (openslr-49<sup>4</sup>), CN-Celeb Corpus [33] (openslr-82<sup>5</sup>) and the split training set of MagicData-RAMC to obtain the speaker embedding extractor.

<sup>3</sup><http://www.openslr.org/68>

<sup>4</sup><http://www.openslr.org/49>

<sup>5</sup><http://www.openslr.org/82>

Table 5: Comparison with several manually labelled datasets

| Language | Datasets              | Duration (hours) | Scenario            | Topic label | #Spkrs | Device              | Sampling Rate (kHz) |
|----------|-----------------------|------------------|---------------------|-------------|--------|---------------------|---------------------|
| EN       | LibriSpeech[8]        | 1000             | Reading             |             | 2484   | mic                 | 16                  |
|          | AMI[16]               | 100              | Conference          | ✓           | 200    | mic array, headsets | 16                  |
|          | Switchboard[19]       | 317              | Conversation        |             | 500    | telephone           | 8                   |
| CN       | THCHS-30[10]          | 33               | Reading             |             | 40     | mic                 | 16                  |
|          | MAGICDATA-READ        | 755              | Reading             | ✓           | 1080   | mobile phone, mic   | 16                  |
|          | HKUST[21]             | 200              | Conversation        | ✓           | 2412   | telephone           | 8                   |
|          | <b>MagicData-RAMC</b> | 180              | <b>Conversation</b> | ✓           | 663    | mobile phone        | <b>16</b>           |

Table 6: ASR Results (CER%) on dev and test set

| Methods       | Dev  | Test |
|---------------|------|------|
| LAS-Conformer | 16.5 | 19.1 |

For training details, the SAD module utilizes a 40-dimensional Mel frequency cepstral coefficients (MFCC) with 25 ms frame length and 10 ms stride as input features to detect the speech activity. ResNet-101 with two fully connected layers is employed to conduct speaker classification task with 64-dimensional filterbank features extracted every 10 ms with 25 ms window, and additive margin softmax [34] is used to get a more distinct decision boundary. The raw waveform is split every 4s (400 dimensions) to form ResNet input. We train the speaker embedding network using stochastic gradient descent (SGD) optimizer with a 0.9 momentum factor and 0.0001 L2 regularization factor.

Besides, 256-dimension embeddings are conducted dimensionality reduction using probabilistic linear discriminant analysis (PLDA) [35] to 128-dimension. Embeddings are extracted on SAD result every 240 ms, and the chunk length is set to 1.5s. For the clustering part, we use Variational Bayes HMM [29] on this task. An agglomerative hierarchical clustering algorithm with VBx is conducted to get the clustering result. In the VBx, the acoustic scaling factor  $Fa$  is set to 0.3, and the speaker regularization coefficient is set to 17. The probability of not switching speakers between frames is 0.99. We present the experimental result in Table 7.

Table 7: Speaker diarization results of VBx system

| Method | Subset | DER         |          | JER   |
|--------|--------|-------------|----------|-------|
|        |        | collar 0.25 | collar 0 |       |
| VBx    | Dev    | 5.57        | 17.48    | 45.73 |
|        | Test   | 7.96        | 19.90    | 47.49 |

### 3.3. Keyword Search

We carry out the KWS task following the DTA Att-E2E-KWS approach proposed in [36] relying on attention-based E2E ASR framework and frame-synchronous phoneme alignments. The KWS system is based on our Conformer-based E2E ASR system described in Sec 3.1. We adopt the dynamic time align-

ment (DTA) algorithm to connect a frame-wise phoneme classifier’s output posteriors and the label-wise ASR result candidates for generating accurate time alignments and reliable confidence scores of recognized words. Keyword occurrences are retrieved within the N-best hypotheses generated in the joint CTC/Att decoding process.

The keyword list is built by picking 200 words from the dev set and provided together with the dataset. In the DTA Att-E2E-KWS system, the frame-wise phoneme classifier of the KWS system shares 12 Conformer blocks with the E2E ASR encoder while retaining the top 2 Conformer blocks unshared. The classifier outputs posteriors of 61 phonemes, including silence and noise. The KWS system is optimized following the setup in Sec 3.1. During the inference stage, we retrieve keywords within ASR 2-best hypotheses. During KWS scoring, a predicted keyword occurrence is considered correct when there is a 50% time overlap at least between the predicted occurrence and a reference occurrence of the same keyword [36]. The results are shown in Table 8.

Table 8: Results on dev and test set for the Conformer-based DTA Att-E2E-KWS system

| Subset | Precision rate | Recall rate | F1 score |
|--------|----------------|-------------|----------|
| Dev    | 0.8698         | 0.8957      | 0.8826   |
| Test   | 0.8587         | 0.8879      | 0.8731   |

## 4. Conclusions

In this paper, we release MagicData-RAMC, a high-quality rich annotated Mandarin conversational speech dataset. It is a freely available high-quality Mandarin corpus specially created for dialog scenarios with rich annotations, including precise voice activity timestamps of each speaker, topic labels, etc. We introduce the collection, structure, and detailed analysis of the dataset. We also conduct speech recognition, speaker diarization, and keyword search tasks based on popular speech toolkits to provide examples of the wide utilization of the corpus. We hope that the MagicData-RAMC speech dataset can enrich the diversity of the speech database and facilitate the applications of various speech-related research.

## 5. References

- [1] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

- [2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4960–4964.
- [3] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [4] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," *Proc. ICASSP*, pp. 5884–5888, 2018.
- [5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [6] H. Miao, G. Cheng, P. Zhang, and Y. Yan, "Online hybrid ctc/attention end-to-end automatic speech recognition architecture," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1452–1465, 2020.
- [7] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. 2nd International Conference on Spoken Language Processing (ICSLP 1992)*, 1992, pp. 899–902.
- [8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.
- [9] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *International conference on speech and computer*. Springer, 2018, pp. 198–208.
- [10] D. Wang and X. Zhang, "Thchs-30 : A free chinese speech corpus," *ArXiv*, vol. abs/1512.01882, 2015.
- [11] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.
- [12] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," *arXiv preprint arXiv:2110.03370*, 2021.
- [13] G. Chen, S. Chai, G.-B. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10, 000 hours of transcribed audio," in *Interspeech*, 2021.
- [14] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *Proc. ICASSP*, vol. 1. IEEE, 2003, pp. I–I.
- [15] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia *et al.*, "The chil audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language resources and evaluation*, vol. 41, no. 3, pp. 389–407, 2007.
- [16] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2007, pp. 238–247.
- [17] 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.
- [18] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.
- [19] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1, 1992, pp. 517–520 vol.1.
- [20] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004.
- [21] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "Hkust/mts: A very large scale mandarin telephone speech corpus," in *International Symposium on Chinese Spoken Language Processing*. Springer, 2006, pp. 724–735.
- [22] S. Kim and F. Metze, "Dialog-context aware end-to-end speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 434–440.
- [23] K. Deng, G. Cheng, H. Miao, P. Zhang, and Y. Yan, "History utterance embedding transformer lm for speech recognition," in *Proc. ICASSP*, 2021, pp. 5914–5918.
- [24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplín, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESNet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.
- [25] A. Gulati, J. Qin, C.-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," *ArXiv*, vol. abs/2005.08100, 2020.
- [26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. Association for Computing Machinery, 2006, p. 369–376.
- [27] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, 2017, pp. 4835–4839.
- [28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.
- [29] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbX) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101254, 2022.
- [30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [32] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *CoRR*, vol. abs/1706.08612, 2017.
- [33] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.
- [34] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.
- [35] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.
- [36] R. Yang, G. Cheng, H. Miao, T. Li, P. Zhang, and Y. Yan, "Keyword search using attention-based end-to-end asr and frame-synchronous phoneme alignments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3202–3215, 2021.