

Evaluating the Robustness of Speaker Verification Models to Age-Related Voice Changes

Idaya Seidu
Carnegie Mellon University
iseidu@andrew.cmu.edu

Heran Equbay
Carnegie Mellon University
hequbay@andrew.cmu.edu

Abstract—Speaker verification systems have achieved remarkable performance under controlled conditions, yet their long-term reliability remains threatened by natural voice changes associated with aging. This study systematically investigates the robustness of two state-of-the-art models, MFA-Conformer and ECAPA-TDNN, against age-related vocal variations. Using a subset of the VoxCeleb dataset, we propose a reproducible pipeline that applies calibrated transformations — modifying pitch, formant structure, and speech rate — to simulate realistic aging effects on speech signals. Extensive evaluations across 5,000 verification trials reveal that while MFA-Conformer outperforms ECAPA-TDNN under original conditions (2.84% vs. 3.64% Equal Error Rate), both models experience severe performance degradation when exposed to age transformations, with EER increasing by up to $7.6\times$. A detailed ablation study isolates the impact of individual transformations, showing that formant modifications cause the most significant decline in verification accuracy. Analysis of score distributions and speaker embeddings indicates that aging primarily shifts genuine scores leftward without affecting impostor distributions, leading to increased overlap and higher error rates. Our findings highlight the urgent need for age-aware speaker verification solutions to ensure the long-term reliability of biometric systems.

Keywords: speaker verification, voice aging, robustness, MFA-Conformer, ECAPA-TDNN, formant shift, embedding analysis

I. INTRODUCTION

Speaker Verification (SV) systems have rapidly evolved over the past decade, powering applications from voice assistants and smart homes to biometric authentication. Advances in deep learning have significantly improved the accuracy and scalability of speaker recognition systems [18]. Among these, text-independent SV—which allows speakers to use arbitrary content during enrollment and testing—has gained prominence due to its flexibility in real-world deployments [12].

Despite these advances, SV systems remain vulnerable to input-level factors that distort vocal signals. Demographic changes, especially those associated with aging, introduce long-term variability that models are rarely trained to accommodate. As individuals age, physiological changes in the vocal tract impact pitch, resonance, and

speaking rate [30], [13], potentially degrading verification accuracy [10], [2].

There has been some work done on modifying model architectures to address these challenges [1], [31]. While effective, such approaches often require full model retraining and may not be feasible for already-deployed systems. In contrast, controlled input-level transformations provide a scalable alternative for evaluating and improving robustness to vocal changes.

In this work, we investigate how age-related voice transformations affect the performance of state-of-the-art speaker verification models. We implement a signal-level age transformation pipeline based on established acoustic literature, simulating aging effects through pitch shifting, formant modification, speech rate adjustment, and vocal tract length perturbation. Using this pipeline, we analyze the behavior of two representative architectures—MFA-Conformer and ECAPA-TDNN—on disjoint subsets of the VoxCeleb dataset.

This paper makes the following contributions:

- We implement a controlled age transformation pipeline using techniques such as pitch shifting, formant modification, speech rate adjustment, and vocal tract length perturbation.
- We evaluate the effects of these transformations on two widely used speaker verification models: MFA-Conformer and ECAPA-TDNN.
- We conduct a detailed ablation study that isolates the impact of individual aging factors (pitch, formants, speaking rate) and their combinations.
- We analyze model behavior using standard evaluation metrics, score distributions, and embedding visualizations to study how aging affects speaker verification.
- We discuss the implications of our findings for designing systems that remain reliable across age-related voice changes.

The remainder of this paper is structured as follows: Section II reviews related work. Section III describes the model architectures and dataset setup. Section IV presents our age transformation techniques. Section V details our experiments and analysis. Section X concludes with insights and future directions.

II. LITERATURE REVIEW

Speaker verification (SV) systems are vulnerable to input distortions that can impact performance across demographic groups. This literature review focuses on three key areas relevant to aging-related vocal changes: general input normalization, aging-specific speech variability, and model architectures designed to mitigate such effects.

A. Age-Related Vocal Variability

Aging impacts the human voice through changes in pitch, formant structure, vocal tract configuration, and speech rate. Xue and Hao [30] documented consistent declines in formant frequencies and vocal tract flexibility with age. Harnsberger et al. [13] examined how these changes affect speech perception, while Jacewicz et al. [15] quantified age-related slowing of articulation.

Mousa [24] explored formant manipulation and time-stretching techniques using PSOLA to simulate age-related modifications, which serve as a foundation for synthetic transformations. These studies motivated our transformation parameters and helped simulate realistic spectral distortions reflective of decades of aging.

B. Speaker Normalization Techniques

Traditional approaches like Vocal Tract Length Normalization (VTLN) [21] and formant-based frequency warping [9] aim to reduce speaker-specific variability, making systems more robust to physiological differences. These methods have been extended through pitch shifting and resampling-based normalization, particularly useful for addressing intra-speaker spectral drift over time.

C. Noise and Feature Normalization

Noise and recording variability are also addressed through spectral subtraction [3], MMSE-STSA [6], and recent deep-learning-based denoising models like MetricGAN [8]. Mel-spectrogram normalization techniques such as CleanMel [26] further improve consistency in speaker embeddings under varying input conditions.

D. Speaker Verification Architectures

Recent model architectures explicitly attempt to improve robustness. The MFA-Conformer model extends the Conformer architecture [11] with multi-scale aggregation, enabling it to encode speaker identity across multiple levels of abstraction [31]. This design improves tolerance to distortions like formant shifts.

The ECAPA-TDNN model [4] incorporates channel attention and multi-scale temporal aggregation using time-delay neural networks. While highly efficient, it appears more sensitive to formant-based aging effects, as our findings also show.

E. Demographic Robustness

Demographic factors such as age, gender, and health status alter both production and perception of speech. Kwak et al. [19] and Herrmann and Cui [14] found that older adults require higher signal-to-noise ratios and may struggle more with AI-generated speech. Leibold and Buss [22] showed that children have greater difficulty identifying consonants in noisy environments, further emphasizing the importance of age-aware design.

F. Gaps in Current Research

Despite many contributions in normalization and architecture, few works focus specifically on longitudinal vocal changes and their implications for speaker verification over a person's lifespan. Our work addresses this gap by simulating controlled age-related transformations and analyzing their effects on modern speaker verification systems.

III. BASELINE MODELS

A. Model Description

In this study, we adopt two state-of-the-art architectures as baselines for speaker verification: the Multi-Frame Attention Conformer (MFA-Conformer) [32] and the Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) [4]. These complementary approaches allow us to investigate aging effects across different architectural paradigms.

1) *MFA-Conformer Architecture*: The MFA-Conformer architecture combines convolutional neural networks and transformer models to capture both local and global patterns in speech signals [32]. The model consists of a conformer encoder with six transformer blocks, processing mel-spectrogram features extracted from audio signals.

A key innovation of the MFA-Conformer is its multi-scale feature aggregation approach, which concatenates outputs from all encoder layers rather than using only the final layer output [32]:

$$x_{out} = \text{Concat}(x_1, x_2, \dots, x_6) \quad (1)$$

where x_i represents the output from the i -th encoder layer. This approach captures speaker characteristics at different levels of abstraction.

Temporal aggregation is handled by an Attentive Statistics Pooling layer, computing weighted statistics across the time dimension:

$$\mu = \sum_t w_t \cdot x_t, \quad \sigma = \sqrt{\sum_t w_t \cdot x_t^2 - \mu^2} \quad (2)$$

where w_t are learned attention weights. The final speaker embedding is produced through batch normalization and a fully connected layer, yielding a fixed dimensional embedding of size 192.

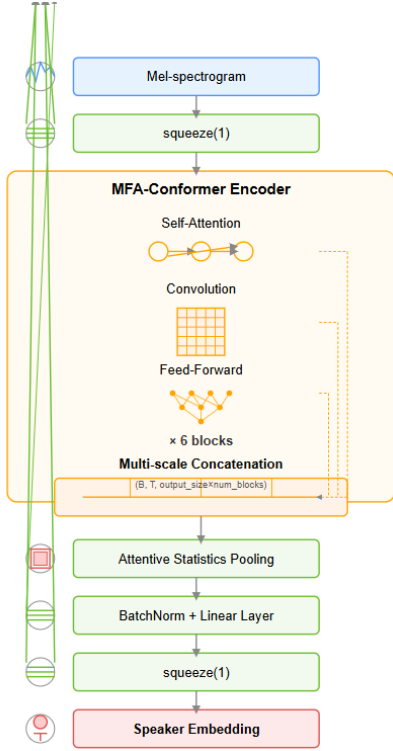


Fig. 1: MFA-Conformer architecture featuring multi-scale feature aggregation across all encoder layers. The model captures both local patterns (via convolutions) and global relationships (via self-attention) for robust speaker verification.

2) *ECAPA-TDNN Architecture*: We adopt the ECAPA-TDNN implementation from SpeechBrain, using a wrapper that:

- Takes mel-spectrogram inputs of shape (batch, 1, time, n_mels)
- Removes unnecessary dimensions through squeeze operations
- Processes features through the core ECAPA-TDNN model
- Produces fixed-dimension speaker embeddings of size 192

The implementation offers two variants: a standard version using 512 channels and a large version using 1024 channels. The ECAPA-TDNN has fewer parameters than the MFA-Conformer while still delivering competitive performance [4].

B. Training Methodology

Both baseline models were trained on the VoxCeleb dataset. We used the VoxCeleb2 development set, comprising 1,211 speakers with 148,641 utterances. The training methodology included:

- **Feature Extraction**: 80-dimensional mel-spectrograms extracted with a window length of 400 samples and hop length of 160 samples.

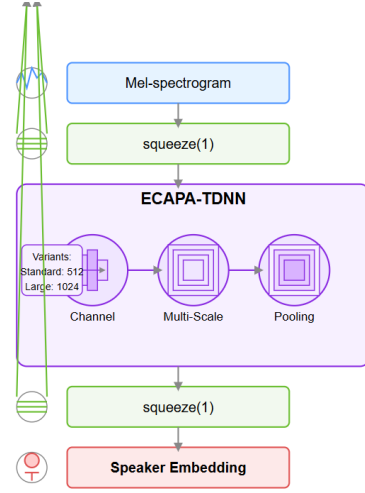


Fig. 2: ECAPA-TDNN architecture with channel attention, multi-scale aggregation, and attentive statistics pooling components. Two variants are available: standard (512 channels) and large (1024 channels).

- **Loss Function**: AM-Softmax loss with a margin of 0.2 and scale factor of 30.
- **Training Framework**: PyTorch Lightning with support for mixed-precision training and automatic checkpoint management.

C. Evaluation Criteria

The performance of the baseline models was evaluated using two primary metrics:

- **Equal Error Rate (EER)**: The threshold where the false acceptance rate equals the false rejection rate. A lower EER indicates better verification performance.
- **Minimum Detection Cost Function (minDCF)**: A weighted measure combining miss and false alarm probabilities:

$$DCF(t) = C_{miss} \cdot P_{miss}(t) \cdot P_{target} + C_{fa} \cdot P_{fa}(t) \cdot (1 - P_{target}) \quad (3)$$

where C_{miss} and C_{fa} are costs for miss and false alarm errors, and P_{target} is the prior probability of the target speaker.

The evaluation protocol uses cosine similarity between L2-normalized embeddings as the scoring metric:

$$score(e_1, e_2) = \frac{e_1 \cdot e_2}{||e_1|| \cdot ||e_2||} \quad (4)$$

where e_1 and e_2 represent speaker embeddings for the enrollment and test utterances, respectively.

D. Training Results

The MFA-Conformer model was trained for 30 epochs and validated on the VoxCeleb1 test set. Table I shows the performance metrics across notable epochs.

TABLE I: MFA-Conformer Baseline Performance Metrics

Epoch	EER (%)	minDCF(0.01)	minDCF(0.001)	Train Loss	Accuracy (%)
0	10.20	0.70	0.81	6.24	17.10
4	4.31	0.41	0.55	2.08	58.50
9	3.62	0.33	0.47	0.52	95.10
13	3.48	0.31	0.44	0.41	95.10
17	3.41	0.35	0.42	0.30	95.10
29	3.49	0.34	0.43	0.27	97.60

Figure 3 illustrates the evolution of verification performance metrics during training. The Equal Error Rate (EER) shows the most dramatic improvement in the first 10 epochs, followed by a more gradual refinement with the best performance of 3.41% achieved at epoch 17.

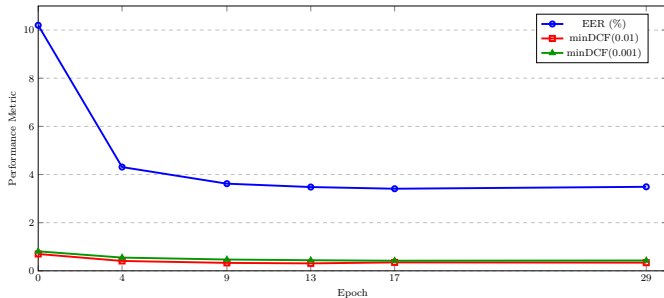


Fig. 3: Evolution of verification performance metrics for MFA-Conformer during training.

The ECAPA-TDNN model was similarly trained for 30 epochs. Table II presents the performance metrics at key epochs during training.

TABLE II: ECAPA-TDNN Performance Metrics

Epoch	EER (%)	minDCF(0.01)	minDCF(0.001)	Train Loss	Accuracy (%)
0	9.53	0.68	0.86	6.23	9.76
4	4.42	0.41	0.47	1.53	65.90
8	4.01	0.38	0.43	0.70	85.40
12	3.93	0.34	0.41	0.35	95.10
20	3.79	0.36	0.40	0.33	95.10
28	3.75	0.37	0.40	0.11	100.00

Figure 4 shows the evolution of ECAPA-TDNN performance metrics during training. Like the MFA-Conformer, it shows rapid initial improvement followed by a more gradual refinement.

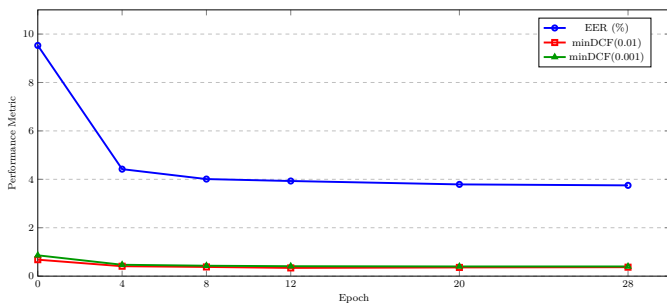


Fig. 4: Evolution of verification performance metrics for ECAPA-TDNN during training.

E. Model Comparison

A comparison between the MFA-Conformer and ECAPA-TDNN training results reveals several interesting insights:

TABLE III: Performance Comparison of MFA-Conformer vs. ECAPA-TDNN

Model	Best EER (%)	minDCF(0.01)	minDCF(0.001)	Parameters
MFA-Conformer	3.41	0.31	0.42	~20M
ECAPA-TDNN	3.75	0.34	0.40	~6.2M

As highlighted before, the MFA-Conformer achieved a slightly better EER (3.41% vs. 3.75%) and minDCF(0.01) (0.31 vs. 0.34), while the ECAPA-TDNN performed marginally better on minDCF(0.001) (0.40 vs. 0.42). Notably, the ECAPA-TDNN has approximately 70% fewer parameters than the MFA-Conformer (6.2M vs. 20M), resulting in a more compact model with faster training and inference times.

Both models showed similar training patterns with rapid initial improvement, followed by a performance plateau. The best performance for MFA-Conformer was achieved at epoch 17, while ECAPA-TDNN reached its best performance at epoch 28, suggesting potentially different convergence properties.

IV. METHODOLOGY

This section outlines our approach for evaluating speaker verification systems under age-related voice transformations. We present a systematic methodology covering dataset preparation, age transformation techniques, and evaluation protocols designed to isolate and measure the impact of vocal aging on verification performance.

A. Dataset Creation and Speaker Selection

To ensure a scientifically valid evaluation, we created a disjoint test set with speakers completely absent from the training data. This approach follows established protocol for speaker verification evaluation, where enrollment and test speakers must be strictly separated from training data to prevent latent memorization effects that could artificially inflate performance metrics [27].

The speaker selection process is mathematically formalized as:

$$S_{\text{test}} = \{s \mid s \in S_{\text{all}} \wedge s \notin S_{\text{train}} \wedge |U_s| \geq U_{\text{min}}\} \quad (5)$$

Where:

- S_{test} represents the set of selected test speakers
- S_{all} represents the set of all available speakers in the VoxCeleb1 test set
- S_{train} represents the set of speakers used during model training
- U_s denotes the set of utterances for speaker s
- U_{min} specifies the minimum required utterances per speaker (established at 10 utterances)

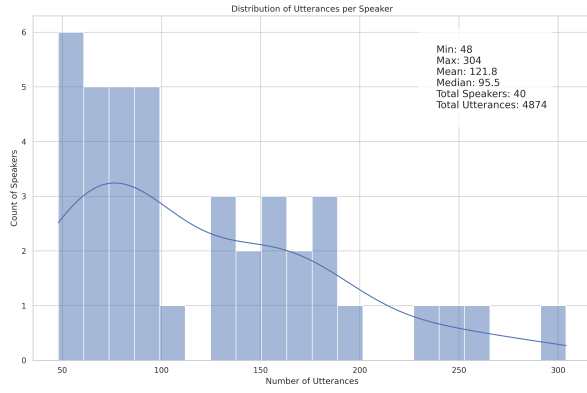


Fig. 5: Distribution of utterances per speaker in the test set.

Through this process, we targeted 45 speakers with a minimum of 10 utterances each, providing sufficient data for reliable verification testing. This approach created a controlled environment that isolates the effects of age transformation from other potential confounding variables. The speaker selection was performed deterministically with a fixed random seed (42) to ensure reproducibility across experiments. Figure 5 shows the distribution of utterances across the 40 speakers in our final test set. The histogram reveals that speakers have between 48 and 304 utterances each (mean=121.8, median=95.5), with most speakers having 50-100 utterances. In total, the test set contains 4,874 utterances distributed across these speakers. This distribution ensures sufficient data for both enrollment and testing phases while maintaining speaker diversity.

The underlying VoxCeleb dataset contains recordings from YouTube videos, primarily consisting of celebrity interviews and public speeches. While these recordings already contain some natural variation in recording conditions and speaking styles, they typically represent speakers at a single point in time rather than across significant age differences. This characteristic makes the dataset ideal for our controlled age transformation experiments, as we can introduce specific age-related modifications to otherwise consistent speaker samples.

B. Age-Related Voice Transformation Techniques

To systematically evaluate the robustness of speaker verification systems against age-related voice changes, we implemented an **AgeShiftAugmentation** module that combines multiple transformation methods to simulate both older and younger voices. Speech aging manifests through multiple acoustic dimensions simultaneously, as supported by findings from Xue and Deliyski [29], Rojas *et al.* [25], and others, showing that age-related changes affect several acoustic measures concurrently rather than independently.

Our approach synthesizes multiple transformation techniques to create realistic age-simulated speech. The transformations target four primary acoustic characteristics that demonstrate consistent change with age according to longitudinal studies [23]:

- 1) Fundamental frequency (F0) changes
- 2) Formant frequency shifts
- 3) Speaking rate alterations
- 4) Vocal tract length variations

1) *Vocal Tract Length Perturbation (VTLP)*: To simulate age-related changes in vocal tract anatomy, we applied Vocal Tract Length Perturbation (VTLP) through frequency warping [30], [16]. The warping function is defined as:

$$\hat{f} = f \cdot \alpha \quad (6)$$

where f is the original frequency, $\hat{f}$ is the warped frequency, and α is the warping factor [5], [16]. Based on physiological studies [30], [23], we used $\alpha = 0.9$ for older voices and $\alpha = 1.15$ for younger voices.

Our implementation employs time-domain processing through phase vocoding that preserves natural phase relationships between harmonics and the harmonic structure critical for speaker identity [7], while ensuring VTLP affects only frequency distribution.

2) *Pitch Shifting*: Fundamental frequency (F0), perceived as pitch, naturally decreases with age in both male and female speakers [13]. This change results from physical alterations in the vocal folds, including decreased elasticity and increased mass with age. We model this through semitone-based pitch shifting:

$$y_{\text{shifted}}(t) = \text{PitchShift}(y(t), n_{\text{steps}}) \quad (7)$$

Where $y(t)$ is the original signal, $y_{\text{shifted}}(t)$ is the pitch-shifted signal, and n_{steps} represents the number of semitones to shift. Based on longitudinal studies of voice aging [13], we calibrated our transformations using $n_{\text{steps}} = -2.0$ for older voices and $n_{\text{steps}} = +1.5$ for younger voices.

The pitch-shifting algorithm maintains the duration of the original signal through phase vocoding [20], which separates the spectral envelope from the excitation source and modifies the latter while preserving the former.

3) *Formant Shifting*: Formant frequencies systematically shift with age due to alterations in vocal tract morphology, tissue elasticity, and articulatory control [23], [30]. We implemented formant shifting through a spectral warping technique [28] that selectively modifies formant structures:

$$M_{\text{warped}}(k, l) = M(k^{\beta}, l) \quad (8)$$

Where $M(k, l)$ is the original magnitude spectrogram, $M_{\text{warped}}(k, l)$ is the warped spectrogram, and β is the shift factor calibrated based on acoustic phonetic research [23]: $\beta = 0.85$ for older voices and $\beta = 1.15$ for younger voices.

This non-linear warping function specifically targets the formant structure by applying greater shifts to higher frequency regions where upper formants reside, consistent with observed patterns in aging voices [30].

4) *Speech Rate Modification*: Speaking rate typically decreases with age due to a combination of neuromotor changes, reduced articulatory precision, and increased pause duration [15]. We implemented time-domain speech rate modification:

$$y_{\text{stretched}}(t) = y(t/\gamma) \quad (9)$$

Where γ is the speed factor calibrated based on empirical studies of age-related changes in speaking rate [15]: $\gamma = 0.9$ for older voices (10% slower) and $\gamma = 1.1$ for younger voices (10% faster).

Our speech rate modification preserves the pitch and formant structure of the original signal through phase vocoding with identity phase locking [20], which preserves phase coherence across frames for natural-sounding time-scale modification.

5) *Combined Age Transformations*: For realistic age transformations, we apply these techniques in sequence to generate composite effects that better represent the multidimensional nature of voice aging. The complete transformation pipeline is defined as:

$$y_{\text{aged}}(t) = T_{\text{speed}}(T_{\text{formant}}(T_{\text{pitch}}(y(t)))) \quad (10)$$

This sequential application ensures that each transformation builds upon the previous ones, creating a more integrated and natural simulation of age-related voice changes. The specific parameters for each transformation type were carefully calibrated based on literature to create realistic age-shifted voices while maintaining speaker identity cues.

Fig. 6 illustrates the spectral effects of these combined transformations through spectrograms of original, older, and younger voice variants of the same utterance. The visible modifications align with established acoustic correlates of aging voices documented in longitudinal studies [13], [23], including changes in frequency distribution, formant structure, and temporal characteristics.

C. Evaluation Framework

To assess the effects of age-related voice transformations on speaker verification performance, we established an evaluation framework that follows the same protocols described in Section III while enabling controlled comparison across original and transformed datasets.

1) *Trial Generation*: For each dataset (original, older voice, younger voice), we generated balanced evaluation trials using:

$$T = \{(e_i, t_j, l_{ij}) \mid e_i \in U, t_j \in U, l_{ij} \in \{0, 1\}\} \quad (11)$$

Where e_i represents an enrollment utterance, t_j represents a test utterance, and l_{ij} is the binary label (1 for

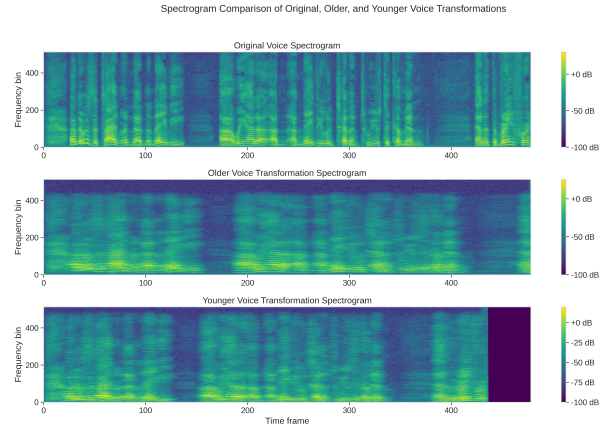


Fig. 6: Spectrograms of the same utterance comparing the original voice (top), its older transformation (middle), and its younger transformation (bottom). The older voice shows lower energy in higher frequencies, downward-shifted formants, and temporal expansion reflecting slower speech rate. The younger voice exhibits increased high-frequency energy, upward-shifted formants, and slightly compressed temporal structure.

same-speaker comparison, 0 for different-speaker comparison). We generated 5,000 trials with a 50/50 split between genuine (same-speaker) and impostor (different-speaker) trials to ensure statistical validity and balanced evaluation.

The trial generation followed a stratified sampling approach to ensure representative coverage across all speakers in the test set. For genuine trials, we randomly selected pairs of utterances from the same speaker, ensuring they were distinct recordings. For impostor trials, we selected utterances from different speakers, balancing the distribution of speaker pairings to prevent any speaker from being overrepresented.

2) *Embedding Extraction and Scoring*: For both verification models (MFA-Conformer and ECAPA-TDNN), we extracted speaker embeddings following a consistent protocol:

$$\mathbf{e} = f_{\theta}(x) \quad (12)$$

Where x represents the input utterance, f_{θ} represents the speaker embedding function with learned parameters θ , and $\mathbf{e}$ represents the resulting speaker embedding vector of dimension 192.

The embedding extraction process began with converting the audio to mono and resampling to 16 kHz when necessary. We then extracted 80-dimensional mel-spectrogram features with a window length of 25ms and hop length of 10ms, consistent with the preprocessing used during model training. These features were passed through the neural network model to obtain the speaker embedding.

Following the standard protocol described in Section III, we applied L2-normalization to the embeddings. Addition-

TABLE IV: Performance Comparison Across Models and Datasets

Model	Dataset	EER (%)	minDCF (0.01)	minDCF (0.001)
MFA-Conformer	Original	2.84	0.3252	0.5224
	Older Voice	18.88	0.9188	0.9188
	Younger Voice	21.48	0.9724	0.9724
ECAPA-TDNN	Original	3.64	0.3160	0.4004
	Older Voice	20.72	0.9304	0.9304
	Younger Voice	23.08	0.9584	0.9584

ally, we implemented center-normalization by subtracting the mean embedding computed across all test utterances:

$$\hat{e}_i = e_i - \frac{1}{N} \sum_{j=1}^N e_j \quad (13)$$

This centering operation compensates for dataset bias and has been shown to improve verification performance by removing common patterns that do not contribute to speaker discrimination [17]. After centering, the embeddings were re-normalized to maintain unit length.

We employed the same scoring method and evaluation metrics described in Section III. Using these consistent protocols allows direct comparison between our baseline performance and results under age transformations.

The following section presents our experimental results, beginning with performance metrics, followed by detailed analysis of score distributions to explain the observed degradation patterns.

V. RESULTS AND ANALYSIS

This section presents the quantitative results of our evaluation and analyzes the patterns of performance degradation introduced by age-related voice transformations.

A. Performance Metrics

Table IV summarizes the verification performance of both the MFA-Conformer and ECAPA-TDNN models across the three datasets: original, older-voice, and younger-voice variants.

The results reveal some notable trends:

- 1) **Baseline performance:** Both models perform well on the original dataset, with MFA-Conformer achieving slightly better results (EER of 2.84% vs. 3.64% for ECAPA-TDNN).
- 2) **Impact of age transformations:** Both models experience substantial degradation when evaluated on age-transformed voices, with EER increasing by approximately 6-7 times compared to baseline performance.
- 3) **Older vs. younger transformations:** For both models, younger voice transformations cause slightly more degradation than older voice transformations. MFA-Conformer’s EER increases from 18.88% (older) to 21.48% (younger), while ECAPA-TDNN’s EER rises from 20.72% (older) to 23.08% (younger).

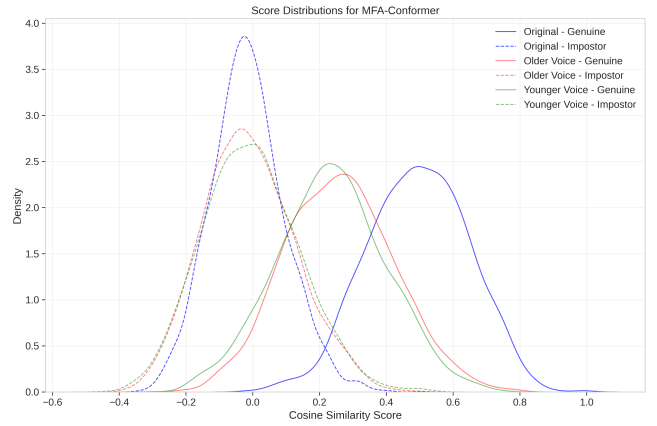


Fig. 7: Score distributions for MFA-Conformer across original, older-voice, and younger-voice datasets. Solid lines represent genuine (same-speaker) score distributions, while dashed lines represent impostor (different-speaker) distributions.

- 4) **rise in minDCF:** Both models show minDCF scores approaching 1.0 under age transformations, suggesting a near-complete inability to distinguish same-speaker from impostor trials under spectral distortions.
- 5) **Consistency across models:** The similar degradation patterns across both architectures indicates that vulnerability to age-related voice changes is a fundamental challenge in speaker verification rather than a model-specific limitation.

B. Score Distribution Analysis

To understand the mechanism behind this performance degradation, we analyzed the distributions of similarity scores for genuine (same-speaker) and impostor (different-speaker) comparisons across all three datasets.

The score distributions were estimated using kernel density estimation with a Gaussian kernel and Scott’s rule for bandwidth selection. Fig. 7 and Fig. 8 show the score distributions for MFA-Conformer and ECAPA-TDNN models, respectively.

These distributions provide critical insights into the mechanisms of performance degradation:

- For both models, age transformations cause genuine score distributions to shift dramatically toward lower similarity values, indicating transformed voices are perceived as different from original utterances.
- The increased overlap between genuine and impostor distributions directly causes the observed rise in error rates. The optimal decision threshold becomes less effective in discriminating genuine from impostor trials.
- Genuine score distributions become wider under age transformation, indicating greater inconsistency in speaker representation.

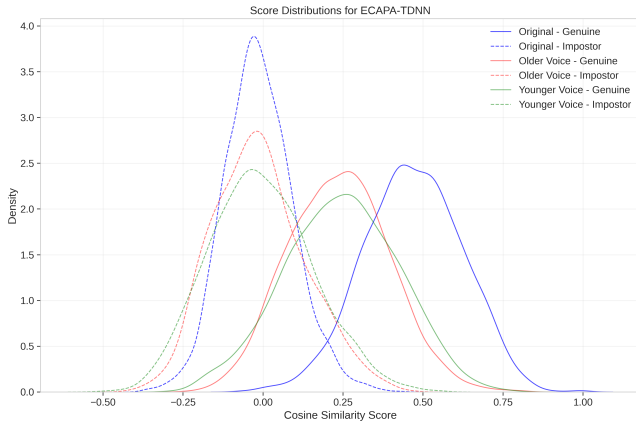


Fig. 8: Score distributions for ECAPA-TDNN across original, older-voice, and younger-voice datasets.

- Both models show similar patterns of degradation, with substantial shifts in their genuine distributions for both older and younger voice transformations.

This analysis confirms that verification degradation under age simulation is primarily driven by reduced similarity between original and transformed utterances from the same speaker, rather than confusion between different speakers.

C. DET Curve Analysis

To further illustrate the problem identified in the score distributions, we examine the Detection Error Tradeoff (DET) curves that provide a detailed view of verification performance across different operating thresholds. Fig. 11 shows DET curves for both models across all datasets.

These curves confirm and extend our previous findings. Both models show significantly better performance on original voices (blue curves) compared to age-transformed voices. The rightward and upward shift of the curves for older voice (red) and younger voice (green) transformations indicates higher error rates at all operating points. The similarity between DET curves for both models suggests that the fundamental challenge of age-related voice transformations affects different architectures in comparable ways. Performance degradation is consistent across the entire operating range, not just at the EER operating point, indicating that threshold adjustment alone cannot compensate for age-related changes. This analysis further confirms that age-related voice transformations fundamentally alter the verification task difficulty, affecting the entire error-rate trade-off curve rather than just specific operating points.

D. Embedding Space Visualization

To gain deeper insights into how age transformations affect speaker representations, we visualized the high-dimensional speaker embeddings using t-SNE. We applied a two-stage dimensionality reduction approach, first using

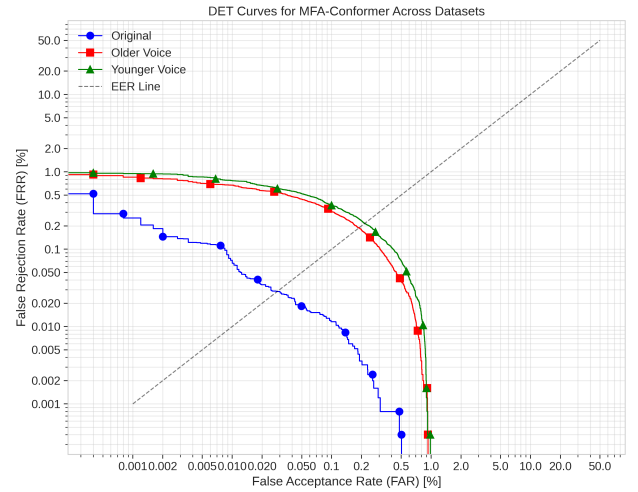


Fig. 9: MFA-Conformer

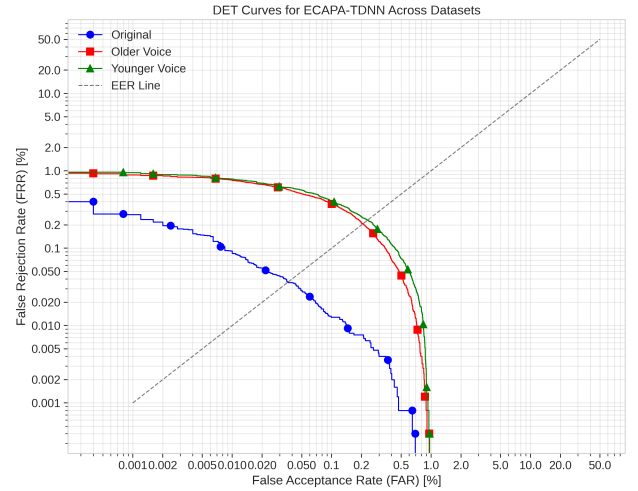


Fig. 10: ECAPA-TDNN

Fig. 11: DET curves for both models across datasets. Lower curves indicate better performance.

PCA to reduce to 50 dimensions (preserving approximately 95% of variance), followed by t-SNE for final reduction to 2 dimensions for visualization.

The embedding visualizations reveal striking differences between original and age-transformed datasets. In the original dataset, both models form distinct, tightly clustered speaker groups with clear separation between different speakers. This explains their strong verification performance on unmodified voices.

In contrast, the age-transformed datasets show significantly disrupted clustering. Speaker points are scattered across the embedding space, with many clusters fragmented into multiple subclusters. The boundaries between different speakers become less defined, with considerable overlap between speaker representations. This visual ev-



Fig. 12: MFA-Conformer (Original)

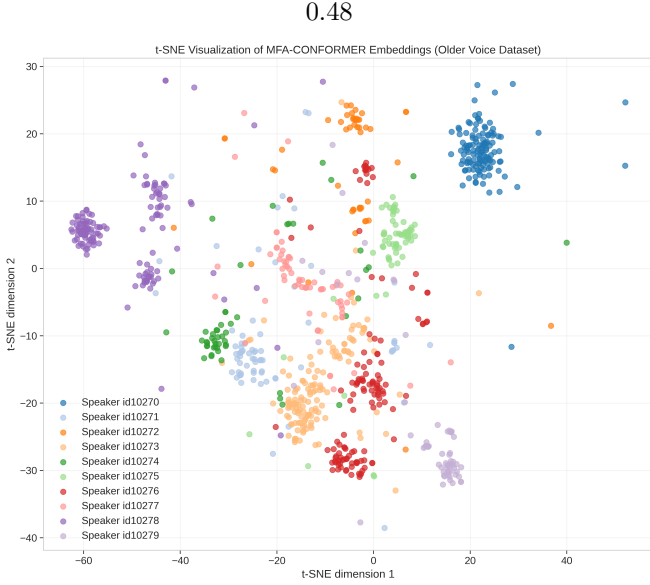


Fig. 13: MFA-Conformer (Older Voice)

Fig. 14: t-SNE visualization of speaker embeddings across datasets. Each color represents a different speaker. Original datasets (top) show clear speaker clustering, while age-transformed datasets (middle, bottom) show disrupted clustering structure.

TABLE V: Ablation Study Results Across Voice Aging Factors

Model	Transformation	EER (%)	minDCF(0.01)	minDCF(0.001)
MFA-Conformer	Older Voice: Pitch Only	7.38	0.4702	0.6137
	Older Voice: Formant Only	9.25	0.4492	0.4762
	Older Voice: Speed Only	5.13	0.3247	0.3262
	Older Voice: Pitch+Formant	11.50	0.6954	0.7312
	Older Voice: Pitch+Speed	9.17	0.5667	0.6538
	Older Voice: Formant+Speed	10.64	0.6302	0.6875
MFA-Conformer	Older Voice: Full	12.38	0.6737	0.6737
	Younger Voice: Pitch Only	6.25	0.4263	0.4263
	Younger Voice: Formant Only	8.50	0.4517	0.4575
	Younger Voice: Speed Only	10.00	0.5942	0.6062
	Younger Voice: Pitch+Formant	9.10	0.5375	0.5375
	Younger Voice: Pitch+Speed	13.75	0.8238	0.8238
ECAPA-TDNN	Younger Voice: Formant+Speed	17.88	0.8522	0.8600
	Younger Voice: Full	16.75	0.9560	0.9650
ECAPA-TDNN	Older Voice: Pitch Only	7.92	0.5852	0.6712
	Older Voice: Formant Only	22.38	0.8863	0.8863
	Older Voice: Speed Only	6.50	0.4125	0.4125
	Older Voice: Pitch+Formant	34.36	0.9912	0.9912
	Older Voice: Pitch+Speed	11.75	0.7772	0.8888
	Older Voice: Formant+Speed	10.88	0.6944	0.7150
ECAPA-TDNN	Older Voice: Full	14.84	0.7230	0.7862
	Younger Voice: Pitch Only	7.25	0.5492	0.7275
	Younger Voice: Formant Only	19.75	0.8697	0.9600
	Younger Voice: Speed Only	11.88	0.6835	0.7225
	Younger Voice: Pitch+Formant	27.00	0.9830	0.9887
	Younger Voice: Pitch+Speed	15.25	0.9212	0.9212
ECAPA-TDNN	Younger Voice: Formant+Speed	18.21	0.8867	0.9800
	Younger Voice: Full	18.14	0.8925	0.8925

idence aligns with our previous observations from score distributions, showing that age transformations fundamentally alter how speaker verification models represent voice identity.

VI. ABLATION STUDY ON VOICE AGING FACTORS

Building on our embedding space analysis, we conducted an ablation study to isolate the effects of individual and combined voice aging factors. We evaluated three primary factors that change with age: pitch (fundamental frequency), formants (vocal tract resonances), and speaking rate. By applying these factors individually and in combinations, we identified their relative contributions to verification performance degradation.

Table V presents the verification performance for both models across all ablation conditions.

The ablation results reveal these key insights:

For MFA-Conformer, changes in speaking rate (speed) caused the least degradation for older voice simulations (5.13% EER) but more substantial effects for younger voice simulations (10.00% EER). Formant shifts consistently affected performance for both older and younger voice transformations. The combination of formant and speed modifications had the most severe impact on younger voice verification (17.88% EER).

ECAPA-TDNN showed high sensitivity to formant modifications, with formant-only transformations causing substantial degradation (22.38% EER for older voices, 19.75% EER for younger voices). The combination of pitch and formant modifications produced the worst performance (34.36% EER for older voices, 27.00% EER for younger voices).

Interestingly, for both models, combinations of voice aging factors did not always result in worse performance than all three factors combined. This non-linear relationship suggests complex interactions between different vocal

aging characteristics that verification systems struggle to accommodate.

VII. DISCUSSION

Our findings reveal that even state-of-the-art speaker verification models remain vulnerable to realistic spectral shifts introduced by aging, despite their strong baseline performance on unaltered voices. This vulnerability highlights a fundamental limitation of current deep speaker embedding systems: they rely heavily on relatively rigid spectral patterns, particularly vocal tract resonances, that are not invariant over a speaker’s lifespan.

The severe degradation associated with formant shifts suggests that models are overly sensitive to the absolute positioning of spectral features rather than learning more abstract speaker representations. This dependence may explain why genuine scores shift significantly while impostor scores remain stable, as aging primarily affects intraspeaker consistency without necessarily making speakers more similar to others.

From a practical standpoint, these findings raise concerns for long-term biometric deployments in real-world applications such as secure authentication, forensic speaker identification, and continuous voice-based monitoring systems. Without compensating mechanisms, speaker verification systems may silently degrade in accuracy as users age, potentially leading to unfair false rejections for elderly populations.

Moreover, our results suggest that efforts in speaker verification research should move beyond simply improving short-term benchmark metrics. Robustness to natural human variability over time — including aging, illness, and emotional state — must become a first-class objective. This shift may require rethinking not just model architectures, but also training paradigms, evaluation protocols, and dataset collection strategies.

Essentially, our study provides evidence that addressing age-related vocal shifts is essential for building speaker verification systems that are both reliable and inclusive over the full human lifespan.

VIII. LIMITATIONS

While our aging transformation pipeline was carefully calibrated using established acoustic and physiological studies, it remains a simplification of natural aging. Real-world vocal aging is gradual, nonlinear, and highly speaker-specific, influenced by factors such as gender, health, and lifestyle. In contrast, we applied fixed transformation parameters uniformly across all speakers to maintain control and comparability. This uniformity omits the natural variability seen in actual longitudinal aging.

Our transformations were not validated using real longitudinal speech data — such as recordings of the same individuals over decades — due to the scarcity of such datasets. Additionally, we did not conduct perceptual

evaluations to assess whether listeners perceive the transformed voices as realistically older or younger, which limits our ability to confirm their ecological validity.

We also applied the transformations sequentially (pitch → formant → speech rate) without investigating whether the order of application introduces interaction effects. While our ablation study isolated the individual and combined impacts of aging factors, it did not explore non-linear dependencies or cumulative distortions from transformation sequencing.

Our analysis emphasized aggregate model performance. We did not investigate whether certain speaker characteristics (e.g., naturally higher formants) predict greater resilience or vulnerability to aging effects. Embedding visualizations did suggest that aging causes cluster fragmentation in some speakers more than others, but these observations were not quantitatively evaluated.

Finally, our evaluation reflects a snapshot of performance under discrete aging shifts rather than a dynamic simulation of long-term aging. Without year-by-year simulations or progressive updates, our framework does not capture how models might behave in long-term deployments with evolving user voices.

IX. FUTURE WORK

To improve ecological validity, future work should validate simulated aging against real longitudinal datasets, where speaker recordings span decades. This would allow direct comparison between synthetic and real aging effects and offer more precise calibration of transformation parameters. In addition, listener studies could assess the perceived age of transformed voices, ensuring perceptual consistency.

We also propose developing probabilistic aging models that generate speaker-specific transformations based on metadata such as age, gender, or vocal baseline. This would allow the simulation of more realistic and diverse aging trajectories, enabling fine-grained studies of model sensitivity.

Exploring aging as a gradual process — via year-by-year simulations — would help determine whether verification accuracy degrades linearly or exhibits thresholds. Similarly, analyzing whether the order of transformations (pitch, formants, speed) affects outcome could reveal subtle interaction effects between acoustic dimensions.

To improve age robustness, we recommend the use of age-augmented training datasets (either synthetic or real), speaker-specific adaptation through progressive enrollment updates, and training models with age-invariant embedding constraints via adversarial learning. Normalization techniques that preserve identity while removing age-related spectral changes also merit exploration.

Architecturally, our results suggest that attention-based and multi-scale designs (e.g., MFA-Conformer) are more robust to vocal resonance shifts. Future work could explore the application of architectures like E-Branchformer,

which combines self-attention and convolutional branches in a lightweight and modular design. This would test whether the robustness trends generalize across similar architectures.

Finally, we recommend expanding future evaluations to diverse demographics, including children and elderly speakers, and across multiple languages. Understanding which speaker traits predict robustness — such as formant ranges, speaking styles, or gender — could help develop personalized or adaptive verification systems.

X. CONCLUSION

In this study, we evaluated the impact of simulated aging effects on speaker verification performance. Using a controlled transformation pipeline grounded in physiological studies, we showed that age-related vocal changes cause substantial degradation in both MFA-Conformer and ECAPA-TDNN models, with Equal Error Rates increasing by up to 7.6 times.

Our ablation study revealed that formant modifications had the greatest individual impact on verification performance, identifying vocal resonance shifts as a key challenge. Score distribution and embedding analyses further confirmed that aging primarily disrupts intra-speaker similarity without affecting impostor separation, making traditional thresholding approaches less effective.

While our aging simulation approach has limitations, it provides important insights into the specific vulnerabilities of current verification models. Our findings emphasize the need for new architectures, training strategies, and evaluation frameworks that explicitly address speaker variability over time.

Ensuring long-term robustness to vocal aging is critical for maintaining the reliability, fairness, and security of speaker verification technologies as they are increasingly deployed in real-world applications.

GitHub: https://github.com/HeranSemere/mfa_conformer_giv

AUTHORS' CONTRIBUTIONS

Idaya Seidu implemented the age transformation pipeline, created the test datasets, and conducted experiments with the MFA-Conformer baseline model. She performed inference on both original and age-transformed (older and younger voice) datasets, generated the score distributions, embeddings visualizations, and DET curves. Idaya also designed and conducted the ablation studies on both models and documented the methodology and results.

Heran Equbay implemented and ran the ECAPA-TDNN baseline model, conducted inference experiments, and tested the model's performance on the age-transformed datasets created by Idaya. Heran documented the ECAPA-TDNN results across all experimental conditions and contributed to the analysis of model performance differences.

Both authors contributed to the report writing, with Idaya focusing on the methodology, results, and ablation studies, while Heran focused on the introduction, literature review, and comparative analysis between the two architectures.

This work was supervised by Massa Baali.

REFERENCES

- [1] Massa Baali, Abdulhamid Aldoobi, Hira Dharmyal, Rita Singh, and Bhiksha Raj. PDAF: A Phonetic Debiasing Attention Framework for speaker verification. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1209–1216, 12 2024.
- [2] Mehdi Bahari, Tali Hadad, and Hugo Van Hamme. Demographic-aware speaker verification using fairness-regularized adaptive learning. In *Proceedings of INTERSPEECH*, pages 3335–3339, 2021.
- [3] Steven F Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.
- [4] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proceedings of INTERSPEECH*, pages 3830–3834, 2020.
- [5] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 346–348 vol. 1, 1996.
- [6] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.
- [7] Daniel Erro, Eva Navas, and Ináki Hernáez. Voice conversion versus speaker identification: analysis of objective and subjective speaker similarity. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1272–1284, 2013.
- [8] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. *Proceedings of the 36th International Conference on Machine Learning*, pages 2031–2040, 2019.
- [9] E. Gouvea and R. M. Stern. Formant-based frequency warping for speaker normalization. In *Proceedings of Eurospeech*, pages 1139–1142, 1997.
- [10] Joanna Grzybowska and Sylwia Kacprzak. Speaker age classification and regression using i-vectors. In *Proceedings of INTERSPEECH*, pages 1402–1406, 2016.
- [11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of INTERSPEECH*, pages 5036–5040, 2020.
- [12] John HL Hansen and Taufiq Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99, 2015.
- [13] James D Harnsberger, Rahul Shrivastav, W S Brown Jr, Howard Rothman, and Harry Hollien. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, 22(1):58–69, 2008.
- [14] B. Herrmann and X. Cui. Impaired prosodic processing rather than hearing function is associated with reduced recognition of ai-generated speech in older adults. *Neurobiology of Aging*, 125:12–20, 2024.
- [15] Ewa Jacewicz, Robert Allen Fox, Caitlin O'Neill, and Joseph Salmons. Articulation rate across dialect, age, and gender. *Language variation and change*, 21(2):233–256, 2009.
- [16] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, volume 31, 2018.
- [17] Patrick Kenny. Bayesian speaker verification with heavy-tailed priors. In *Odyssey*, volume 14, 2010.

- [18] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.
- [19] H. Kwak, W. Han, and S. Lee. Effect of aging on speech recognition in noise. *Journal of Audiology & Otology*, 22(4):183–189, 2018.
- [20] Jean Larocche and Mark Dolson. Improved phase vocoder time-scale modification of audio. In *IEEE Transactions on Speech and Audio Processing*, volume 7, pages 323–332. IEEE, 1999.
- [21] L. Lee and R. C. Rose. Vocal tract length normalization for large vocabulary continuous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 353–356, 1996.
- [22] Lori J Leibold and Emily Buss. Children’s identification of consonants in a speech-shaped noise or a two-talker masker. *The Journal of the Acoustical Society of America*, 133(3):1586–1596, 2013.
- [23] Sue Ellen Linville. *Vocal Aging*. Singular/Thomson Learning, San Diego, CA, 2001. Print.
- [24] Ahmed E. D. Mousa. Voice conversion using psola and resampling for speaker normalization. *International Journal of Speech Technology*, 14(3):223–232, 2011.
- [25] S. Rojas, E. Kefalianos, and A. P. Vogel. How does our voice change as we age? a systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research*, 63(2):533–551, 2020.
- [26] Nian Shao, Rui Zhou, Pengyu Wang, Xian Li, Ying Fang, Yujie Yang, and Xiaofei Li. Cleanmel: Mel-spectrogram enhancement for improving both speech quality and asr, 2025.
- [27] 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, 2018.
- [28] D. Sundermann, H. Ney, and H. Hoge. Vtln-based cross-language voice conversion. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 676–681, 2003.
- [29] S. A. Xue and D. D. Deliyski. Effects of aging on selected acoustic voice parameters: Preliminary normative data. *Educational Gerontology*, 27(2):159–168, 2001.
- [30] Steve An Xue and Grace Jianping Hao. Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study. *Journal of Speech, Language, and Hearing Research*, 46(3):689–701, 2003.
- [31] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-Yi Lee, and Helen Meng. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification. In *Proceedings of INTERSPEECH*, pages 3553–3557, 2022.
- [32] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-Yi Lee, and Helen Meng. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification. In *Proceedings of INTERSPEECH*, pages 3553–3557, 2022.