

---

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

---

Yang Chen<sup>1\*</sup>   Hui Wang<sup>1\*</sup>   Shiyao Wang<sup>1</sup>   Junyang Chen<sup>1</sup>   Jiabei He<sup>1</sup>

Jiaming Zhou<sup>1</sup>   Xi Yang<sup>2</sup>   Yequan Wang<sup>2</sup>   Yonghua Lin<sup>2</sup>

Yong Qin<sup>1†</sup>

<sup>1</sup>College of Computer Science, Nankai University,

<sup>2</sup>Beijing Academy of Artificial Intelligence, Beijing, China,

## 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. Code is available at <https://github.com/flageval-baai/SeniorTalk> and data at <https://huggingface.co/datasets/evan0617/seniortalk>.

## 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 [1, 2]. 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 [3, 4]. However, many state-of-the-art speech systems struggle to perform effectively within the elderly population, exhibiting biases on elderly vocal patterns. [5, 6]. For instance, existing speech recognition systems often show poor performance with older users, influenced by factors such as speech deterioration, hearing loss, health issues [7], and the diversity of speech patterns among older adults [8, 9]. A key underlying reason for this challenge is the lack of datasets specifically addressing the unique needs of ultra-high-age individuals [9], 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 [10, 11], 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

---

\*Equal Contribution. † Corresponding author.

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

Dataset	Mean Age	Style	Annotation Features					Age Group Distribution (%)		
			Region	Transcript	Timestamp	Accent	Sound	55-65	65-75	75-85
ASR0060	60.42	Reading	North/South/Other	✓	N/A	×	×	56.7	42.5	0.8
MECSD	67.27	Reading	×	✓	N/A	×	✓	17.6	72.9	9.4
SeniorTalk	79.5	Conversation	Provincial	✓	✓	✓	✓	0	0	100

accents. For example, both the AISHELL-ASR0060 and MECSD datasets [12] adopt an unusually low age threshold of 55 years for defining elderly participants. In particular, these data sets show extremely low proportions of participants aged 75 to 85 years: only 0.8% for AISHELL-ASR0060 and 9.4% for MECSD, which means that the oldest population (75+ years) is severely underrepresented. Similarly, the S-JNAS corpus of elderly Japanese speech reports a mean speaker age of 67.6 years [11]. 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 [13].

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 [12, 14], 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 [12], 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**, the only freely available open-source Mandarin speech 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) [18] 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 [17] advance French elderly speech analysis through distress detection benchmarks for ambient assisted living systems. Further corpus development has continued with

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.

Corpus	Language	Age	Style	# Spks.	Dur.(hrs)	Year	Avail.
BraPoRus [10]	Brazilian Portuguese-Russian	59-98	Monologue, interview, ...	1,500	170	2024	N
EARS [15]	Japanese	70-99	Reading	123	13.4	2023	N
Improving S-JNAS [11]	Japanese	65-99	Reading	221	31.7	2020	Y
AISHELL-ASR0060	Mandarin	55+	Reading	503	793	2019	Y
MECSD [12]	Mandarin	55-85	Reading	85	110	2019	P
elderLUCID [16]	English	19-84	Reading	83	N	2017	N
Develop S-JNAS [14]	Japanese	60-98	Reading	100	9.2	2015	Y
ERES38 [17]	French	68-98	Interview	22	17	2013	N
AD80 [17]	French	62-94	-	43	4.7	2013	N
CCC [18]	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

the creation of specialized datasets, including a corpus of 100 elderly Japanese speakers designed to enhance human-robot interaction in elder care [14]. The elderLUCID project [16] 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) [12] and the AISHELL-ASR0060 database<sup>2</sup> 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 [11]. The Elderly Multimodal Interpersonal Conversation (E-MIC)<sup>3</sup> 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 [15], has continued to refine acoustic modeling techniques for super-elderly Japanese speakers, while the BraPoRus corpus [10] 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.

#### 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 1a illustrates the gender distribution across different age groups. The dataset includes 67 male and 135 female

<sup>2</sup> [https://www.aishelltech.com/General\\_Datasets](https://www.aishelltech.com/General_Datasets)

<sup>3</sup> <https://ai4robot.github.io/emic-en/>

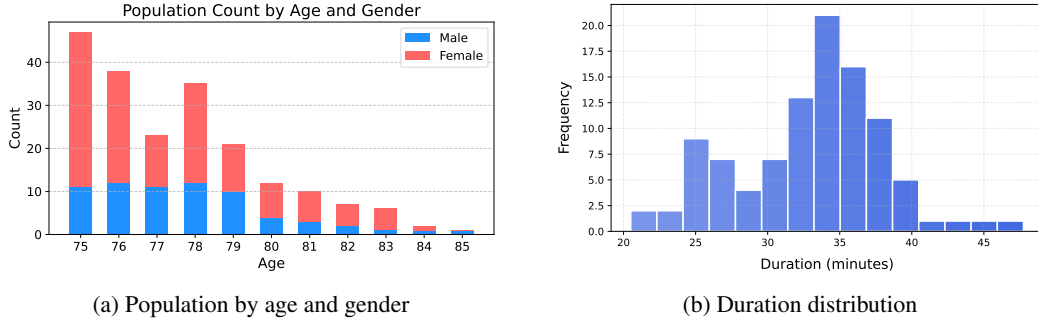


Figure 1: Data analysis: (a) Age-gender structure; (b) Duration distribution.

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. 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 1b illustrates the distribution of session durations, which range from 25 to 50 minutes, with the majority of sessions clustering around the 35-minute mark.

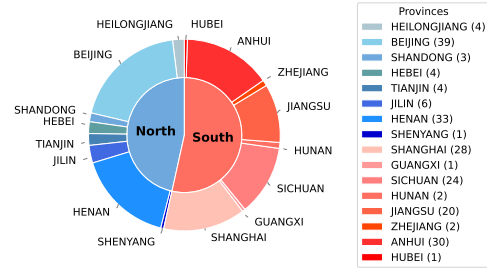


Figure 2: This histogram visualizes the region distribution of the given dataset, showing how the values are spread across different region.

**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.4.

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.

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.

Model	# Params	Dim	Dev (%)	PLDA		Cosine similarity	
				EER (%)	minDCF	EER (%)	minDCF
x-vector	4.2M	512	12.04	17.53	0.9032	22.38	0.8964
ResNet-TDNN	15.5M	256	<b>4.37</b>	11.39	0.7259	13.15	0.8225
ECAPA-TDNN	20.8M	192	5.85	<b>8.56</b>	<b>0.6263</b>	<b>5.66</b>	<b>0.6121</b>

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

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	<b>27.84</b>	<b>11.52</b>	<b>22.85</b>	<b>11.31</b>

## 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) [19] and cosine similarity, with evaluation based on two metrics: (1) *Equal Error Rate (EER)*: We define a threshold  $\tau$  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 [20] through domain-specific fine-tuning via SpeechBrain [21]: x-vector architecture [22],<sup>4</sup> ECAPA-TDNN [23],<sup>5</sup> and ResNet-TDNN [24].<sup>6</sup> 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 finetuned models’ performance relative to the pretrained model validates the suitability of our dataset for aging voice biometric tasks. This suggests that age-related vocal degradation (e.g., pitch instability, articulatory imprecision) introduces distinct challenges compared to pediatric voices, potentially affecting gender differentiation and speaker discriminability. The pretrained model results without fine-tuning are provided in Appendix B.3.2 for comparison.

Second, despite achieving the best performance on the test set, the ECAPA-TDNN underperforms the ResNet-TDNN on the development set, indicating potential overfitting risks for elderly speech data. This emphasizes the need for data augmentation strategies, regularization techniques and hyperparameter tuning when deploying deep speaker models in geriatric voice applications.

<sup>4</sup> <https://huggingface.co/speechbrain/spkrec-xvect-voxceleb>

<sup>5</sup> <https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

<sup>6</sup> <https://huggingface.co/speechbrain/spkrec-resnet-voxceleb>

## 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:

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

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 [25, 26].<sup>7</sup> 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,<sup>8</sup> 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<sup>9</sup> with fine-tuned versions of x-vector, ECAPA-TDNN, 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 improvement at 0.25 collar. This superiority demonstrates the effectiveness of our dataset for the diarization of elderly speakers.

Second, as shown in Table 4, ECAPA-TDNN achieves significantly superior performance over both ResNet-TDNN and x-vector-based baselines on the speaker diarization task. Moreover, consistent with these findings, the ECAPA-TDNN demonstrates consistently lower EER across both PLDA and cosine similarity scenarios on the speaker verification task as shown in table 3. These results highlight the ECAPA-TDNN’s robustness, confirming its strong generalization ability not only in clean, utterance-level speaker verification tasks but also in more challenging, real-world conversational speaker diarization scenarios involving increased noise and overlapping speech.

<sup>7</sup> <https://github.com/pyannote/pyannote-audio>

<sup>8</sup> <https://huggingface.co/pyannote/segmentation-3.0>

<sup>9</sup> <https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM>

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 (1:2 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.

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

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

### 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:

$$CER = \frac{S + D + I}{N}, \quad (2)$$

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 [27] as our training framework, selecting Transformer, Conformer, and E-Branchformer as our baselines. All models are trained using 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 [28] model integrates convolutions with self-attention for ASR, sandwiched between two feed-forward layers.
- **E-Branchformer:** Proposed by Kwangyoun Kim et al. [29], E-Branchformer builds upon the Branchformer [30], 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. [31], Paraformer is a fast and accurate parallel transformer model that leverages a continuous integrate-and-fire (CIF) [32] 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 [33]<sup>10</sup> is a Transformer-based multilingual ASR model developed by OpenAI, trained on 680,000 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. The duration distribution of utterances across different accent levels is presented in Appendix A.3.

*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 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.3.1.

Table 6: Character Error Rate (CER) (%) of the pretrained Paraformer-large model, along with various sizes of Whisper models (tiny, base, small, medium, and large-v3) under both zero-shot and fine-tuning settings.

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

**Finetuned Models** Table 6 presents comparative CER results for two model families: 1) our Wenet-finetuned Paraformer-large architecture [27], and 2) a series of Whisper models [33] adapted through whisper-flamingo fine-tuning.<sup>11</sup> 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 elderly speech recognition.

<sup>10</sup> <https://github.com/openai/whisper>

<sup>11</sup> <https://github.com/roudimit/whisper-flamingo>



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

Method	MCD ( $\downarrow$ )	STOI ( $\uparrow$ )	PESQ ( $\uparrow$ )
CampNet	7.302	0.220	1.291
EditSpeech	6.225	0.514	1.363
A <sup>3</sup> T	5.851	0.586	1.455
FluentSpeech	5.811	0.627	1.645

Table 8: Subjective Evaluation results of speech editing models trained from scratch.

Method	Boundary-MOS	MOS
CampNet	1.25 $\pm$ 0.33	1.17 $\pm$ 0.27
EditSpeech	2.42 $\pm$ 0.98	1.82 $\pm$ 0.91
A <sup>3</sup> T	3.00 $\pm$ 0.89	2.04 $\pm$ 0.74
FluentSpeech	4.72 $\pm$ 0.44	4.53 $\pm$ 0.57

## 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 and Metrics

We employ the Speech Editing Toolkit<sup>12</sup> framework to implement our models, which include CampNet[34], EditSpeech[35], A<sup>3</sup>T[36], and FluentSpeech[37]. The detailed hyperparams are described in Appendix B.2.4.

For objective evaluation metrics, we utilize Mel-Cepstral Distortion (MCD) [38], Short-Time Objective Intelligibility (STOI) [39], and Perceptual Evaluation of Speech Quality (PESQ) [40], which are commonly used objective metrics for assessing speech generation quality.

For subjective evaluation metrics, we conducted comprehensive human perceptual evaluation to assess speech editing quality. We randomly selected 50 audio samples and recruited three evaluators to rate each sample on a 1-5 scale using two metrics:

- **MOS (Mean Opinion Score):** Evaluates naturalness of edited speech segments
- **Boundary-MOS:** Our specialized metric measuring transition smoothness at editing boundaries

### 4.4.2 Performance Analysis

The results presented in Table 7 and 8 indicate that FluentSpeech achieved the best performance, with all objective metrics and subjective metrics falling within acceptable ranges. In relevant research [37, 36, 34, 41], these experimental values are widely recognized as acceptable. This suggests that our dataset is suitable for generative tasks.

## 5 Limitations and Ethical Considerations

While SeniorTalk provides important data for elderly speech processing, it has notable limitations. The dataset includes 202 participants from 16 provinces but displays biases in gender, regional, and accent distributions. Additionally, the participant age range does not fully cover the super-elderly population (85+ years), and the overall dataset size remains relatively modest, which may restrict generalizability across diverse vocal characteristics. From an ethical perspective, rigorous safeguards

<sup>12</sup> <https://github.com/Zain-Jiang/Speech-Editing-Toolkit>

were implemented: informed consent was obtained, recordings were conducted in controlled environments, and all personally identifiable information was anonymized. Participants received fair compensation in line with local economic conditions. To mitigate risks associated with misuse such as malicious voice synthesis, data access is limited to verified academic researchers under a non-commercial CC BY-NC-SA 4.0 license.

## 6 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.

## Acknowledgement

This work has been supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSFChina (Grant No.62271270).

## References

- [1] Narumol Chumuang, Thittaporn Ganokratanaa, Patiyuth Pramkeaw, Mahasak Ketcham, Suvil Chomchaiya, and Worawut Yimyam. Voice-activated assistance for the elderly: Integrating speech recognition and iot. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–4. IEEE, 2024.
- [2] Angelo Scuteri and Peter M. Nilsson. 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)*, pages 35–39. Academic Press, Boston, second edition edition, 2024.
- [3] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17:127–144, 2013.
- [4] Victoria Young and Alex Mihailidis. 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, 2010.
- [5] Ajinkya Kulkarni, Anna Tokareva, Rameez Qureshi, and Miguel Couceiro. The balancing act: Unmasking and alleviating ASR biases in Portuguese. In Bharathi Raja Chakravarthi, Bharathi B, Paul Buitelaar, Thenmozhi Durairaj, György Kovács, and Miguel Ángel García Cumberas, editors, *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 31–40, St. Julian’s, Malta, March 2024. Association for Computational Linguistics.
- [6] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*, 2021.
- [7] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer’s disease in narrative speech. *Journal of Alzheimer’s disease*, 49(2):407–422, 2015.
- [8] Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. 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, 2022.

- [9] Shujie Hu, Xurong Xie, Mengzhe Geng, Zengrui Jin, Jiajun Deng, Guinan Li, Yi Wang, Mingyu Cui, Tianzi Wang, Helen Meng, and Xunying Liu. Self-supervised asr models and features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3561–3575, 2024.
- [10] Irina A Sekerina, Anna Smirnova Henriques, Aleksandra S Skorobogatova, Natalia Tyulina, Tatiana V Kachkovskaia, Svetlana Ruseishvili, and Sandra Madureira. Brazilian portuguese-russian (braporus) corpus: automatic transcription and acoustic quality of elderly speech during the covid-19 pandemic. *Linguistics Vanguard*, 9(s4):375–388, 2024.
- [11] Meiko Fukuda, Hiromitsu Nishizaki, Yurie Iribe, Ryota Nishimura, and Norihide Kitaoka. Improving speech recognition for the elderly: A new corpus of elderly Japanese speech and investigation of acoustic modeling for speech recognition. In Nicoletta Calzolari, Frédéric B  chet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H  l  ne Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6578–6585, Marseille, France, May 2020. European Language Resources Association.
- [12] Tianqi Wang, Quanlei Yan, Jingshen Pan, Feiqi Zhu, Rongfeng Su, Yi Guo, Lan Wang, and Nan Yan. Towards the speech features of early-stage dementia: Design and application of the mandarin elderly cognitive speech database. In *Interspeech 2019*, pages 4529–4533, 2019.
- [13] Hajime Orimo, Hideki Ito, Takao Suzuki, Atsushi Araki, Takayuki Hosoi, and Motoji Sawabe. Reviewing the definition of “elderly”. *Geriatrics & gerontology international*, 6(3):149–158, 2006.
- [14] Yurie Iribe, Norihide Kitaoka, and Shuhei Segawa. Development of new speech corpus for elderly japanese speech recognition. In *2015 International Conference Oriental COCOSA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSA/CASLRE)*, pages 27–31, 2015.
- [15] Meiko Fukuda, Ryota Nishimura, Hiromitsu Nishizaki, Koharu Horii, Yurie Iribe, Kazumasa Yamamoto, and Norihide Kitaoka. A new speech corpus of super-elderly japanese for acoustic modeling. *Computer Speech & Language*, 77:101424, 2023.
- [16] Valerie Hazan, Outi Tuomainen, Jeessun Kim, and Christopher Davis. elderlucid: London ucl older adults’ clear speech in interaction database, 2017.
- [17] Fr  d  ric Aman, Michel Vacher, Solange Rossato, and Fran  ois Portet. 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, 2013.
- [18] Charlene Pope and Boyd H. Davis. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161, 2011.
- [19] Simon JD Prince and James H Elder. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.
- [20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620, 2017.
- [21] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.
- [22] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.
- [23] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

- [24] 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. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026, 2020.
- [25] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.
- [26] Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.
- [27] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Interspeech 2021*, pages 4054–4058, 2021.
- [28] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040, 2020.
- [29] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. E-branchformer: Branchformer with enhanced merging for speech recognition, 2022.
- [30] Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding, 2022.
- [31] Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pages 2063–2067, 2022.
- [32] Linhao Dong and Bo Xu. 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, 2020.
- [33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- [34] Tao Wang, Jiangyan Yi, Ruibo Fu, Jianhua Tao, and Zhengqi Wen. 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, 2022.
- [35] Daxin Tan, Liqun Deng, Yu Ting Yeung, Xin Jiang, Xiao Chen, and Tan Lee. 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, 2021.
- [36] He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. A<sup>3</sup>T: 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, 17–23 Jul 2022.
- [37] Ziyue Jiang, Qian Yang, Jialong Zuo, Zhenhui Ye, Rongjie Huang, Yi Ren, and Zhou Zhao. Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11655–11671, 2023.
- [38] R. Kubichek. Mel-cestral 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, 1993.

- [39] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 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, 2010.
- [40] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 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, 2001.
- [41] Rui Liu, Jiatian Xi, Ziyue Jiang, and Haizhou Li. Fluenteditor: Text-based speech editing by considering acoustic and prosody consistency. *Proc. InterSpeech2024*, 2024.

## A Supplementary Dataset Information

Table 9: Accent Intensity Annotation Guidelines

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

### 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 3. 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.

### A.3 Dataset Accent Distribution Analysis

This section presents the distribution characteristics of accent severity levels within our dataset. The accent annotation criteria follow the standardized framework detailed in table 9. As illustrated in Table 12, the distribution exhibits distinct patterns across different severity levels:

- Light accent samples constitute the largest temporal proportion (110,419s) and account for 42,790 samples.

Table 10: Annotation Levels and Their Associated Tasks for Dataset Analysis

Annotation Level	Annotation Dimension	Associated Tasks	Representative Instances
Speaker Metadata	Demographic Age	Elderly Speech Analysis	75
	Geographic (Province)		Jiangsu, Henan
	ID Card Gender		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]

- Moderate accent demonstrates the second-largest distribution with 17,782s duration and 8,226 samples.
- Samples with no accent (5,951s/8,056) and heavy accent (370s/142) show minor representation.

The analysis reveals a pronounced concentration of data in the light accent category, which comprises over half of the total samples, followed by moderate, no, and heavy accents.

#### A.4 Annotation

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

##### A.4.1 Annotator

As shown in Table 11, 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.

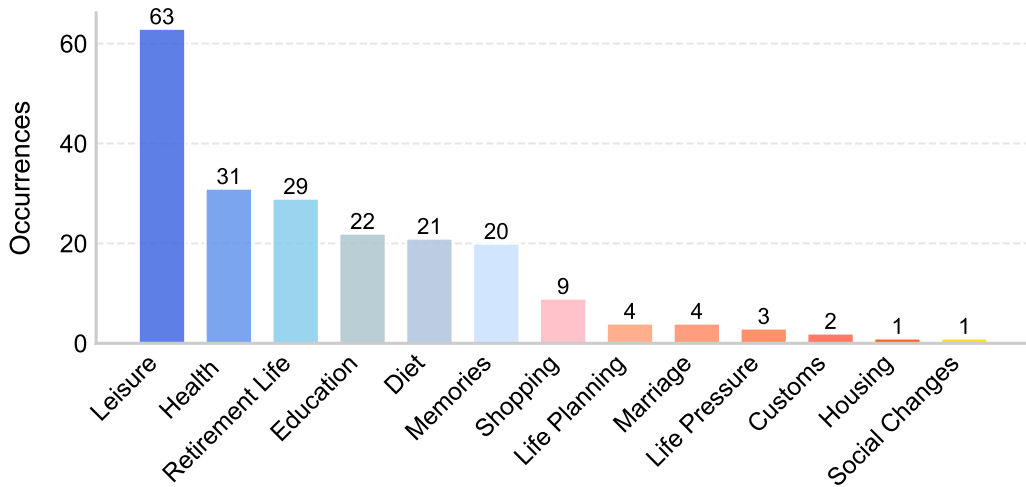


Figure 3: This distribution of topics.

#### A.4.2 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 9) 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.4.3 Annotation Information

Table 10 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, 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.

#### A.5 Licenses for assets

This work adheres to strict licensing compliance for all third-party assets. Below are the licensing terms for codebases and datasets utilized in our speech dataset creation paper:

- **WeNet:** The WeNet framework [27] for end-to-end speech recognition is licensed under Apache 2.0, permitting modification and commercial use with attribution.
- **Whisper Flamingo:** The Whisper Flamingo framework [33] for Whisper adaptation operates under BSD 3-Clause License, requiring copyright retention in derivative works.
- **SpeechBrain:** SpeechBrain toolkit [21] utilizes Apache 2.0.

Table 11: Annotator Information

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 12: Accent distribution

Accent	Total Duration (seconds)	Number of Samples
No	5,951	8,056
Light	110,419	42,790
Moderate	17,782	8,226
Heavy	370	142

- **PyAnnote:** PyAnnote toolkit[25, 26] utilizes MIT License which allows users to freely use, modify, and distribute the software, provided that the copyright notice and permission notice are retained.
- **Speech-Editing-Toolkit:** The Speech-Editing-Toolkit [37] includes no explicit software license but enforces strict usage prohibitions through appended declaration:  
“Any synthesis of recognizable voices (e.g., politicians, celebrities) without explicit consent violates ethical standards and copyright laws. Violators assume full legal liability.”
- **VoxCeleb Dataset:** Training data for speaker modeling is governed by CC BY 4.0 [20], mandating proper attribution for academic/commercial redistribution.

For reproducibility and open science, we release all components of our SeniorTalk under the following licenses:

- **SeniorTalk Dataset:** Our conversational speech dataset is released under CC BY-NC-SA 4.0, restricting commercial applications while requiring share-alike terms for derivatives.
- **Codebase:** Core infrastructure follows Apache License 2.0. Embedded third-party code retains original licensing.

## 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

Table 13: 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.

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	162	47,269	29.95	2.28
Dev	20	6,891	4.09	2.14
Test	20	5,869	3.77	2.31
Sum	202	60,029	37.81	2.27

Table 14: 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.

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	0-182	48,591	30.47	2.26
Dev	0-182	5,398	3.40	2.27
Test	182-202	6,040	3.95	2.35
Sum	0-202	60,029	37.82	2.27

### 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 table 14. 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 14. For speaker diarization, the dataset split is identical to that of the speaker verification task, except that we use dialogues instead of sentences.

Table 15: 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.

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	162	47,269	29.95	2.28
Dev	20	6,891	4.09	2.14
Test	20	5,869	3.77	2.31
Sum	202	60,029	37.81	2.27

### 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 15.

### 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 17. 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  $1e-3$  for Transformer and Conformer models, while it was  $5e-4$  for the E-Branchformer model. A warmup period of 5,000 steps was used for all models. The models were trained for a maximum of 100 epochs (Transformer) and 60 epochs (Conformer and E-Branchformer).

Table 16: Summary of dataset splits, including the utterances (# Utt.), total duration (Dur.), and average utterance length (Avg.) for speech editing tasks.

Split	# Utt.	Dur. (hrs)	Avg. (s)
Train	44,954	32.42	2.60
Test	4,994	3.68	2.65
Sum	49,948	36.10	2.60

Table 17: Hyperparameters for training ASR models from scratch.

Encoder	Accum Grad	Batch size	LR	Warmup	Epochs
Transformer	4	4	1e-3	5,000	100
Conformer	4	4	1e-3	5,000	60
E-Branchformer	4	4	5e-4	5,000	60

Table 18: Hyperparameters for fine-tuning pre-trained ASR models.

Model	Accum Grad	Batch size	Learning rate	Warmup	Training steps
Paraformer-large	4	28	5e-4	5,000	101,579
Whisper-large	1	8	5e-6	1,000	90,000
Whisper-medium	1	3	6.25e-6	1,000	225,000
Whisper-small	1	4	1.25e-5	1,000	225,000

### B.2.2 ASR Model Fine-tuning

Table 18 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 5e-6 to 5e-4. A warmup period of 1,000 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 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 19. 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 5e-3 and a base learning rate of 1e-8. The x-vector model utilized a linear learning rate schedule with an initial learning rate of 5e-3 and a base learning rate of 1e-4. All models were trained for 20 epochs.

### 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 20. 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 2e-4 for all models. A warmup period of 8,000 steps was used, and all models were trained for 200,000 steps.

## B.3 Extra Experiment Analysis

### B.3.1 Detailed results of Wenet models on ASR task

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 22 and Table 21. In Table 22, we can observe that the

Table 19: Hyperparameters for training speaker verification models.

Model	Batch size	LR Schedule	Init LR	Base LR	Epochs
ECAPA-TDNN	128	Cyclic	5e-3	1e-8	20
ResNet-TDNN	128	Cyclic	5e-3	1e-8	20
x-vector	128	Linear	5e-3	1e-4	20

Table 20: Hyperparameters for training speaker editing models.

Model	Params	Batch size	LR	Warmup	Training steps
CampNet	21.22M	16	2e-4	8,000	200,000
EditSpeech	48.15M	16	2e-4	8,000	200,000
A <sup>3</sup> T	14.86M	16	2e-4	8,000	200,000
FluentSpeech	23.86M	30	2e-4	8,000	200,000

Table 21: Model Performance in Southern Provinces

Model	Shanghai	Guangxi	Sichuan	Anhui	Zhejiang	Jiangsu	Hunan
Paraformer-fintuned	13.68	4.88	9.62	16.11	12.42	16.57	14.60
Paraformer	13.26	6.24	10.07	17.91	12.30	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	54.90

Table 22: Model Performance in Northern Provinces

Model	Beijing	Heilongjiang	Liaoning	Hebei	Henan
Paraformer-fintuned	7.38	5.85	27.08	7.71	20.93
Paraformer	6.80	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

CERs of Beijing, Hebei, and Heilongjiang are relatively low, while the recognition difficulty for Liaoning and Henan increases. Similarly, in Table 21, 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.

We have conducted further analysis on age effects. Table 23 and 24 are our experimental results for the Paraformer model before and after fine-tuning. Our analysis indicates that the difficulty of speech recognition increases significantly with speaker age, which underscores the necessity of our dataset in addressing this challenge.

### B.3.2 Results of baselines on the speaker verification task

This section evaluates three speaker verification models trained exclusively on the VoxCeleb dataset without SeniorTalk-based fine-tuning, as summarized in Table 25. After applying our proposed fine-tuning strategy using the SeniorTalk dataset, the models exhibit significant performance improvements on elderly speaker verification tasks, highlighting the critical role of domain-specific adaptation for aging voice biometrics.

## C 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

Table 23: Paraformer Model Performance across age

model	age	S	D	I	N	wer(%)	num_utterances	num_speakers
Paraformer	75	1020	217	195	14452	9.91	1278	5
	76	2008	405	451	21247	13.48	2072	7
	78	1233	263	176	8546	19.56	1010	3
	80	769	102	65	5430	17.24	602	2
	81	918	153	153	6331	19.33	621	2
	82	489	51	91	2761	22.85	286	1

Table 24: Paraformer-finetuned Model Performance

model	age	S	D	I	N	wer(%)	num_utterances	num_speakers
Paraformer-finetuned	75	906	268	231	14452	9.72	1278	5
	76	1994	520	496	21247	14.17	2072	7
	78	986	288	223	8546	17.52	1010	3
	80	734	107	85	5430	17.05	602	2
	81	749	167	170	6331	17.15	621	2
	82	407	42	91	2761	19.56	286	1

Table 25: Results of baselines on the speaker verification task

Model	PLDA EER(%)	Cosine EER(%)
x-vector	18.13	30.81
ResNet-TDNN	15.81	<b>17.50</b>
ECAPA-TDNN	<b>11.40</b>	20.88

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 while balancing scientific utility with participant protection in vulnerable populations.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We summarise limitations in Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We publicly release the source code on GitHub (<https://github.com/flageval-baai/SeniorTalk>). In Section 4, we comprehensively introduce the models employed in our study along with the associated repositories. Subsequently, in Appendix B.2, we delve into a detailed discussion of the hyperparameter settings utilized in our experiments. Moreover, Appendix B.1 is dedicated to an exploration of the dataset partitioning strategies adopted for our experimental evaluations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In abstract, we release the code and data link. In Section 4, we introduce the baseline models along with the associated repositories. Subsequently, in Appendix B.2, we delve into a detailed discussion of the hyperparameter settings utilized in our experiments. Moreover, Appendix B.1 is dedicated to an exploration of the dataset partitioning strategies adopted for our experimental evaluations.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: in Appendix B.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported in our study due to the substantial computational expense involved, particularly given the large number of experimental spacecraft we have utilized. Incorporating error bars would necessitate considerable additional time and resources, which are not feasible within the scope of our current research constraints.

Guidelines:

- The answer NA means that the paper does not include experiments.



- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Supplementary Experiments B, we conduct a detailed exploration of the experimental setup, including the hardware configuration of the machines utilized and the specific training steps.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: In Section 5, we provide a concise overview of the topic, while a comprehensive discussion is presented in Appendix C.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: in Appendix C.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: We discuss the safeguards in Appendix C and all the data and code are released under the CC BY-NC-SA 4.0 license.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: in appendix A.5

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: Compliance documentation for voice data collection ethics (including participant consent protocols) and licensing specifications are detailed in Appendix A. Implementation specifics, including hyperparameter configurations and training procedures, are provided in Supplementary Materials (Appendix B).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[Yes\]](#)

Justification: in appendix A and appendix C

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[No\]](#)

Justification: While formal IRB approval was not obtained for this study, all data collection strictly followed ethical research practices including explicit participant consent, voice data anonymization, and compliance with international privacy regulations (GDPR/CCPA). We implemented robust safeguards against potential risks while adhering to the NeurIPS Code of Ethics throughout the research process.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not use the LLM.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.