SeniorTalk: A Chinese Conversation Dataset with Rich Annotations for Super-Aged Seniors

Yang Chen[1+], Hui Wang[1+], Shiyao Wang[1], Junyang Chen[1],Jiabei He[1], Jiaming Zhou[1], Xi Yang[2], Yequan Wang[2], Yonghua Lin[2], Yong Qin[1*]

[1]College of Computer Science, Nankai University,
[2]Beijing Academy of Artificial Intelligence, Beijing, China ,
Correspondence: 2120230617@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

While voice technologies increasingly serve aging populations, current systems exhibit significant performance gaps due to inadequate training data capturing elderly-specific vocal characteristics like presbyphonia and dialectal variations. The limited data available on super-aged individuals in existing elderly speech datasets, coupled with overly simple recording styles and annotation dimensions, exacerbates this issue. To address the critical scarcity of speech data from individuals aged 75 and above, we introduce SeniorTalk, a carefully annotated Chinese spoken dialogue dataset. This dataset contains 55.53 hours of speech from 101 natural conversations involving 202 participants, ensuring a strategic balance across gender, region, and age. Through detailed annotation across multiple dimensions, it can support a wide range of speech tasks. We perform extensive experiments on speaker verification, speaker diarization, speech recognition, and speech editing tasks, offering crucial insights for the development of speech technologies targeting this age group.

1 Introduction

The rapid global aging population presents both significant challenges and opportunities in the development of technologies specifically designed for older adults, particularly those aged 75 and over (Scuteri and Nilsson, 2024; Chumuang et al., 2024). As the number of people in this ultra-high-age group continues to grow, it becomes increasingly important to enhance the accessibility and inclusion of speech-based technologies for this demographic (Young and Mihailidis, 2010; Portet et al., 2013). However, many state-of-the-art speech systems struggle to perform effectively within the elderly population, exhibiting biases on elderly vocal patterns. (Kulkarni et al., 2024; Feng et al., 2021). For instance, existing speech recognition systems often show poor performance with older users, influenced by factors such as speech deterioration, hearing loss, health issues (Fraser et al., 2015), and the diversity of speech patterns among older adults (Geng et al., 2022; Hu et al., 2024). A key underlying reason for this challenge is the lack of datasets specifically addressing the unique needs of ultra-high-age individuals (Hu et al., 2024), which hampers the development of robust foundation models and the application of tailored solutions.

Although existing studies have made some efforts to collect geriatric speech data (Sekerina et al., 2024; Fukuda et al., 2020), significant limitations remain. First, current speech corpora predominantly focus on younger adults or healthy senior populations, with samples that consist mainly of scripted speech and exhibit standardized accents. For example, both the AISHELL-ASR0060 and MECSD datasets (Wang et al., 2019) adopt an unusually low age threshold of 55 years for defining elderly participants. Similarly, the S-JNAS corpus of elderly Japanese speech reports a mean speaker age of 67.6 years (Fukuda et al., 2020). These fall below the World Health Organization's (WHO) geriatric classification designating 75+ years as the late elderly phase which is a period associated with progressive age-related decline in physiological function (Orimo et al., 2006).

Second, the current collection paradigms and annotation methods of existing datasets further limit their practical applicability. A large portion of speech resources for the elderly primarily focuses on reading style (Wang et al., 2019; Iribe et al., 2015), which do not reflect the everyday communication scenarios that elderly people encounter in real life. Furthermore, many corpora are tailored for narrowly defined tasks such as automatic speech recognition (ASR) or pathology detection

* Yong Qin is the corresponding author.
First Author and Second Author contribute equally to this work.

Dataset	Age	Style	Annotation Features				
			Speaker Region	Transcript	Timestamp	Accent Level	Sound Event
ASR0060	55+	Reading	North/South/Other	✓	N/A	×	×
MECSD	55+	Reading	×	✓	N/A	×	✓
SeniorTalk	75+	Conversation	Provincial	✓	✓	✓	✓

Table 1: Comparison of Chinese elderly speech datasets, including AISHELL-ASR0060 (marked in the table as ASR0060), MECSD, and SeniorTalk, across annotation features.

(Wang et al., 2019), thereby preventing comprehensive characterization of age-related vocal variations. These datasets also lack key features such as speaker diarization or dialect labels, which restrict their ability to support a wider array of use cases and hinder their robustness in addressing the diverse challenges of speech processing.

To address these limitations, we introduce **SeniorTalk**, a Mandarin dataset consisting of spontaneous conversations among individuals aged 75 and older. As shown in Table 1, this dataset comprises natural conversational recordings from 202 native Chinese speakers, representing a rich diversity of regional, age, and gender demographics, and captured in authentic, real-world interaction settings. It effectively addresses the current gaps and limitations in datasets focused on elderly populations, particularly the underrepresentation of super-aged seniors and the lack of diversity in recording styles and annotation dimensions. Moreover, we conduct extensive experiments across various speech tasks, providing a benchmark specifically for the elderly population. By open-sourcing this corpus along with fine-grained metadata, we aim to bridge the vocal age gap and promote the development of equitable voice technologies for aging societies. SeniorTalk makes three key contributions to the field:

- We recruit 202 super-elderly speakers from 16 provinces in China, ensuring balance across gender, age, and geography. This resulted in SeniorTalk, a dataset with 55.53 hours of data from 101 conversations.
- We provide detailed, multi-dimensional annotations for the dataset, encompassing speaker information, transcriptions, timestamps, and more. These annotations enable comprehensive speech signal analysis and support a wide range of speech-related tasks for the elderly.
- We conduct a comprehensive series of experiments using the dataset, covering tasks such as speaker verification, speaker diarization, speech recognition, and others. This establishes a solid benchmark for evaluating models across various speech-related tasks.

2 Related Work

Several corpora have been developed to address the specific acoustic characteristics of elderly speech. Early work in this area includes the Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS), a foundational resource for Japanese elderly speech research. Subsequent research has broadened the scope of investigation, with efforts such as the Carolinas Conversations Collection (CCC) (Pope and Davis, 2011) focusing on multiethnic elderly speakers with chronic conditions, offering valuable insights into how sociolinguistic factors and health status influence speech production. The AD80 and ERES38 corpora (Aman et al., 2013) advance French elderly speech analysis through distress detection benchmarks for ambient assisted living systems. Further corpus development has continued with the creation of specialized datasets, including a corpus of 100 elderly Japanese speakers designed to enhance human-robot interaction in elder care (Iribe et al., 2015). The elderLUCID project (Hazan et al., 2017) examines the complexities of speech communication in older adults, considering the interplay of hearing loss, phonation, articulation, and cognitive factors.

More recently, researchers have focused on specific languages and conditions, as exemplified by the Mandarin Speech Database for Early Dementia Detection (MECSD) (Wang et al., 2019) and the AISHELL-ASR0060 database for elderly Mandarin speech. Improvements to existing corpora, such as S-JNAS, have also been explored, including the creation of acoustic models specifically designed for "super-elderly" speakers (Fukuda et al., 2020). The Elderly Multimodal Interpersonal Con-

https://www.aishelltech.com/General_Datasets

Corpus	Language	Age	Style	# Spks.	Dur.(hrs)	Year	Avail.
BraPoRus (Sekerina et al., 2024)	Brazilian Portuguese-Russian	59-98	Monologue, interview, ...	1,500	170	2024	N
EARS (Fukuda et al., 2023)	Japanese	70-99	Reading	123	13.4	2023	N
Improving S-JNAS (Fukuda et al., 2020)	Japanese	65-99	Reading	221	31.7	2020	Y
AISHELL-ASR0060	Mandarin	55+	Reading	503	793	2019	Y
MECSD (Wang et al., 2019)	Mandarin	55-85	Reading	85	110	2019	P
elderLUCID (Hazan et al., 2017)	English	19-84	Reading	83	N	2017	N
Develop S-JNAS (Iribe et al., 2015)	Japanese	60-98	Reading	100	9.2	2015	Y
ERES38 (Aman et al., 2013)	French	68-98	Interview	22	17	2013	N
AD80 (Aman et al., 2013)	French	62-94	-	43	4.7	2013	N
CCC (Pope and Davis, 2011)	English	65+	Interview	600+	800+	2011	Y
S-JNAS	Japanese	60-90	Reading	301	-	2007	Y
E-MIC	Korean	65-85	conversation	100	3	-	Y

Table 2: Summary of related elderly speech datasets. Key characteristics include the age ranges of speakers (Age), the number of speakers (# Spks.), publication year (Year), and availability status (Avail.), where 'P' indicates partial availability.

versation (E-MIC) dataset expands the scope of analysis by incorporating multimodal data, including video and audio, to study turn-taking in elderly conversations. Further work, such as EARS (Fukuda et al., 2023), has continued to refine acoustic modeling techniques for super-elderly Japanese speakers, while the BraPoRus corpus (Sekerina et al., 2024) highlights the importance of preserving heritage languages and the challenges of remote data collection during the COVID-19 pandemic.

3 Dataset description

3.1 Dataset overview

SeniorTalk is designed specifically for the ultra-elderly population aged 75 and above, offering a comprehensive collection of spoken dialogue data aimed at supporting various speech-related tasks. The dataset includes a total of 101 recorded speech dialogues, representing a diverse range of linguistic characteristics. It spans 55.53 hours of speech data, recorded from 202 participants, and features 60,029 individual utterances. Additionally, the dataset is enriched with annotations across 8 distinct dimensions, ensuring its suitability for training and evaluating robust models across multiple speech-processing tasks.

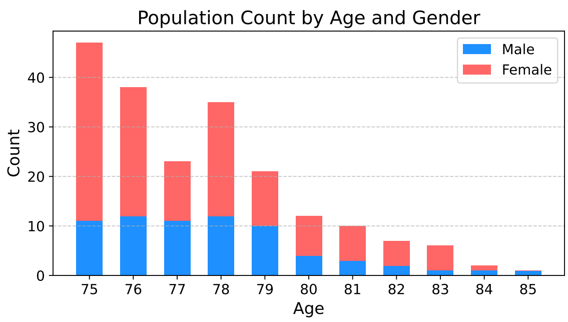

Figure 1: A stacked bar chart showing the population count by age and gender. The chart illustrates the distribution of male and female populations across different age groups.

3.2 Statistics

Participants We recruit 202 participants aged between 75 and 85 years, ensuring a diverse representation across different segments of the ultra-elderly population, and obtain the necessary consent for their participation. Specific authorization details are provided in Appendix A.1. Figure 1 illustrates the gender distribution across different age groups. The dataset includes 69 male and 131 female speakers, with a higher proportion of females. This gender imbalance stems from the relative ease of recruiting female participants during the data collection process, likely due to the higher life expectancy of women in the targeted age range.

https://ai4robot.github.io/emic-en/

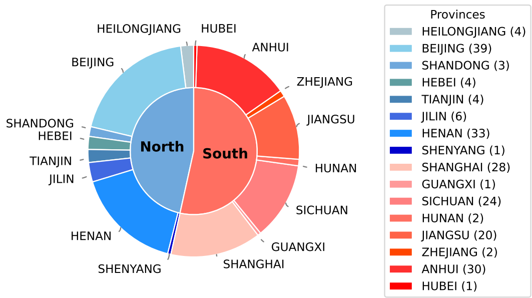

Figure 2: A dual-layer pie chart representing the distribution of provinces across the North and South regions. The inner layer highlights the overall proportions of North and South, while the outer layer shows the specific distribution of provinces within each region.

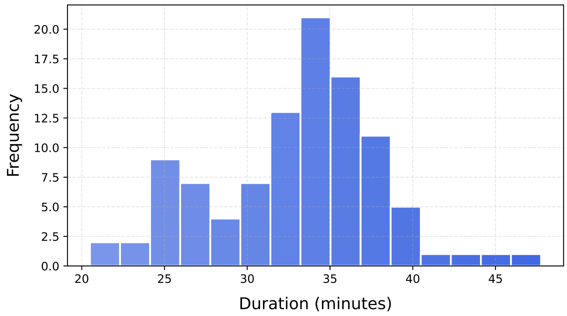

Figure 3: This histogram visualizes the frequency distribution of the given dataset, showing how the values are spread across different ranges.

Geographically, as shown in Figure 2, 94 participants are from northern China, while 108 are from southern China, covering regions such as Beijing, Shanghai, and Sichuan. This regional diversity enhances dialect recognition models by exposing them to a wide range of linguistic patterns.

Recording The dataset comprises spontaneous speech dialogues that span a wide range of topics, aimed at capturing the natural flow and diversity of real-life conversations. These dialogues are recorded using various mobile devices, with a distribution of 70% from Android devices and 30% from iOS smartphones. The topics of these dialogues are specifically chosen to address relevant issues for older adults, such as health, pets, retirement, and other related matters. Each conversation typically covers between one and three topics, with the distribution of topics visualized in Appendix A.2. Figure 3 illustrates the distribution of session durations, which range from 25 to 50 minutes, with the majority of sessions clustering around the 35-minute mark.

Annotation The dataset is annotated by a team of 14 trained annotators who undergo unified training. During the training, they are instructed on a well-defined annotation workflow and strict standards to ensure consistency and accuracy. Detailed guidelines on the annotation process and standards are provided in Appendix A.3.

Annotations are made across four main dimensions. At the speaker profile level, attributes such as age, gender, and origin are annotated. These annotations are particularly useful for analyzing elderly speech signals, helping to study age-related speech characteristics. The session level includes annotations for temporal segmentation and overlapping speech, which are essential for segmenting speech and identifying overlaps in multi-speaker scenarios. At the utterance level, transcriptions and accent intensity are annotated. Transcriptions support ASR. Finally, at the token level, special sound events like laughter are marked, providing insight into non-verbal communication.

4 Experiments

In this section, we assess our dataset across various tasks, including speaker verification, speaker diarization, speech recognition, and speech editing.

4.1 Speaker Verification

This section introduces the Speaker Verification (SV) task, which is essential for verifying the identity of speakers within the geriatric population to ensure their financial security. To facilitate this task, we implement a data partitioning strategy, with further details provided in Appendix B.1.1.

4.1.1 Metrics

We adopt two scoring approaches: probabilistic linear discriminant analysis (PLDA) (Prince and Elder, 2007) and cosine similarity, with evaluation based on two metrics: (1) *Equal Error Rate (EER)*: We define a threshold τ where the miss probability equals the false alarm probability. Specifically, if the similarity score is above this threshold, the system accepts that the speakers are the same person; if it is below this threshold, the system rejects the claim. This threshold is selected when the false acceptance rate equals the false rejection rate. (2) *Minimum Detection Cost (minDCF)*: A cost-sensitive metric for evaluating speaker verification systems under application-specific conditions.

4.1.2 Baselines

We adapt three state-of-the-art speaker embedding systems pre-trained on VoxCeleb (Nagrani

Model	# Params	Dim	Dev (%)	PLDA		Cosine similarity	
				EER (%)	minDCF	EER (%)	minDCF
x-vector	4.2M	512	12.04	14.63	0.9768	19.26	0.9598
ResNet-TDNN	15.5M	256	4.372	10.88	0.8450	11.50	0.9196
ECAPA-TDNN	20.8M	192	8.86	11.54	0.9214	10.24	0.9582

Table 3: Results of fine-tuning baselines on the speaker verification task, where Dim indicates the dimension of the extracted embeddings and Dev represents the EER on the validation set.

et al., 2017) through domain-specific fine-tuning via SpeechBrain (Ravanelli et al., 2021): x-vector architecture (Snyder et al., 2018), ECAPA-TDNN (Desplanques et al., 2020), and ResNet-TDNN (Villalba et al., 2020). Detailed hyperparameters are listed in Appendix B.2.3.

4.1.3 Results and Analysis

Table 3 demonstrates two critical findings from our geriatric voice analysis: First, the strong baseline performance confirms our dataset's suitability for aging voice biometrics. However, age-related vocal degradation (e.g., pitch instability, articulatory imprecision) introduces distinct challenges compared to pediatric voices, potentially affecting gender differentiation and speaker discriminability.

Second, the ECAPA-TDNN's relative underperformance versus parameter-efficient model ResNet using PLDA testing method suggests susceptibility to overfitting in limited elderly speech data scenarios. This emphasizes the need for data augmentation strategies, regularization techniques and hyperparameter finetuning when deploying deep speaker models in geriatric voice applications.

4.2 Speaker Diarization

This section introduces the speaker diarization task, which entails partitioning audio recordings into segments that correspond to individual speakers. This task is essential for assessing the performance of various speaker models, as detailed in Section 4.2.3. The data split is the same as speaker verification task.

4.2.1 Metrics

We employ the Diarization Error Rate (DER) as the evaluation metric for the speaker diarization task. The computation of DER is defined as follows:

$$DER = \frac{FA + MD + Conf}{T}, \qquad (1)$$

where FA denotes the number of segments incorrectly identified as speech when no speaker is present, MD represents the number of segments where speech is present but not detected. $Conf$ refers to the number of segments where detected speech is attributed to the wrong speaker, and T indicates the total number of speech segments in the reference transcript.

In Table 4, we explore two collar settings: 0 seconds and 0.25 seconds. The collar parameter defines a time window around the detected speaker boundaries, allowing for a margin of error in segment alignment. A collar value of 0 seconds requires exact matching of boundaries, while a value of 0.25 seconds introduces a quarter-second tolerance to accommodate minor discrepancies in detection.

4.2.2 Baselines

For the speaker diarization task, we use the PyAnnote toolkit (Bredin, 2023; Plaquet and Bredin, 2023) . This speaker diarization pipeline consists of three primary components: Voice Activity Detection (VAD), the speaker extractor, and clustering methods (e.g., K-Nearest Neighbors).

- **VAD**: We employ the pyannote/segmentation-3.0 model , an end-to-end neural architecture for joint speech activity detection and speaker segmentation. Since false alarm and missed detection rates remain consistent across experimental conditions due to fixed VAD parameters, Table 4 exclusively reports confusion errors and DER metrics.
- **Speaker extractor**: For the speaker extractor module, we replace the default ResNet-34-LM extractor from PyAnnote with fine-tuned versions of x-vector, ECAPA-TDNN,

https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
https://huggingface.co/speechbrain/spkrec-resnet-voxceleb

https://github.com/pyannote/pyannote-audio
https://huggingface.co/pyannote/segmentation-3.0
https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

Model	# Params	Dim	collar=0		collar=0.25	
			DER (%)	Confusion (%)	DER (%)	Confusion (%)
ResNet-34-LM	15.5M	256	33.14	16.82	28.39	16.85
x-vector	4.2M	512	53.01	36.69	49.82	38.28
ResNet-TDNN	15.5M	256	43.44	27.13	39.58	28.03
ECAPA-TDNN	20.8M	192	27.84	11.52	22.85	11.31

Table 4: Results of fine-tuning speaker embedding extraction models and pretrained model ResNet-34-LM on the speaker diarization task.

and ResNet-TDNN architectures, all adapted for speaker verification. This yields four distinct experimental configurations, as detailed in Table 4.

- **Clustering method**: For the clustering method, we employ the default Spectral Clustering, as proposed by PyAnnote.

4.2.3 Results and Analysis

Our experimental analysis yields three principal conclusions regarding elderly speaker diarization:

First, after fine-tuning with the elderly dataset, the ECAPA-TDNN model outperforms the PyAnnote default model ResNet-34-LM by 5.3% DER improvement at 0 collar and 5.54% DER imporvement at 0.25 collar. This superiority demonstrates the effectiveness of our dataset for the diarization of elderly speakers.

Second, our empirical analysis on speaker diarization reveals that ECAPA-TDNN significantly outperforms both ResNet-TDNN and x-vector-based systems. However, Table 3 demonstrates that these performance gaps diminish in speaker verification tasks, with ResNet-TDNN even surpassing ECAPA-TDNN under PLDA scoring. We attribute this discrepancy to two primary factors: the shift to a conversational dataset and ECAPA-TDNN's inherent architectural advantages in noise resilience. While ECAPA-TDNN is typically trained on clean speech corpora, our experiments evaluate its robustness in realistic scenarios characterized by background noise and overlapping speech. The model's noise-resilient components enhance its feature extraction capabilities under such adverse conditions, which proves critical in diarization tasks but less impactful in controlled speaker verification settings. This contrast highlights the importance of conversational datasets in assessing model robustness, as they better reflect real-world challenges that differentiate model performance.

Third, our experiments reveal notably high DER and confusion errors. Our analysis attributes this phenomenon to two dataset-specific factors: (1) a pronounced gender imbalance (1:3 female-to-male ratio) contrasting standard benchmarks' balanced distributions, and (2) age-related vocal changes in elderly speakers that reduce the saliency of secondary sexual voice characteristics. These combined effects create systematic challenges for speaker identity separation, particularly in conversational contexts where demographic diversity and physiological aging patterns naturally occur.

4.3 Speech Recognition

Automatic speech recognition entails transcribing spoken language into text, and recognizing elderly speech patterns is crucial in emergency response scenarios due to age-related vocal characteristics that can affect system reliability. This section presents an empirical evaluation of the collected elderly speech corpus, with data partitioning strategies detailed in Appendix B.1.2 .

4.3.1 Metrics

The experimental results demonstrate the performance on the test dataset after training with the train dataset, using Character Error Rate (CER) as the evaluation metric, which is computed by the following equation:

$$CER = \frac{S + D + I}{N}, \tag{2}$$

where S, D, and I respectively signify the quantities of substitutions, deletions, and insertions. denotes the cumulative number of characters within the reference text. When evaluating character-level transcription accuracy, a system featuring a lower CER is typically regarded as more proficient.

4.3.2 Baselines

We employ the open-source wenet toolkit (Yao et al., 2021) as our training framework, selecting Transformer, Conformer, and E-Branchformer as our baselines. All models are trained using

Encoder	# Params	CER	Accent				Region	
			No	Light	Moderate	Heavy	South	North
Transformer	14.1M	48.99	22.58	49.05	51.07	80.95	48.5	50.24
Conformer	15.7M	34.61	21.23	34.21	37.62	59.52	34.55	34.74
E-Branchformer	16.9M	33.25	20.71	33.03	35.32	64.29	32.97	33.94

Table 5: Decoding performance (CER, %) of Transformer, Conformer, and E-Branchformer models using Attention rescoring, with accent differentiation and region categorization.

Model	# Params	Zero-shot	Fine-tuning
Paraformer-large	232M	14.91	14.41
Whisper-tiny	39M	92.20	58.80
Whisper-base	74M	64.02	38.17
Whisper-small	244M	55.83	28.69
Whisper-medium	769M	60.47	25.77
Whisper-large-v3	1,550M	57.74	23.84

Table 6: Character Error Rate (CER) (%) of supervised pre-trained baselines in both zero-shot and fine-tuning settings.

a combined Connectionist Temporal Classification (CTC) and Attention-based Encoder-Decoder (AED) approach. We fine-tune the hyperparameters of these three models to ensure comparable parameter counts while achieving optimal performance. Detailed hyperparameter configurations can be found in the Appendix B.2.1 and B.2.2.

The following models are considered in the context of training from scratch:

- **Transformer**: The standard Transformer architecture employs both CTC and AED objectives, establishing a widely-adopted baseline for ASR.
- **Conformer**: The Conformer (Gulati et al., 2020) model integrates convolutions with self-attention for ASR, sandwiched between two feed-forward layers.
- **E-Branchformer**: Proposed by Kwangyoun Kim et al. (Kim et al., 2022), E-Branchformer builds upon the Branchformer (Peng et al., 2022), which attains performance levels comparable to Conformer. E-Branchformer improves on this framework by implementing a novel merging strategy and integrating additional point-wise modules.

In addition to the three models trained from scratch, we fine-tune two pre-trained models: Paraformer, which employs the Wenet framework, and Whisper, whose hyperparameters and code base are described in Appendix B.2.2.

- **Paraformer**: Proposed by Gao et al. (Gao et al., 2022), Paraformer is a fast and accurate parallel transformer model that leverages a continuous integrate-and-fire (CIF) (Dong and Xu, 2020) predictor to estimate the number of tokens and generate hidden representations. This pre-trained model is trained on a non-public, industry-grade dataset comprising 60,000 hours of Chinese ASR data.
- **Whisper**: Whisper (Radford et al., 2023) is a Transformer-based multilingual ASR model developed by OpenAI, trained on 68,000 hours of labeled speech data. We examine various versions of Whisper, ranging from tiny to large-v2, with model sizes varying from 39M to 1.55B.

4.3.3 Results and Analysis

Models Trained from Scratch We analyze models trained from scratch across three aspects: baseline performance, accent intensity impact, and regional variations.

Baseline Model Performance: Table 5 compares three models, namely E-Branchformer, Conformer, and Transformer, which are trained from scratch and utilize the attention rescoring decoding method. The E-Branchformer achieves the lowest overall CER, outperforming Conformer by 1.36% and Transformer by 15.74%. This performance advantage holds consistently across all tested conditions, including varying accent intensities and geographical regions.

Accent Intensity Impact: As shown in Table 5, CER increases with accent intensity: Moderate Accent yields higher CER than Light Accent, which in turn exceeds No Accent. This trend highlights the growing recognition challenge as accents become more pronounced.

Regional Variations: The E-Branchformer achieves a 0.97% lower CER in the South than in the North, with the Conformer and Transformer showing improvements of 0.19% and 1.74%, respectively. These differences indicate slightly higher recognition difficulty in the North, although

https://github.com/openai/whisper

overall performance remains comparable. At the provincial level, substantial variations in CER are observed across different provinces. For detailed analysis, please refer to the AppendixB.3.

Finetuned Models Table 6 presents comparative CER results for two model families: 1) our Wenet-finetuned Paraformer-large architecture (Yao et al., 2021), and 2) a series of Whisper models (Radford et al., 2023) adapted through whisper-flamingo fine-tuning. The analysis reveals two key findings:

First, although Whisper models show modest CER improvements with larger parameter counts (except for small), all variants have CERs exceeding 50% in zero-shot senario. This performance gap suggests potential domain mismatch between Whisper's training data distribution and elderly speech characteristics, possibly due to underrepresentation of senior voices in Whisper's pretraining corpus. Such mismatch may explain the frequently observed hallucination patterns in elderly speech recognition results. Notably, our targeted fine-tuning reduces CERs substantially, improving cer of Whisper-large-v3 from 57.74 to 23.84%, demonstrating both the challenge level of our elderly speech dataset and its utility for domain adaptation.

Second, the Paraformer-large model achieves 14.91% CER in zero-shot evaluation and 14.41% CER after fine-tuning , outperforming all Whisper variants by significant margins. This advantage likely stems from Paraformer's pretraining on 60,000 hours of proprietary Chinese speech data encompassing diverse regional accents, age groups, and speaking styles. This demonstrates that Paraformer has better generalization ability in the field of Chinese speech recognition.

Method	**MCD** (↓)	**STOI** (↑)	**PESQ** (↑)
CampNet	7.302	0.220	1.291
EditSpeech	6.225	0.514	1.363
A^3T	5.851	0.586	1.455
FluentSpeech	5.811	0.627	1.645

Table 7: Evaluation results of speech editing models trained from scratch.

4.4 Speech Editing

Speech editing is a generative task that modifies corresponding speech based on text alterations, and editing speech from elderly individuals is particularly useful for editing interviews with seniors. To facilitate this task, we implement a new data split, as detailed in Appendix B.1.3. However, utilizing this data for generative tasks poses risks related to elder fraud, particularly in the context of telemarketing scams targeting senior individuals.

4.4.1 Baselines

We employ the Speech Editing Toolkit framework to implement our models, which include CampNet(Wang et al., 2022), EditSpeech(Tan et al., 2021), A^3T(Bai et al., 2022) , and FluentSpeech(Jiang et al., 2023). The detailed hyperparams are described in Appendix B.2.4.

4.4.2 Metrics

We utilize Mel-Cepstral Distortion (MCD) (Kubichek, 1993), Short-Time Objective Intelligibility (STOI) (Taal et al., 2010), and Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001), which are commonly used objective metrics for assessing speech generation quality.

4.4.3 Performance Analysis

The results presented in Table 7 indicate that FluentSpeech achieved the best performance, with all three objective metrics falling within acceptable ranges. In relevant research (Jiang et al., 2023; Bai et al., 2022; Wang et al., 2022; Liu et al., 2024), these experimental values are widely recognized as acceptable. This suggests that our dataset is suitable for generative tasks.

5 Conclusion

In this paper, we present SeniorTalk, a Mandarin dataset featuring spontaneous conversations among individuals aged 75 and older. With 55.53 hours of data from 202 speakers across 16 provinces in China, this dataset offers valuable resources for developing voice technologies for aging populations. By providing detailed annotations and conducting extensive experiments across key speech tasks, we establish SeniorTalk as a benchmark for evaluating models for elderly speakers, aiming to bridge the vocal age gap and promote the development of more inclusive voice technologies.

Limitations

While SeniorTalk advances research on elderly speech processing, several limitations should be

https://github.com/roudimit/whisper-flamingo

https://github.com/pyannote/pyannote-audio noted. First, while our dataset includes 202 participants across 16 provinces, its scale remains modest compared to the linguistic diversity and population size of China. Expanding the dataset to include thousands of speakers would better capture regional accents, dialects, and individual vocal variability. Second, the current age range could be extended to include more super-elderly individuals (e.g. 85+ or centenarians), as advanced aging often correlates with unique vocal patterns due to physiological decline, cognitive changes, or age-related pathologies. Third, geographic coverage could be broadened to encompass underrepresented regions (e.g., rural areas or ethnic minority regions) to mitigate potential demographic biases. Additionally, longitudinal data capturing vocal aging trajectories could further enhance the dataset's utility. These extensions would strengthen the generalizability of voice technologies for aging populations and address the complex interplay of age, health, and regional linguistic diversity.

Ethics Statement

This study strictly adhered to rigorous ethical protocols to safeguard elderly participants' rights. Audio recordings were conducted in quiet indoor environments at senior care facilities. To accommodate potential cognitive decline among participants, multiple conversation topics were provided during dyadic interactions, with recording devices positioned equidistant between paired participants.

Prior to each session, informed consent was obtained after explaining the research objectives and data collection parameters, including voice characteristics, conversational content, age documentation, and accent analysis. Participants were compensated monetarily, with amounts ranging from 330 to 400 RMB. This compensation was calibrated according to local purchasing power, thereby ensuring equitable remuneration tailored to the specific geographical locations of the participants. All personal identifiers (e.g., national ID numbers, full names) were systematically de-identified by replacing them with unique speaker identifiers, despite initial age verification requiring temporary ID inspection.

The dataset carries inherent risks requiring stringent governance. Potential malicious exploitation for voice synthesis could potentially exacerbate elderly-targeted telecommunications fraud through voice spoofing. Accordingly, access is strictly restricted to vetted academic researchers through institutional credential verification, with legally binding agreements prohibiting commercial use or redistribution.

Key ethical safeguards implemented include: 1) Explicit participant consent through verbal and written confirmation, 2) Comprehensive privacy preservation via anonymization protocols, 3) Fair compensation aligned with regional economic standards, 4) Institutional Review Board approval for all data collection procedures, 5) Multi-layered access controls preventing unauthorized usage.

This framework ensures compliance with international research ethics standards (Declaration of Helsinki, GDPR principles) while balancing scientific utility with participant protection in vulnerable populations.

References

Frédéric Aman, Michel Vacher, Solange Rossato, and François Portet. 2013. Speech recognition of aged voice in the aal context: Detection of distress sentences. In *2013 7th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. 2022. A^3T: Alignment-aware acoustic and text pretraining for speech synthesis and editing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1399–1411. PMLR.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Yang Chen, Yuhang Jia, Shiwan Zhao, Ziyue Jiang, Haoran Li, Jiarong Kang, and Yong Qin. 2024. Diffeditor: Enhancing speech editing with semantic enrichment and acoustic consistency. *Preprint*, arXiv:2409.12992.

Narumol Chumuang, Thittaporn Ganokratanaa, Patiyuth Pramkeaw, Mahasak Ketcham, Suvil Chomchaiya, and Worawut Yimyam. 2024. Voice-activated assistance for the elderly: Integrating speech recognition and iot. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–4. IEEE.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Linhao Dong and Bo Xu. 2020. Cif: Continuous integrate-and-fire for end-to-end speech recognition.

In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422.

Meiko Fukuda, Ryota Nishimura, Hiromitsu Nishizaki, Koharu Horii, Yurie Iribe, Kazumasa Yamamoto, and Norihide Kitaoka. 2023. A new speech corpus of super-elderly japanese for acoustic modeling. *Computer Speech & Language*, 77:101424.

Meiko Fukuda, Hiromitsu Nishizaki, Yurie Iribe, Ryota Nishimura, and Norihide Kitaoka. 2020. Improving speech recognition for the elderly: A new corpus of elderly Japanese speech and investigation of acoustic modeling for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6578–6585, Marseille, France. European Language Resources Association.

Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pages 2063–2067.

Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. 2022. Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2597–2611.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Valerie Hazan, Outi Tuomainen, Jeesun Kim, and Christopher Davis. 2017. elderlucid: London ucl older adults' clear speech in interaction database.

Shujie Hu, Xurong Xie, Mengzhe Geng, Zengrui Jin, Jiajun Deng, Guinan Li, Yi Wang, Mingyu Cui, Tianzi Wang, Helen Meng, and Xunying Liu. 2024. Self-supervised asr models and features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3561–3575.

Yurie Iribe, Norihide Kitaoka, and Shuhei Segawa. 2015. Development of new speech corpus for elderly japanese speech recognition. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 27–31.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. 2024. Audioeditor: A training-free diffusion-based audio editing framework. *Preprint*, arXiv:2409.12466.

Ziyue Jiang, Qian Yang, Jialong Zuo, Zhenhui Ye, Rongjie Huang, Yi Ren, and Zhou Zhao. 2023. Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11655–11671.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. 2022. E-branchformer: Branchformer with enhanced merging for speech recognition. *Preprint*, arXiv:2210.00077.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

Ajinkya Kulkarni, Anna Tokareva, Rameez Qureshi, and Miguel Couceiro. 2024. The balancing act: Unmasking and alleviating ASR biases in Portuguese. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 31–40, St. Julian's, Malta. Association for Computational Linguistics.

Rui Liu, Jiatian Xi, Ziyue Jiang, and Haizhou Li. 2024. Fluenteditor: Text-based speech editing by considering acoustic and prosody consistency. *Proc. InterSpeech2024*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

Hajime Orimo, Hideki Ito, Takao Suzuki, Atsushi Araki, Takayuki Hosoi, and Motoji Sawabe. 2006. Reviewing the definition of "elderly". *Geriatrics & gerontology international*, 6(3):149–158.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. *Preprint*, arXiv:2207.02971.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. 2013. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17:127–144.

Simon JD Prince and James H Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2.

Angelo Scuteri and Peter M. Nilsson. 2024. Chapter 4 - aging population: Challenges and opportunities in a life course perspective. In Pedro Guimarães Cunha, Peter M. Nilsson, Michael Hecht Olsen, Pierre Boutouyrie, and Stéphane Laurent, editors, *Early Vascular Aging (EVA) (Second Edition)*, second edition edition, pages 35–39. Academic Press, Boston.

Irina A Sekerina, Anna Smirnova Henriques, Aleksandra S Skorobogatova, Natalia Tyulina, Tatiana V Kachkovskaia, Svetlana Ruseishvili, and Sandra Madureira. 2024. Brazilian portuguese-russian (braporus) corpus: automatic transcription and acoustic quality of elderly speech during the covid-19 pandemic. *Linguistics Vanguard*, 9(s4):375–388.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217.

Daxin Tan, Liqun Deng, Yu Ting Yeung, Xin Jiang, Xiao Chen, and Tan Lee. 2021. Editspeech: A text based speech editing system using partial inference and bidirectional fusion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 626–633. IEEE.

Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak. 2020. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026.

Tao Wang, Jiangyan Yi, Ruibo Fu, Jianhua Tao, and Zhengqi Wen. 2022. Campnet: Context-aware mask prediction for end-to-end text-based speech editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2241–2254.

Tianqi Wang, Quanlei Yan, Jingshen Pan, Feiqi Zhu, Rongfeng Su, Yi Guo, Lan Wang, and Nan Yan. 2019. Towards the speech features of early-stage dementia: Design and application of the mandarin elderly cognitive speech database. In *Interspeech 2019*, pages 4529–4533.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Interspeech 2021*, pages 4054–4058.

Victoria Young and Alex Mihailidis. 2010. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112.

A Supplementary Dataset Information

A.1 Authorization

This study employs time-series biometric data collected under a formal ethical authorization framework developed in collaboration with a third-party AI data service provider specializing in biometric acquisition. The framework ensures methodological transparency and regulatory alignment through the following mechanisms:

Legally Binding Consent Protocols Participants provide informed consent for the collection of vocal and physiological time-series data, explicitly authorizing its use in AI research and derivative model development.

Rights Management All datasets and derived models remain the exclusive intellectual property of the anonymized provider. Participants retain conditional rights to access, modify, or request data deletion, contingent on technical feasibility (deletion requests that invalidate associated research outputs may require proportional compensation).

Cross-Jurisdictional Compliance Third-party data sharing requires explicit opt-in consent, with passive approval mechanisms activated only after a 72-hour objection period following notification.

Biometric Corpus Specifications The dataset includes anonymized temporal features such as age, gender, regional dialect markers, and sequential vocal patterns (e.g., pitch dynamics, spectral entropy trajectories).

A.2 Topic

This section presents the frequency statistics of all topics in the dataset, as illustrated in Figure 4. In total, there are 13 major categories, encompassing 58 distinct topic labels. The distribution of these 13 categories is depicted in the figure. These topics primarily reflect the everyday concerns of elderly individuals, with particular emphasis on areas such as Leisure, Health, and Retirement Life, which appear most frequently. These topics align closely with the key interests and priorities of older adults. Understanding and focusing on these subjects is crucial, as it can greatly benefit the performance of ASR systems, especially when catering to elderly users in relevant contexts.

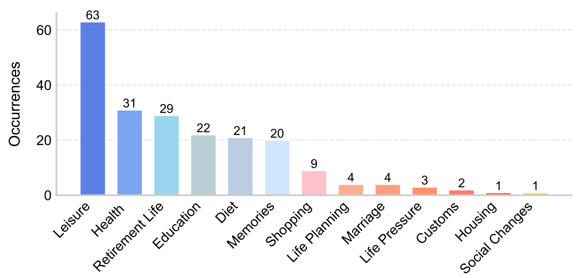

Figure 4: This distribution of topics.

A.3 Annotation

This section presents information facilitating annotation, covering details about annotators, the annotation process, and the content being annotated.

A.3.1 Annotator

As shown in Table 10, We have a total of 14 annotators. Eleven of them are between 23 and 30 years old, and three are between 30 and 50 years old. All annotators hold undergraduate degrees. Moreover, they are proficient in their local dialects, which facilitates the transcription of accented speech.

A.3.2 Annotation workflow

The annotation process utilizes a proprietary cloud-based platform developed by a third-party data service provider. Annotators access tasks through this platform, which facilitates manual segmentation of conversational data into discrete utterances followed by automated speech recognition to generate preliminary transcripts. These baseline outputs undergo human revision alongside annotation of linguistic features including accent intensity levels (see Table 8) and paralinguistic markers such as [NOISE] or [MUSIC]. A multi-tier quality assurance protocol is implemented through the same platform, where dedicated reviewers systematically validate annotation consistency and accuracy prior to dataset finalization.

A.3.3 Annotation Information

Table 9 presents our four-tier annotation framework comprising speaker, session, utterance, and token levels, with each layer supporting distinct speech processing tasks through systematically designed metadata.

The speaker-level annotations capture demographic attributes (geographic origin, gender, age) to enable geriatric voice analysis.

At the session level, comprehensive temporal annotations of turn-taking boundaries and overlap detection facilitate speaker diarization and speech separation benchmarking.

Utterance-level transcriptions incorporate accent intensity labels and orthographic normalization,

Level	Annotation Criteria
No Accent (0)	*Phonetics*: Fully standard pronunciation with clear articulation *Comprehensibility*: Effortlessly understood by non-native listeners *Regional Features*: No detectable regional phonological characteristics
Light Accent (1)	*Phonetics*: Occasional non-standard vowels/consonants (<20% utterances) *Comprehensibility*: Minor listening effort required for full understanding *Regional Features*: Subtle but identifiable regional speech patterns
Moderate Accent (2)	*Phonetics*: Frequent non-standard prosody/lexical stress (20-50% utterances) *Comprehensibility*: Requires focused attention, occasional repetition needed *Regional Features*: Strong regional phonological markers affecting intelligibility
Heavy Accent (3)	*Phonetics*: Pervasive non-standard articulation (>50% utterances) *Comprehensibility*: Frequent breakdowns requiring contextual guessing *Regional Features*: Severely divergent from standard phonological norms

Table 8: Accent Intensity Annotation Guidelines

Annotation Level	Annotation Dimension	Associated Tasks	Representative Instances
Speaker Metadata	Demographic Age Geographic Origin (Province) ID Card Gender	Elderly Speech Analysis	75 Jiangsu, Henan Female/Male
Session	Temporal Segmentation Overlapping Speech	Speaker Diarization Speech Separation	[48.475 - 73.582] `spk_001` `trans1(trans2)[+]`
Utterance	Raw Transcription Accent Intensity (0-3)	Speech Recognition Ordinal Classification	`[Mandarin utterance]` Neutral (0) / Strong (3)
Token	Special Markers	Paralinguistic Analysis	`[MUSIC],[NOISE],[LAUGHTER]`

Table 9: Annotation Levels and Their Associated Tasks for Dataset Analysis

jointly supporting automatic speech recognition and ordinal classification of vocal aging patterns.

The token tier extends this granularity with paralinguistic markers, including [MUSIC], [NOISE], [LAUGHTER], and [SONANT], enabling quantitative analysis of non-lexical vocal characteristics critical for elderly communication studies.

ID	Region	Age	Gender	Education
1	Henan	36	Male	Bachelor
2	Hebei	23	Female	Bachelor
3	Hebei	24	Female	Bachelor
4	Hebei	24	Female	Bachelor
5	Hebei	30	Female	Bachelor
6	Fujian	27	Male	Bachelor
7	Fujian	23	Male	Bachelor
8	Fujian	23	Female	Bachelor
9	Fujian	38	Female	Bachelor
10	Chuzhou	22	Male	Bachelor
11	Chuzhou	27	Male	Bachelor
12	Chuzhou	23	Female	Bachelor
13	Yunnan	27	Male	Bachelor
14	Hunan	47	Male	Bachelor

Table 10: Annotator Information

B Supplementary Experiments

This section mainly presents supplementary experimental details, including the dataset partitioning for each task and the hyperparameters. The fine-tuning and zero-shot tests of the Whisper model are performed using an NVIDIA A800, while all other experiments are carried out with an NVIDIA GeForce RTX 3090.

B.1 Data Split

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	0-182	48591	30.47	2.26
Dev	0-182	5398	3.40	2.27
Test	182-202	6040	3.95	2.35
Sum	0-202	60029	37.82	2.27

Table 11: Summary of dataset splits, including the speakers range (# Spk.), utterances (# Utt.), total duration (Dur.), and average utterance length (Avg.) for speaker verification and speaker dirazation tasks.

Encoder	Accum Grad	Batch size	LR	Warmup	Epochs
Transformer	4	4	1.00E-03	5000	100
Conformer	4	4	1.00E-03	5000	60
E-Branchformer	4	4	5.00E-04	5000	60

Table 12: Hyperparameters for training ASR models from scratch.

Model	Accum Grad	Batch size	Learning rate	Warmup	Training steps
Paraformer-large	4	28	5.00E-04	5000	101579
Whisper-large	1	8	5.00E-6	1000	90000
Whisper-medium	1	3	6.25e-6	1000	225000
Whisper-small	1	4	1.25e-5	1000	225000

Table 13: Hyperparameters for fine-tuning pre-trained ASR models.

B.1.1 Speaker Verification and Diarization

For the speaker verification experiments, we first segment the original dialogue dataset into sentence-level units. Given that the dataset includes precise timestamps for each sentence and specific audio markers for events such as overlapping speech and noise, we extract clean audio segments devoid of these special sound events and speaker overlaps. From the 101 dialogues in the dataset, we randomly select 10 dialogues, using the corresponding segmented audio at the sentence level as our test dataset. The remaining audio is split, with 10% randomly allocated for the validation set and 90% designated for the training set. The detailed information of the datasplit is shown in 11. After the datasplit, we create 20,000 carefully balanced verification pairs (50% genuine vs. 50% impostor pairs) from the test set. This trial composition ensures uniform coverage of both intra-speaker (spk_i, spk_j) and inter-speaker pairs (spk_i, spk_j). The detailed datasplit information is shown in table 11. For speaker diarization, the dataset split is identical to that of the speaker verification task, except that we use dialogues instead of sentences.

B.1.2 Speech Recognition

For the elderly dataset, we initially segment the dialogue - based data into clean sentence - level data following the annotations. This process is similar to the data split in the speaker verification task described in Appendix B.1.1. Subsequently, we partition the data by speaker and randomly divide it into three subsets: the training set, the development set (dev), and the test set, with a ratio of 8:1:1. The detailed data split information is presented in Table 18.

B.1.3 Speech Editing

After converting the conversation into sentences using the method described in the speaker verification data split in Appendix B.1.1, we filter the sentences. Only those with more than four characters are included for training and testing. We then divide the data into a training-to-testing ratio of 9:1.

B.2 Hyperparams

B.2.1 ASR Model Training from Scratch

The hyperparameters used for training ASR models from scratch are presented in Table 12. We experimented with three different encoder architectures: Transformer, Conformer, and E-Branchformer. For all models, we used an accumulation gradient of 4 and a batch size of 4. The initial learning rate was set to 1.00E-03 for Transformer and Conformer models, while it was 5.00E-04 for the E-Branchformer model. A warmup period of 5000 steps was used for all models. The models were trained for a maximum of 100 epochs (Transformer) and 60 epochs (Conformer and E-Branchformer).

B.2.2 ASR Model Fine-tuning

Table 13 outlines the hyperparameters used for fine-tuning pre-trained ASR models. Our experiments covered Paraformer-large and multiple Whisper variants (large/medium/small). The batch size varied depending on the model, ranging from 3 to 28. The learning rates were carefully chosen for each model, spanning from 5.00E-06 to 5.00E-04. A warmup period of 1000 steps was used across all fine-tuned models. The total training steps also varied significantly depending on the model, reflecting the different sizes and pre-training strategies. All

Model	Batch size	LR Schedule	Init LR	Base LR	Epochs
ECAPA-TDNN	128	Cyclic	5.00E-03	1.00E-08	20
ResNet-TDNN	128	Cyclic	5.00E-03	1.00E-08	20
x-vector	128	Linear	5.00E-03	1.00E-04	20

Table 14: Hyperparameters for training speaker verification models.

Model	Params	Batch size	LR	Warmup	Training steps
CampNet	21.22M	16	2.00E-04	8000	200000
EditSpeech	48.15M	16	2.00E-04	8000	200000
A^3T	14.86M	16	2.00E-04	8000	200000
FluentSpeech	23.86M	30	2.00E-04	8000	200000

Table 15: Hyperparameters for training speaker editing models.

Model	Shanghai	Guangxi	Sichuan	Anhui	Zhejiang	Jiangsu	Hunan
Paraformer-fintuned	13.68	4.88	9.62	16.11	12.42	16.57	14.6
Paraformer	13.26	6.24	10.07	17.91	12.3	16.61	13.18
Brachformer	27.37	10.82	25.94	36.58	29.83	36.68	39.94
Conformer	29.11	11.91	27.11	38.01	30.17	39.44	40.65
Tranformer	42.27	26.91	41.78	52.3	46.53	51.82	54.9

Table 16: Model Performance in Southern Provinces

Model	Beijing	Heilongjiang	Liaoning	Hebei	Henan
Paraformer-fintuned	7.38	5.85	27.08	7.71	20.93
Paraformer	6.8	5.38	23.85	9.41	24.19
Brachformer	23.65	21.21	55.31	19.63	40.85
Conformer	25.15	20.27	56.88	20.44	41.59
Tranformer	43.01	39.61	67.58	36.61	56.03

Table 17: Model Performance in Northern Provinces

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	162	47269	29.95	2.28
Dev	20	6891	4.09	2.14
Test	20	5869	3.77	2.31
Sum	202	60029	37.81	2.27

Table 18: Summary of dataset splits, including the number of speakers (# Spk.) and utterances (# Utt.), total duration (Dur.), and average utterance length (Avg.) for speech recognition task.

the hyperparameters of whisper models are the default setting of whisper-flamingo .

B.2.3 Speaker Verification Model Training

Training details for our speaker verification models – including ECAPA-TDNN, ResNet-TDNN, and x-vector architectures – are summarized in Table 14. All models used a batch size of 128. For ECAPA-TDNN and ResNet-TDNN, a cyclic learning rate schedule was employed with an initial learning rate of 5.00E-03 and a base learning rate of 1.00E-08. The x-vector model utilized a linear learning rate schedule with an initial learning rate of 5.00E-03 and a base learning rate of 1.00E-04. All models were trained for 20 epochs.

https://github.com/roudimit/whisper-flamingo

Split	# Utt.	Dur. (hrs)	Avg. (s)
Train	44954	32.42	2.60
Test	4994	3.68	2.65
Sum	49948	36.10	2.60

Table 19: Summary of dataset splits, including the utterances (# Utt.), total duration (Dur.), and average utterance length (Avg.) for speech editing tasks.

B.2.4 Speech Editing Model Training

For training our second set of Speech Editing models(CampNet, EditSpeech, A3T, and FluentSpeech), we utilized the hyperparameters summarized in Table 15. The number of trainable parameters for each model is also provided. All models, except FluentSpeech, used a batch size of 16. FluentSpeech used a batch size of 30. The initial learning rate was set to 2.00E-04 for all models. A warmup period of 8000 steps was used, and all models were trained for 200,000 steps.

B.3 Extra Experiment Analysis

This section primarily presents the detailed CER of the models trained with Wenet in the speech recognition experiments. Specifically, the specific recognition results of these models for sentences from different provinces are shown in Table 17 and Table 16. In Table 17, we can observe that the CERs of Beijing, Hebei, and Heilongjiang are relatively low, while the recognition difficulty for Liaoning and Henan increases. Similarly, in Table 16, Guangxi has the lowest recognition difficulty among all provinces. Moreover, the order of recognition difficulty, that is, the ranking of CERs, among other provinces is roughly the same. This indicates that when we break down the regions to the provincial level, the recognition difficulty varies across different provinces. (Chen et al., 2024) (Jia et al., 2024)