applied sciences

Article

Age Prediction from Korean Speech Data Using Neural Networks with Diverse Voice Features

Hayeon Ku [1], Jiho Lee [2], Minseo Lee [1], Seulgi Kim [1] and Janghyeok Yoon [1,*]

[1] Department of Industrial Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea; hayean2000@konkuk.ac.kr (H.K.); kiyomi103@konkuk.ac.kr (M.L.); tmfrl716@konkuk.ac.kr (S.K.)

[2] Neopons Inc., 465 Dongdaegu-ro, Dong-gu, Daegu 41260, Republic of Korea; jiholee255@neopons.com

* Correspondence: janghyoon@konkuk.ac.kr

Abstract: A person's voice serves as an indicator of age, as it changes with anatomical and physiological influences throughout their life. Although age prediction is a subject of interest across various disciplines, age-prediction studies using Korean voices are limited. The few studies that have been conducted have limitations, such as the absence of specific age groups or detailed age categories. Therefore, this study proposes an optimal combination of speech features and deep-learning models to recognize detailed age groups using a large Korean-speech dataset. From the speech dataset, recorded by individuals ranging from their teens to their 50s, four speech features were extracted: the Mel spectrogram, log-Mel spectrogram, Mel-frequency cepstral coefficients (MFCCs), and ΔMFCCs. Using these speech features, four deep-learning models were trained: ResNet-50, 1D-CNN, 2D-CNN, and a vision transformer. A performance comparison of speech feature-extraction methods and models indicated that MFCCs + ΔMFCCs was the best for both sexes when trained on the 1D-CNN model; it achieved an accuracy of 88.16% for males and 81.95% for females. The results of this study are expected to contribute to the future development of Korean speaker-recognition systems.

Keywords: age prediction; speaker recognition; voice feature extraction; convolutional neural network; vision transformer

Academic Editor: Dongsuk Yook

Received: 23 December 2024
Revised: 24 January 2025
Accepted: 26 January 2025
Published: 27 January 2025

Citation: Ku, H.; Lee, J.; Lee, M.; Kim, S.; Yoon, J. Age Prediction from Korean Speech Data Using Neural Networks with Diverse Voice Features. *Appl. Sci.* **2025**, *15*, 1337. https://doi.org/10.3390/app15031337

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A person's voice is a behavioral characteristic used for identification and authentication. It is often termed a "voiceprint," owing to its variability in response to the speaker's pronunciation, intonation, and speech patterns. These vocal characteristics change throughout a person's life, owing to anatomical and physiological factors, making the human voice a reliable indicator of age [1]. Age, a significant personal attribute and valuable signal, has garnered attention in various fields, including biometrics and computer vision [2].

Recognizing a speaker's age from their voice has applications across multiple domains, including commercial services, forensic investigations, and healthcare. For instance, in commercial services, call-center systems can be designed to identify a caller's age and match them with the most suitable agent. Additionally, by recognizing a customer's age and sex, services can be enhanced through more targeted advertising and marketing strategies [3]. In forensic investigations for crimes such as kidnapping and blackmail, estimating the age of a speaker from voice recordings can assist in identifying the perpetrator [4]. In healthcare, voice analysis can be used to estimate a person's vocal age compared to their physical age. However, estimating a speaker's age from short-term utterances remains challenging and interest in addressing this issue has recently increased.

Early research on speaker age estimation set the foundation for modern approaches. Bahari, McLaren [5] applied an i-vector-based method, demonstrating the feasibility of automated solutions. Subsequently, OSMAN, Büyük [6] examined the effects of utterance length and the number of frames on least-squares support-vector regression models, highlighting how speech duration and feature granularity affect predictive accuracy. At around the same time, A Badr and K Abdul-Hassan [7] introduced a bidirectional gated recurrent neural network (G-RNN) approach that leveraged various features—such as Mel-frequency cepstral coefficients (MFCCs), spectral subband centroids (SSCs), linear-predictive coefficients (LPCs), and formants—to capture the frequency-sensitive elements of speech. Li, Han [8] further advanced these techniques by demonstrating how acoustic and prosodic feature fusion can bolster model performance.

As deep-learning techniques gained momentum, Avikal, Sharma [9] used linear-prediction cepstral coefficients (LPCCs) and Gaussian mixture models (GMMs) to group ages in five-year increments between 5 and 50. Focusing specifically on Korean speech, So, You [10] proposed a deep artificial neural network trained on MFCCs, achieving notable accuracy gains for men in their 20s, 30s, and 50s, as well as for women in their 20s, 40s, and 50s. Further refinements emerged through transformer-based architectures and self-supervised methods, as seen in the work of Gupta, Truong [11] and Burkhardt, Wagner [12], both of which reported enhanced predictive accuracy via bi-encoder transformer models and robust speech representations.

The most recent wave of research has explored specialized convolutional neural networks (CNNs) and attention mechanisms. Tursunov, Mustaqeem [3] introduced a CNN with a multi-attention module (MAM) on speech spectrograms generated via short-time Fourier transform (STFT), showing that accurate age classification is possible, even across multiple languages and varying age brackets. Truong, Anh [13] compared seven self-supervised learning (SSL) models for joint age estimation and gender classification on the TIMIT corpus and demonstrated that an attention-based prediction model outperformed wav2vec 2.0 in both clean and 5dB signal-to-noisy conditions, achieving more robust and accurate speech representations. However, several studies, including those of Grzybowska and Kacprzak [1] and Kalluri, Vijayasenan [14], emphasize that domain-specific adaptations—particularly the integration of diverse feature sets—are crucial for capturing the full breadth of language- and culture-specific vocal nuances.

However, these studies have several limitations that create significant practical challenges. First, studies using non-Korean datasets cannot be directly applied to Korean speakers because of language-specific vocal characteristics. The unique features of Korean speech, such as pitch patterns and consonant tensing, significantly affect age-related voice characteristics, leading to poor performance when models trained on other languages are used. Second, the few existing studies on Korean speech either lack comprehensive age coverage or use broad and imprecise age categories that limit practical applications. This makes it difficult to provide accurate age-specific services, such as precise content recommendations or customer-service matching.

Finally, most studies have relied on single-feature extraction methods or limited model architectures, potentially missing important vocal characteristics specific to Korean speakers. This approach fails to capture the complex interactions between various acoustic features unique to Korean age-related speech patterns, resulting in suboptimal real-world performance.

To address these limitations, this study proposes a comprehensive framework that is specifically designed to overcome each challenge. For a robust analysis of Korean speech characteristics, we utilized a large-scale dataset, comprising approximately 2200 speech samples per age, collected from various regions of Korea to ensure a balanced representation across different dialects and speaking styles. To enable precise age-specific services, we implemented fine-grained classification in five-year increments across the ages of 11–59, resulting in 10 distinct age groups.

To capture various age-related voice characteristics, we employed complementary feature-extraction methods (Mel spectrogram and log-Mel spectrogram for frequency-domain analysis, MFCCs and ΔMFCCs for temporal dynamics) and analyzed the characteristics using various deep-learning architectures (ResNet-50, 1D-CNN, 2D-CNN, and a vision transformer). Notably, we trained all models from scratch rather than using pre-trained models because existing pre-trained models optimized for general speech or other languages may not effectively capture unique Korean phonological characteristics [15,16].

This study makes three major contributions to existing literature. First, we established an effective framework for Korean speech-based age prediction by systematically evaluating various combinations of voice features and deep-learning models. Our experiments showed that the combination of MFCCs + ΔMFCCs features with a 1D-CNN architecture achieves the best performance for both sexes. Second, we demonstrated superior classification accuracy (88.16% for males and 81.95% for females), while using more detailed age groups than in previous studies, enabling more precise age-sensitive applications, such as personalized customer-service matching. Finally, through a detailed performance analysis, we identified specific challenges in age prediction, particularly in distinguishing voices in the 30–39 age range, providing crucial insights for future research in Korean speech processing.

The remainder of this paper is organized as follows: Section 2 describes the dataset, data preprocessing, and feature-extraction methods used in this study. Section 3 provides an overview of the neural network architectures employed in the experiments. Section 4 presents the evaluation metrics, detailed results, and comparative analyses for age prediction using voice features and neural networks. Section 5 concludes the paper with a summary and suggestions for future research.

2. Data and Feature Extraction

2.1. Data Preparation

In this study, a conversational speech dataset provided by the AI-Hub (https://www.aihub.or.kr, accessed on 1 December 2024) of the National Information Society Agency (NIA) of Korea was used. The dataset comprised recordings from 2547 speakers, both male and female, aged 11–59. Data were collected from various regions of Korea to capture regional accents. The recordings were collected through both online voice chat systems and offline studio environments, following a standardized collection protocol. The audio data were stored in PCM WAV format at a 16 kHz sampling rate. For recordings originally made at 44 kHz, downsampling to 16 kHz was performed to maintain consistency across the dataset. Quality assurance was maintained through a three-stage verification process including worker review, manager inspection, and final validation checks to ensure data integrity.

To prevent age-biased training, datasets for males and females were constructed by randomly selecting approximately 2200 speech samples for each age. As a result, approximately 11,000 speech samples were obtained for each sex in each 5-year age group (e.g., 20–24, 25–29), forming 10 distinct groups spanning ages 11–59. The data were trimmed to remove silent segments that fell below the typical background noise levels to prevent silence in the audio data from affecting the audio quality and model performance. Sounds under 60 dB—such as whispers or quiet conversations—were considered background noise and excluded to ensure that only meaningful audio was retained in each clip [17].

Only the first three seconds of the trimmed audio data were used for the experiments. Audio samples that were shorter than three seconds after trimming were excluded to ensure that all speech data samples utilized in this study were exactly three seconds long. The speech data preprocessing is illustrated in Figure 1, and the detailed compositions of the datasets used in this study are listed in Table 1.

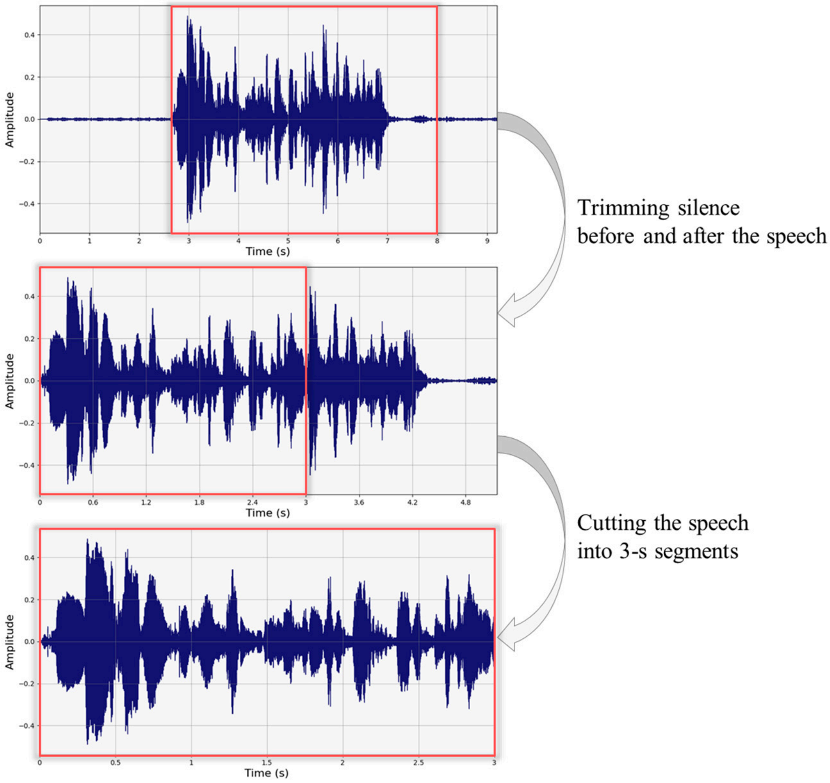


Figure 1. Speech data preprocessing procedures.

Table 1. Composition of the training, validation, and testing datasets by sex and age group.

Age Group	Training		Validation		Testing	
	Male	Female	Male	Female	Male	Female
11–14	6756	7637	764	860	799	797
15–19	9308	8535	1013	922	987	1000
20–24	8822	8833	1000	1034	982	991
25–29	8993	9004	979	980	999	999
30–34	8859	8908	1012	1041	984	997
35–39	8867	8942	1025	977	982	990
40–44	8897	8800	1015	998	990	983
45–49	8884	8370	1011	926	991	993
50–54	8907	8987	950	975	989	998
55–59	8987	9012	929	957	991	996
Total	87,280	87,028	9698	9670	9694	9744

2.2. Feature Extraction

Feature extraction is the process of converting a raw speech signal into acoustic feature vectors that capture the specific characteristics of the speaker [18]. The performance and prediction accuracy can vary, depending on the method used for extracting the features from the speech signal [19]. All audio in this study was sampled at a rate of 16 kHz and consists of 3 s segments. Features frequently used in speech and emotion recognition, such as the Mel spectrogram, log-Mel spectrogram, MFCCs, and delta MFCCs, were extracted and used for model training.

The spectrogram represents the amplitude or intensity of an audio signal at various frequencies over time. It is extracted by dividing the audio signal into short time segments and then performing a Fourier transform on each segment to detect the frequency components within that segment. In the plot, the x-axis represents time and the y-axis represents frequency, with the amplitude of each frequency shown as a heatmap. Lighter colors in the plot indicate higher energy at that frequency, whereas darker colors represent empty or dead sounds.

However, because the human auditory system can perceive only a limited range of frequencies and amplitudes, spectrograms do not capture all the information necessary for human-level sound. To mimic the response of the human ear to sound, a Mel filter bank was applied. This represents the amplitude and frequency based on the Mel scale, which corresponds to the frequencies perceived by humans. It is converted from a frequency using Equation (1).

$$Mel(f) = 2595\, log_{10}\left(1 + \frac{frequency}{700}\right) \tag{1}$$

A Mel spectrogram combines the Mel scale with a spectrogram to provide a visual representation in the frequency and time domains. In the plot, the x-axis represents time, and the y-axis represents the Mel-scaled frequency. Instead of showing the amplitude or intensity of the signal, the colors are displayed using a decibel scale.

The log-Mel spectrogram is a Mel spectrogram in which a logarithmic transformation has been applied to the Mel-scale frequency, which is the y-axis of the Mel spectrogram [20]. These spectrogram-based features have been used in fields such as speech emotion recognition [21] and environmental sound classification [22]. Figure 2 presents the spectrogram, Mel spectrogram, and log-Mel spectrogram images generated from the audio.

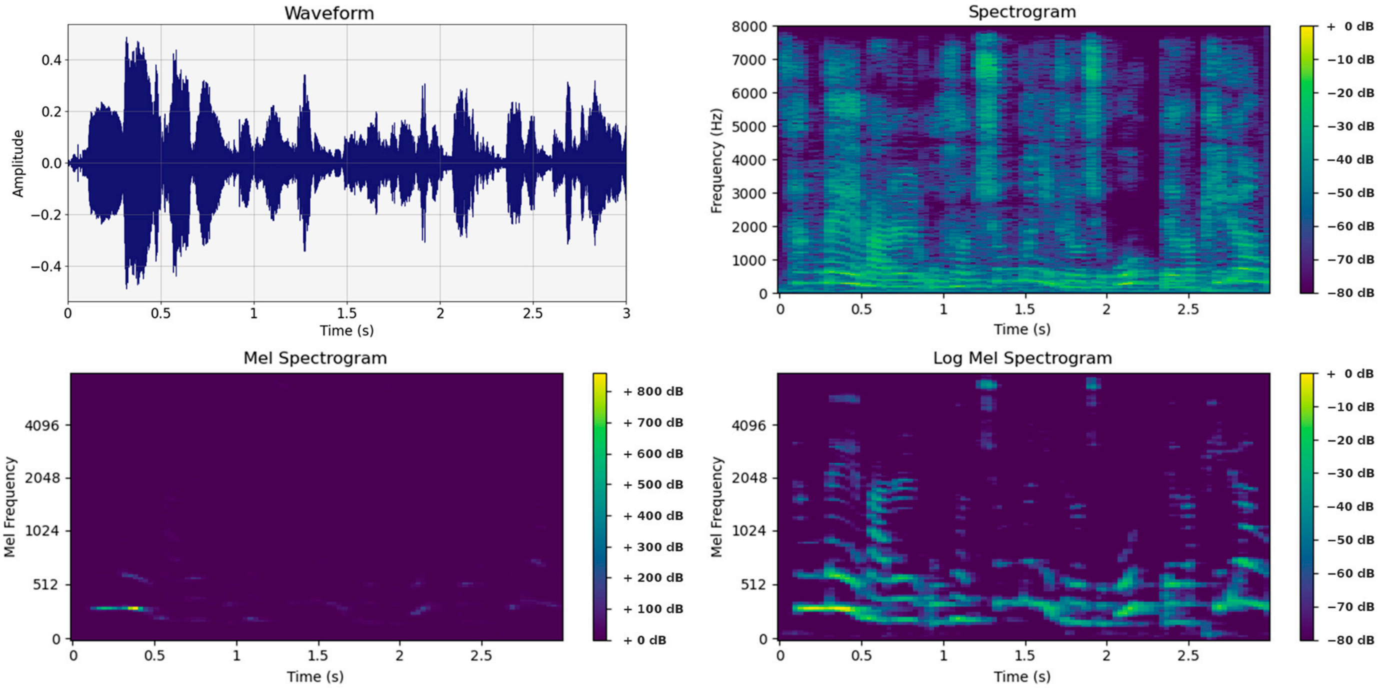


Figure 2. Visualization of a spectrogram, Mel spectrogram, and log-Mel spectrogram in audio.

MFCCs are a set of characteristic coefficients that represent audio signals based on a Mel filter bank and reflect the properties of the human auditory frequency range. The MFCCs apply a discrete cosine transform (DCT) to the log-Mel spectrum and transform it into the time domain, as shown in Equation (2). Here, X_m denotes the log energy of the mth Mel spectrogram, and c denotes the index of the cepstral coefficient. MFCCs have been applied in various fields, including speech emotion recognition [23] and music genre classification [24].

$$MFCCs = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_m(i) \cos\left(\frac{c\pi\left(m - \frac{1}{2}\right)}{M_m}\right) \quad (2)$$

Additionally, the delta coefficients of MFCCs are often used in speech recognition systems to capture the rate of change or trajectories of MFCCs [25]. These coefficients were extracted by computing the n-order difference cepstral coefficients by applying Equation (3) to the MFCCs.

$$d_i(n) = \frac{\sum_{n=1}^{N} n(c_{i+n} - c_{i-n})}{2\sum_{n=1}^{N} n^2}, \quad (3)$$

where $d_i(n)$ denotes the delta coefficient, calculated from static coefficients c_{i+n} and c_{i-n} at frame i. When n = 1, the velocity coefficient (delta) is obtained, indicating a first-order difference. When n = 2, the acceleration coefficient (delta-delta) is derived, representing the second-order difference [26]. Delta MFCCs are commonly used alongside MFCCs and have been applied to tasks such as disguised voice classification [27] and speaker recognition [28]. In this study, MFCCs + ΔMFCCs, the stacked combination of MFCCs and ΔMFCCs, were used for model training. Figure 3 provides a visualization of the MFCCs and delta MFCCs converted from the audio.

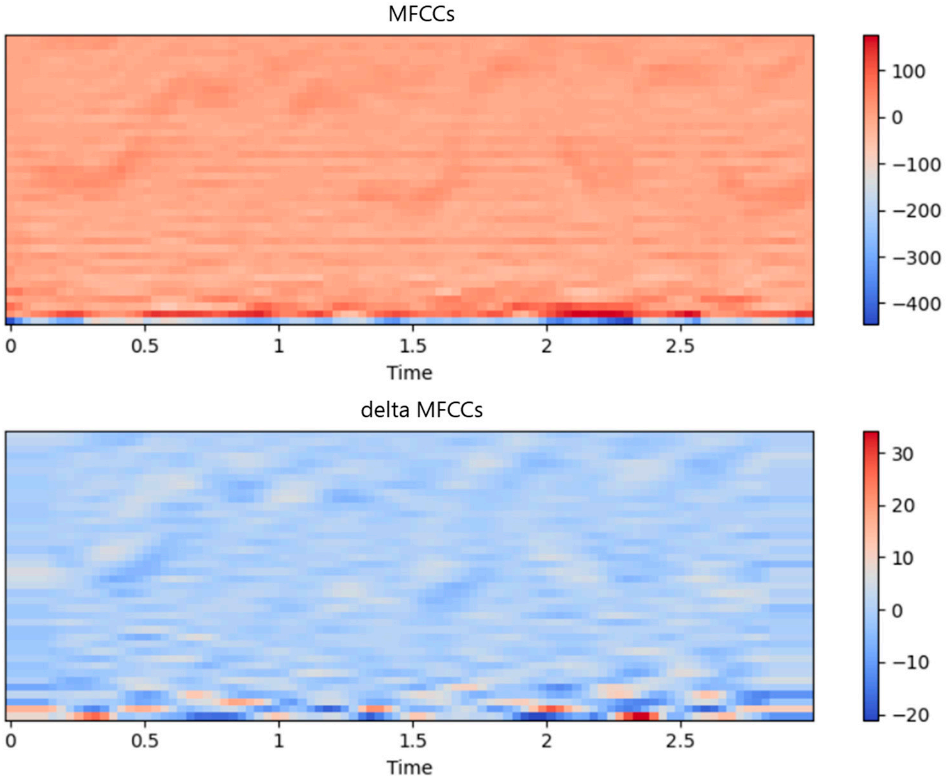


Figure 3. Representation of MFCCs and delta MFCCs in audio.

3. Neural Network Architectures

Two main frameworks are typically employed for audio data classification. The first involves extracting voice quality features and applying them to deep neural networks (DNNs). Although this approach allows for an intuitive interpretation of the results, it may not fully capture the complex characteristics inherent in speech signals. The second framework converts audio data into spectrogram-based images using feature extraction techniques and applies them directly to DNNs. This method preserves both temporal and frequency information, enabling the model to learn intricate patterns within the speech signals. However, this method requires significant computational resources and can complicate the interpretation of results.

Convolutional neural networks (CNNs) were developed primarily to solve image-related problems and have demonstrated strong performance in various speech-related tasks, such as automatic speech recognition, speech emotion recognition, and speaker recognition [29,30]. Recent developments in computer vision have led to the introduction of various techniques, among which the vision transformer (ViT) has gained attention for its exceptional image classification performance. Unlike traditional CNNs, ViTs can capture global information and context by processing the entire input holistically, which has resulted in attempts to use ViTs for classifying speech images [15,31].

The image-based approach effectively preserves the time frequency information in speech and captures subtle variations in speech patterns across age groups. In this study, three types of CNN-based models and transformer-based ViTs were trained, and their performances were compared. Detailed descriptions of the models are provided below.

3.1. ResNet-50

ResNet is an architecture composed of residual layers, with its main components being residual blocks and skip connections between layers that allow information to bypass one or more layers [32]. This architecture effectively addresses the issue of vanishing gradients as the depth of the neural network increases, unlike other models, where the performance tends to degrade with deeper networks. In addition, it reduces the computational complexity and enhances the network's training capacity. ResNet-50, which is a variant of ResNet, comprises 50 layers and more than 25.6 million parameters.

In this study, the ResNet-50 architecture was trained on a high-quality Korean-speech dataset instead of a pre-trained model, in order to build a model specialized for age classification. The structure of the model is illustrated in Figure 4. The architecture begins with a 7 × 7 convolutional layer with stride 2, followed by five sequential sets of residual blocks: three blocks with 64 filters (×3), four blocks with 128 filters (×4), six blocks with 256 filters (×6), and three blocks with 512 filters (×3). Each residual block incorporates 1 × 1 and 3 × 3 convolutions, complemented by skip connections that preserve feature information and improve gradient flow. These skip connections are particularly crucial for capturing subtle differences in age-related voice characteristics. The network concludes with an average pooling layer, followed by a fully connected layer with softmax activation, which performs the final classification into ten age groups.

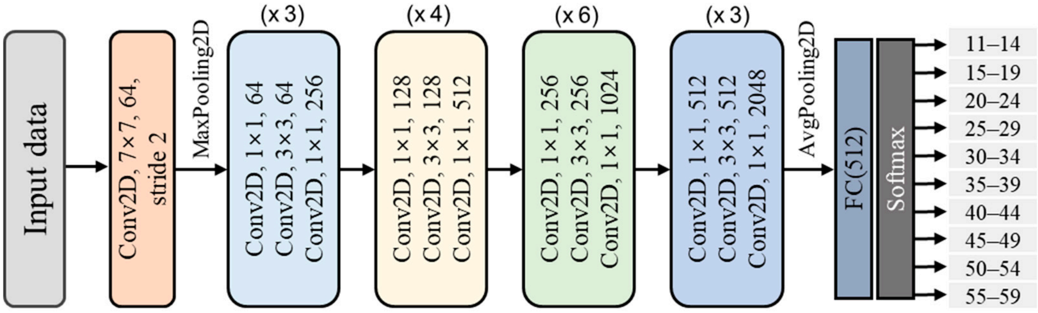


Figure 4. ResNet-50 architecture.

3.2. 1D-CNN

1D-CNNs have been effective in various fields, such as speech emotion recognition [33], music genre classification [34], and biomedical signal classification [35]. In this study, transposed speech data were used as input for the 1D-CNN model to capture the temporal patterns of frequency changes over timestamps.

The structure of the model is illustrated in Figure 5. The architecture consists of five sequential convolutional blocks, each employing a kernel size of 3, with progressively increasing filter sizes: 32, 64, 128, 256, and 512. Each block comprises a Conv1D layer followed by batch normalization and ReLU activation. To prevent overfitting while maintaining the model's ability to learn time-dependent variations effectively, dropout layers (rate = 0.2) are strategically placed after selected convolutional blocks. The network culminates in a global max pooling layer followed by fully connected layers, with a final softmax activation layer for age group classification.

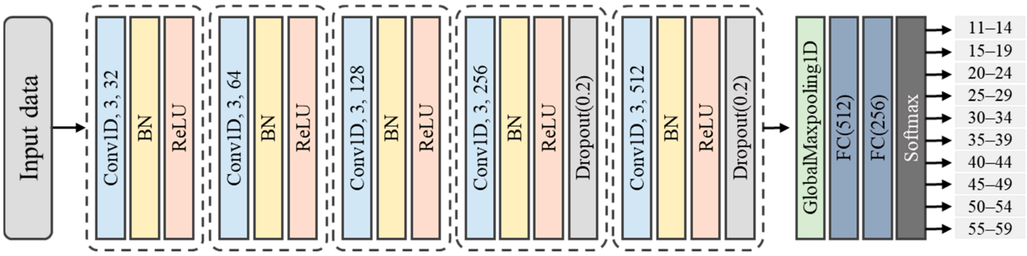


Figure 5. 1D-CNN architecture.

3.3. Multi-Kernel 2D-CNN

A 2D-CNN with three parallel convolutional layers, each employing different kernel sizes, was used to extract the time- and frequency-domain features from the audio feature map. Typically, the receptive field of a CNN expands with additional layers. However, increasing the number of layers increases the number of parameters, which can lead to overfitting. Previous studies have demonstrated that the model performance can be improved by expanding the receptive field using different kernel sizes without increasing the number of layers [36,37].

The overall architecture, shown in Figure 6, employs three parallel branches at its input stage. Each branch consists of a Conv2D layer with distinct kernel sizes (11 × 1, 1 × 9, and 3 × 3, as illustrated in Figure 7), followed by batch normalization, ReLU activation, and dropout layers. This parallel structure enables the simultaneous extraction of various temporal and frequency features at different scales. The features from these parallel branches are concatenated and processed through four additional convolutional blocks, each comprising Conv2D, batch normalization, ReLU activation, and dropout layers. This design allows the network to capture a comprehensive range of acoustic cues, from local patterns to extended temporal dependencies, which is particularly effective for distinguishing subtle age-related variations in Korean speech. The architecture concludes with a global max pooling layer and fully connected layers for final classification.

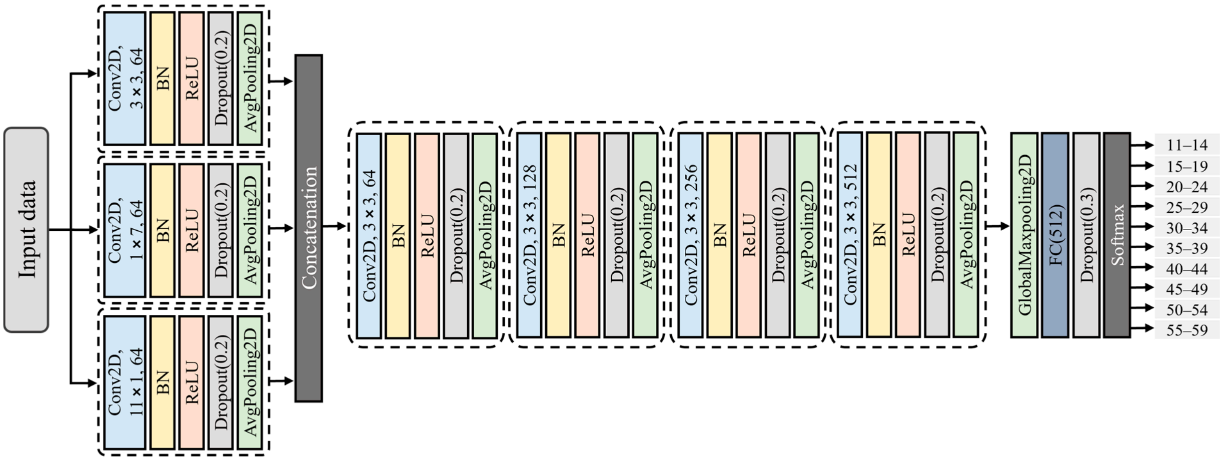


Figure 6. Multi-kernel 2D-CNN architecture.

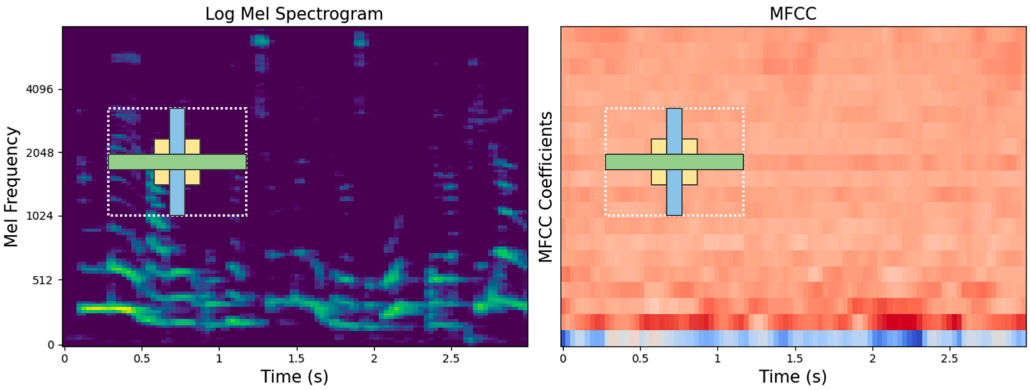


Figure 7. Receptive field (dotted rectangle) from parallel convolution layers (blue, green, and yellow rectangles).

3.4. Vision Transformer

The transformer is an end-to-end natural language processing (NLP) model proposed in 2017 that employs a self-attention mechanism to capture global dependencies and facilitate parallel processing [38]. Traditionally, CNNs have been regarded as the fundamental architecture in vision applications; however, transformers have recently emerged as a promising alternative [39]. The vision transformer (ViT) was the first transformer architecture applied to computer vision tasks and has been used for image classification, object detection, and video processing [40].

In speech research, ViT has been applied to tasks such as speech emotion recognition [41] and biomedical signal classification [42] because it can capture both local and global features from speech represented as images. The architecture of our implemented model, shown in Figure 8, begins with the division of input spectrograms into smaller patches. These patches undergo linear embedding and are augmented with learned positional encodings before being processed by the transformer encoder. The encoder consists of multiple transformer blocks, each containing layer normalization followed by multi-head self-attention mechanisms and MLP layers with skip connections. This structure enables the model to learn relationships between patches at multiple scales, capturing both fine-grained details and global patterns in the speech signal. The final classification is performed through an MLP head with softmax activation, which produces age group predictions based on the encoded features.

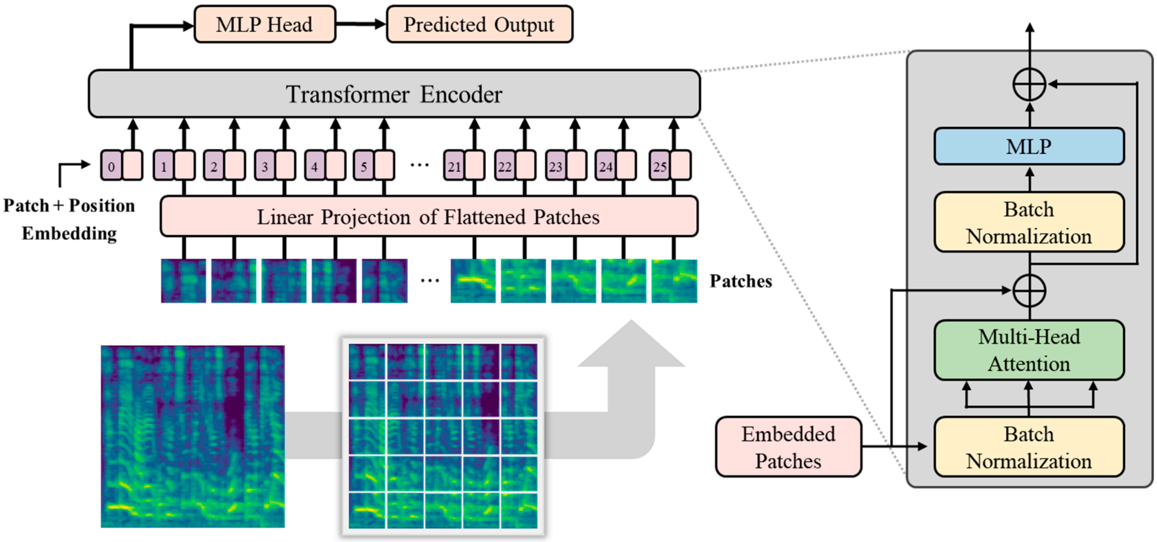


Figure 8. Vision transformer architecture.

4. Results and Analyses

Figure 9 illustrates the process overview of this study. As shown in the figure, the Korean speech dataset preprocessing included sampling and trimming. After preprocessing, speech features such as the Mel spectrogram, log-Mel spectrogram, MFCCs, and MFCCs + ΔMFCCs were extracted, and the ResNet-50, 1D-CNN, 2D-CNN, and ViT models were trained using the extracted features. Finally, the performances were compared to identify the most suitable combination of speech features and deep learning models for Korean speech. The age was predicted separately for male and female speech data, and the combination of voice features and deep learning models that achieved the highest accuracy for each sex was examined. The computational environment used in the experiment is listed in Table 2.

Table 2. Experimental environment.

Hardware		Software	
CPU	AMD $Ryzen^{TM}$ 9 7950 @ 4.5 GHz (AMD, Santa Clara, CA, USA)	OS	Linux Ubuntu 20.04
GPU	NVIDIA GeForce RTX 4090 (NVIDIA Corporation, Santa Clara, CA, USA)	Programming Language	Python 3.8.16
RAM	64 GB		

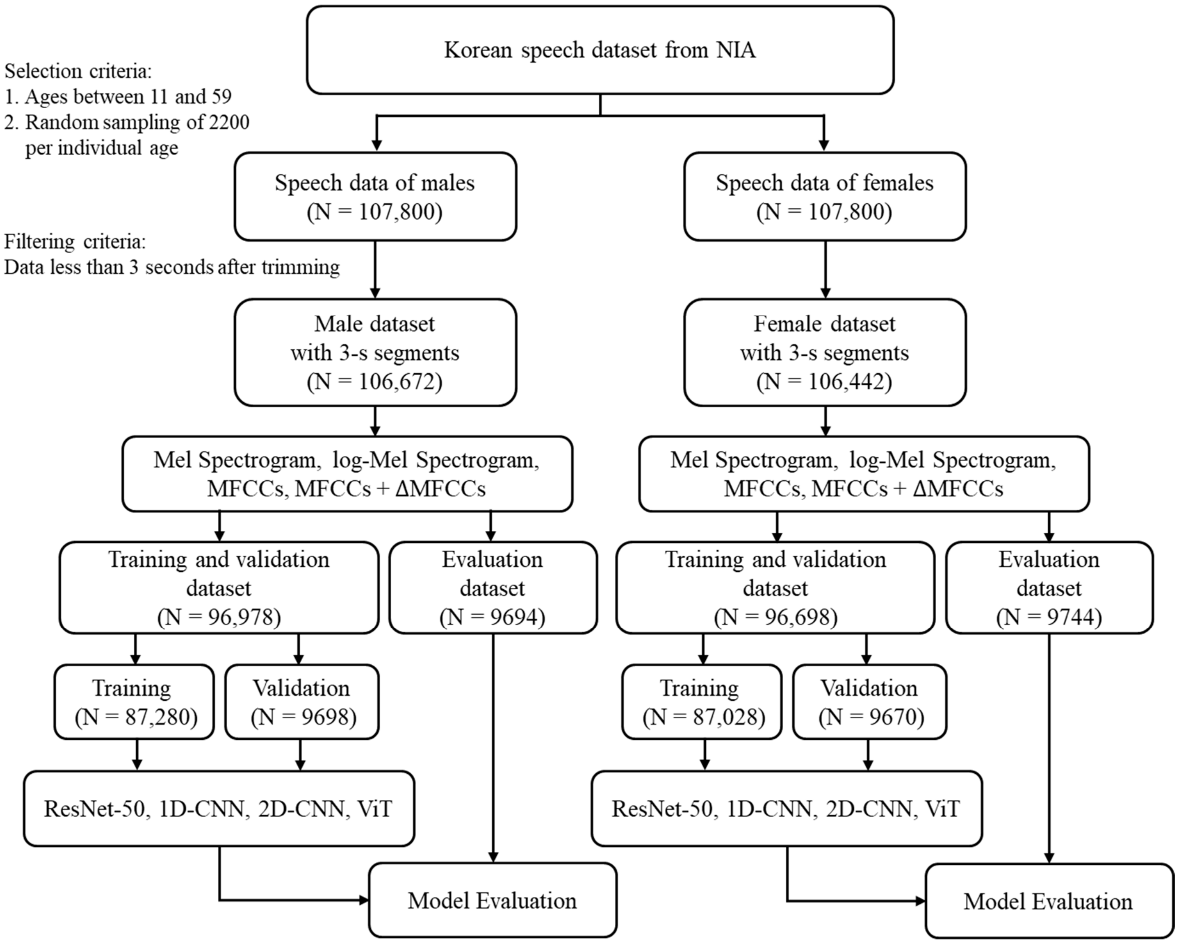

Korean speech dataset from NIA
Selection criteria:
1. Ages between 11 and 59
2. Random sampling of 2200 per individual age
Speech data of males (N = 107,800)
Speech data of females (N = 107,800)
Filtering criteria:
Data less than 3 seconds after trimming
Male dataset with 3-s segments (N = 106,672)
Female dataset with 3-s segments (N = 106,442)
Mel Spectrogram, log-Mel Spectrogram, MFCCs, MFCCs + ΔMFCCs
Mel Spectrogram, log-Mel Spectrogram, MFCCs, MFCCs + ΔMFCCs
Training and validation dataset (N = 96,978)
Evaluation dataset (N = 9694)
Training and validation dataset (N = 96,698)
Evaluation dataset (N = 9744)
Training (N = 87,280)
Validation (N = 9698)
Training (N = 87,028)
Validation (N = 9670)
ResNet-50, 1D-CNN, 2D-CNN, ViT
ResNet-50, 1D-CNN, 2D-CNN, ViT
Model Evaluation
Model Evaluation

Figure 9. Process overview.

4.1. Evaluation Metrics

Multiple evaluation metrics were used to assess the performance of the age prediction model. These metrics are described by Equations (4)–(7). When the model correctly predicts positive data, it is referred to as a true positive (TP). When the model incorrectly predicts data as positive that are actually negative, this is called a false positive (FP). A false negative (FN) occurs when the model predicts a negative for data that are actually positive, whereas a true negative (TN) occurs when the model correctly predicts data that are actually negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1-score = \frac{2\times TP}{2\times TP+2\times FN} \quad (7)$$

The accuracy, calculated using Equation (4), represents the proportion of correctly predicted samples out of the total samples. The precision, calculated using Equation (5), is the proportion of samples predicted to be positive that are actually positive. The recall, calculated using Equation (6), represents the proportion of actual positives that are correctly predicted. Finally, the F1-score, calculated using Equation (7), is the harmonic mean of the precision and recall.

4.2. Results

Tables 3 and 4 present the prediction accuracies for the 10 age groups using the male and female speech datasets, respectively. The hyperparameters for each combination are presented in Appendix A. For males, the ViT model exhibits the best average performance across all four speech features. The top performing input feature, on average, across the four models was MFCCs + ΔMFCCs. The best combination of speech features and deep-learning models for predicting male ages was MFCCs + ΔMFCCs and the 1D-CNN model, achieving an accuracy of 88.16%.

Table 3. Age prediction accuracy (%) for males.

Input Feature Model	Mel Spectrogram	log-Mel Spectrogram	MFCCs	MFCCs + ΔMFCCs	Average
ResNet-50	66.05	85.73	87.00	87.23	81.50
1D-CNN	65.83	82.02	87.05	88.16	80.77
2D-CNN	61.11	85.44	86.80	85.89	79.81
ViT	78.84	86.36	85.15	85.55	83.98
Average	67.96	84.89	86.50	86.71	81.52

For females, the ViT model also showed the best average performance across all four speech features. The top performing input feature, on average, across the four models was MFCCs. In conclusion, the combination of MFCCs + ΔMFCCs and the 1D-CNN model achieved the highest accuracy for predicting female ages, with a result of 81.95%. In summary, ViT demonstrated the best average performance for age prediction in both males and females, and the combination of MFCCs + ΔMFCCs and 1D-CNN provided the best results for both sexes.

Table 4. Age prediction accuracy (%) for females.

Input Feature Model	Mel Spectrogram	log-Mel Spectrogram	MFCCs	MFCCs + ΔMFCCs	Average
ResNet-50	55.31	78.45	79.64	78.19	72.90
1D-CNN	48.47	70.89	80.78	81.95	70.52
2D-CNN	32.15	77.91	81.07	79.33	67.62
ViT	66.20	74.74	78.43	77.75	74.28
Average	50.53	75.50	79.98	79.31	71.33

Table 5 provides detailed results for the 1D-CNN and MFCCs + ΔMFCCs combination, which achieved an accuracy of 88.16% for male age prediction. It includes specific details on the precision, recall, and F1-score for each age group. Class @11–14′ achieved 97% precision, 97% recall, and 97% F1-score. Class @55–59′ surpassed the other age groups with 96% precision, 97% recall, and 96% F1-score. Figure 10 illustrates the confusion matrix for the male age prediction results.

Table 5. Detailed performance for male age prediction using 1D-CNN with MFCCs + ΔMFCCs.

Age Group	Precision	Recall	F1-Score	Support
11–14	0.97	0.97	0.97	799
15–19	0.91	0.90	0.91	987
20–24	0.86	0.93	0.89	982
25–29	0.81	0.91	0.86	999
30–34	0.84	0.68	0.75	984
35–39	0.83	0.78	0.80	982
40–44	0.83	0.86	0.85	990
45–49	0.91	0.90	0.90	991
50–54	0.92	0.94	0.93	989
55–59	0.96	0.97	0.96	991
Weighted avg	0.88	0.88	0.88	9694
Accuracy		0.8816		

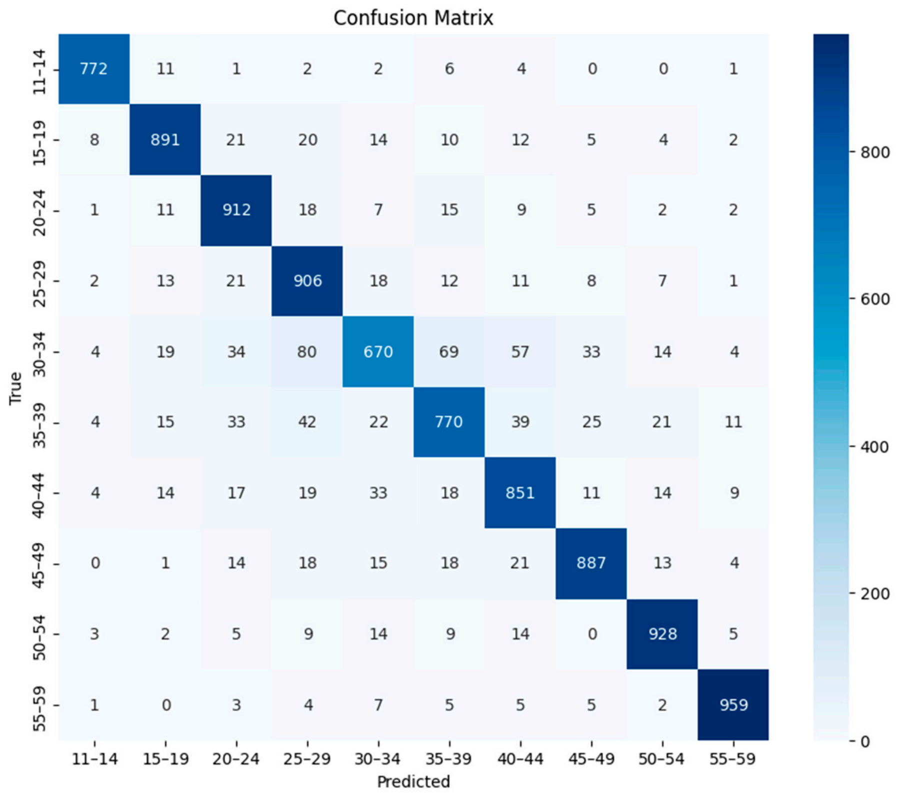
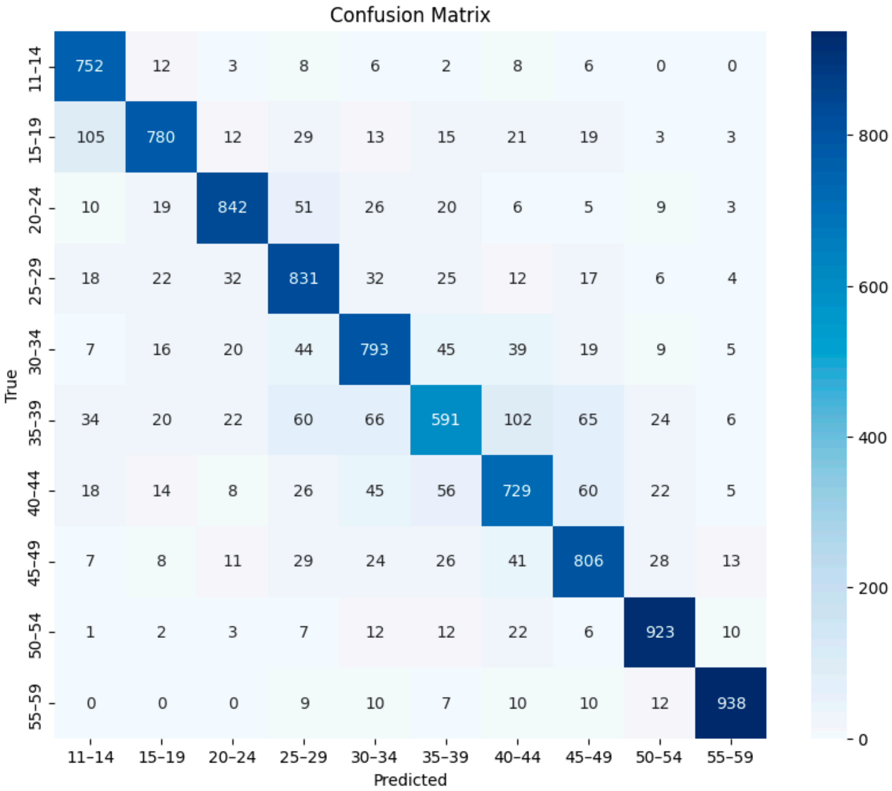


Figure 10. Confusion matrix for male age prediction using 1D-CNN with MFCCs + ΔMFCCs.

Table 6 provides detailed results for the 1D-CNN and MFCCs + ΔMFCCs combination, which achieved the highest accuracy of 81.95% in the female age prediction. Class ‘55–59’ recorded 95% precision, 94% recall, and 95% F1-score, surpassing other age groups. However, Class ‘35–39’ showed a relatively low prediction performance, with 74% precision, 60% recall, and 66% F1-score. Figure 11 illustrates the confusion matrix for the female age-prediction results, indicating that Class ‘35–39’ was often misclassified as Class ‘40–44’.

Table 6. Detailed performance for female age prediction using 1D-CNN with MFCCs + ΔMFCCs.

Age Group	Precision	Recall	F1-Score	Support
11–14	0.79	0.94	0.86	797
15–19	0.87	0.78	0.82	1000
20–24	0.88	0.85	0.87	991
25–29	0.76	0.83	0.79	999
30–34	0.77	0.80	0.78	997
35–39	0.74	0.60	0.66	990
40–44	0.74	0.74	0.74	983
45–49	0.80	0.81	0.80	993
50–54	0.89	0.92	0.91	998
55–59	0.95	0.94	0.95	996
Weighted avg	0.82	0.82	0.82	9744
Accuracy		0.8195		



	11–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59
11–14	752	12	3	8	6	2	8	6	0	0
15–19	105	780	12	29	13	15	21	19	3	3
20–24	10	19	842	51	26	20	6	5	9	3
25–29	18	22	32	831	32	25	12	17	6	4
30–34	7	16	20	44	793	45	39	19	9	5
35–39	34	20	22	60	66	591	102	65	24	6
40–44	18	14	8	26	45	56	729	60	22	5
45–49	7	8	11	29	24	26	41	806	28	13
50–54	1	2	3	7	12	12	22	6	923	10
55–59	0	0	0	9	10	7	10	10	12	938

Figure 11. Confusion matrix for female age prediction using 1D-CNN with MFCCs + ΔMFCCs.

4.3. Comparative Analyses

4.3.1. Performance Benchmarking Against Previous Studies

Recent studies have demonstrated various approaches to speech-based age prediction, with promising results. So and You [10] focused on three broad age groups, achieving accuracies of 78.6% for men and 71.9% for women using traditional speech processing techniques. Tursunov and Mustaqeem [3] explored different datasets and classification granularities, reporting 72% accuracy on the Common Voice dataset for six age groups (teens through sixties) and notably higher accuracy (96%) on a Korean speech recognition dataset when classifying just three broad categories (children, adults, elderly). Al-Maashani, Mendonça [43] achieved a remarkable performance, with 97% accuracy, by combining CNN-based Mel-Spectrogram analysis with comprehensive acoustic feature extraction (including MFCCs, spectral contrast, roll-off, and bandwidth) for six age groups. Derdour, Henni [44] demonstrated robust performance across different demographic categories, reporting accuracy exceeding 85% across various accent groups and approximately 88% for gender-based classifications.

While our achieved accuracies of 88.16% for males and 81.95% for females might appear lower than some previous studies, our research makes significant contributions by addressing more challenging and practical aspects of age prediction. Specifically, our approach successfully distinguishes between ten distinct five-year intervals, representing a considerably more complex task than the broader categorizations of three to six groups used in previous studies. This fine-grained classification capability using MFCCs + ΔMFCCs with 1D-CNN demonstrates the robustness of our approach in handling the nuanced characteristics of Korean speech, while maintaining reasonable accuracy levels. Our results establish an important foundation for applications requiring detailed age prediction, such as personalized service delivery and age-sensitive content adaptation, particularly within Korean language contexts. This work not only extends the boundaries of previous research but also addresses the growing demand for more precise age classification in real-world applications.

4.3.2. Gender-Based Performance Analysis

The results indicated a noticeable performance difference between male and female voices. This section presents a comparative analysis of the results. For the 16 combinations, the average prediction performance for males was 81.52%, whereas that for females was 71.33%. In all combinations, male performance surpassed female performance. This suggests that Korean male voices exhibit more distinct characteristics across five-year age intervals than female voices.

This performance disparity between male and female voices can be attributed to several physiological and acoustic factors [45]. Males typically exhibit more pronounced anatomical changes in their vocal apparatus with age, particularly in laryngeal structure and vocal fold characteristics. These changes result in more distinct and progressive alterations in fundamental frequency, harmonic structure, and overall voice quality across age groups [46]. Furthermore, male voices generally show more consistent patterns of age-related changes in terms of pitch lowering and resonance modifications, making it easier for the models to learn and classify age-specific characteristics. The relatively stable progression of these changes in males contributes to the higher prediction accuracy.

In contrast, female voices present more complex patterns of age-related changes, influenced by both physiological and sociocultural factors. The effects of hormonal changes throughout life stages, particularly during and after menopause, create more variable patterns in voice characteristics [47]. Additionally, female speakers often demonstrate greater variability in speaking styles, intonation patterns, and voice modulation, which can obscure age-specific features [48]. This variability is further complicated by social and professional factors, as women may consciously or unconsciously modify their speaking patterns to conform to various social contexts. These combined factors result in more overlapping voice characteristics between different age groups in females, especially in the middle-age ranges (30–39 years), making accurate age classification more challenging for the models. The lower prediction accuracy for females thus reflects the inherent complexity and greater variability in female voice aging patterns.

Analyzing performance variations across age groups revealed that, for both sexes, the accuracy tended to be higher in younger groups (e.g., 11–14 and 15–19) and older groups (e.g., 55–59), while middle-aged groups, such as 30–39, showed slightly lower accuracy. This may suggest that vocal changes in middle-aged speakers are less distinct, potentially due to overlapping pitch and tone characteristics during this period of life. Additionally, the difference in performance between male and female speakers was particularly pronounced in these middle-aged groups, possibly indicating more gradual or subtle vocal changes in females than in males as they aged.

4.3.3. Feature Extraction and Model Performance

Another observation is that, across all input features, MFCCs and MFCCs + ΔMFCCs consistently provided the highest accuracy for both sexes, particularly when used with the 1D-CNN model. This consistent result across features and sex groups emphasizes the robustness of MFCC-based features in capturing age-related information in Korean speech, regardless of sex. However, the Mel spectrograms, especially for female speakers, yielded lower accuracies. This suggests that the Mel spectrogram may lack the detail necessary for accurately capturing age-related differences in vocal characteristics, especially when compared with more nuanced features, such as MFCCs or the log-Mel spectrogram.

The findings underscore not only the sex-specific patterns in age prediction accuracy but also the essential importance of selecting feature extraction methods attuned to the linguistic and demographic nuances of the dataset. Notably, the consistent accuracy of MFCC-based features across the sexes indicates their efficacy in capturing the age-related characteristics of Korean speech. This result suggests that tailoring feature extraction approaches to a dataset's linguistic characteristics and demographic distinctions can enhance predictive performance, especially in capturing nuanced age progression within specific sex groups.

Figure 12 shows the prediction accuracy of each model based on the input features. In the male age prediction results, when comparing the performance across input features, the Mel spectrogram showed a significantly lower average performance than the other input features. The log-Mel spectrogram, which applies a logarithmic scale to the Mel spectrogram, demonstrated higher performance, suggesting that the Mel spectrogram is less effective at distinguishing the characteristics of male and female voices. The log-Mel spectrogram achieved the highest performance with ViT for male age prediction and ResNet-50 for female age prediction. The MFCCs performed best with a 1D-CNN for males and a 2D-CNN for females. For both sexes, the best performance was obtained when training a 1D-CNN with MFCCs + ΔMFCCs.

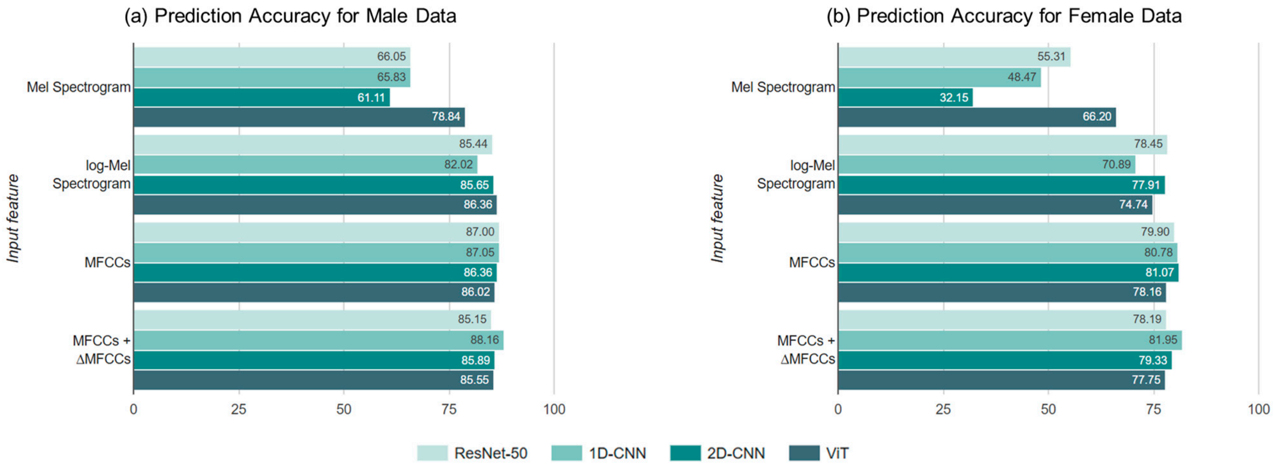


Figure 12. Comparison of age prediction accuracy across different models by input feature for (**a**) male and (**b**) female data.

Figure 13 shows the prediction accuracy of each input feature by model. ResNet-50 and 2D-CNN achieved the best performance for both males and females when trained with MFCCs, while 1D-CNN performed best when using MFCCs + ΔMFCCs. For ViT, the best performance was achieved using the log-Mel spectrogram for males and MFCCs for females. The models trained with Mel spectrograms—ResNet-50, 1D-CNN, and 2D-CNN—achieved an accuracy of up to 66.05% for male age prediction and up to 55.31% for female age prediction. In contrast, ViT achieved higher accuracies: 78.84% for males and 66.20% for females. This demonstrates that ViT is a robust model that is capable of capturing key features, even in data where feature identification is challenging.

Although transformers, including ViTs, have demonstrated remarkable performance across various computer vision tasks, they inherently lack certain inductive biases found in CNNs. CNNs are designed with the assumption that images or speech signals contain rich and localized information, which allows them to effectively capture regional details through spatial hierarchies. This attribute is advantageous for tasks such as age prediction, in which local patterns in voice data can reveal subtle age-related features. In contrast, ViT divides images into small patches and uses self-attention mechanisms to capture the context among patches, thereby facilitating comprehensive feature interaction across the input. However, this approach may not emphasize local structures as effectively as a CNN.

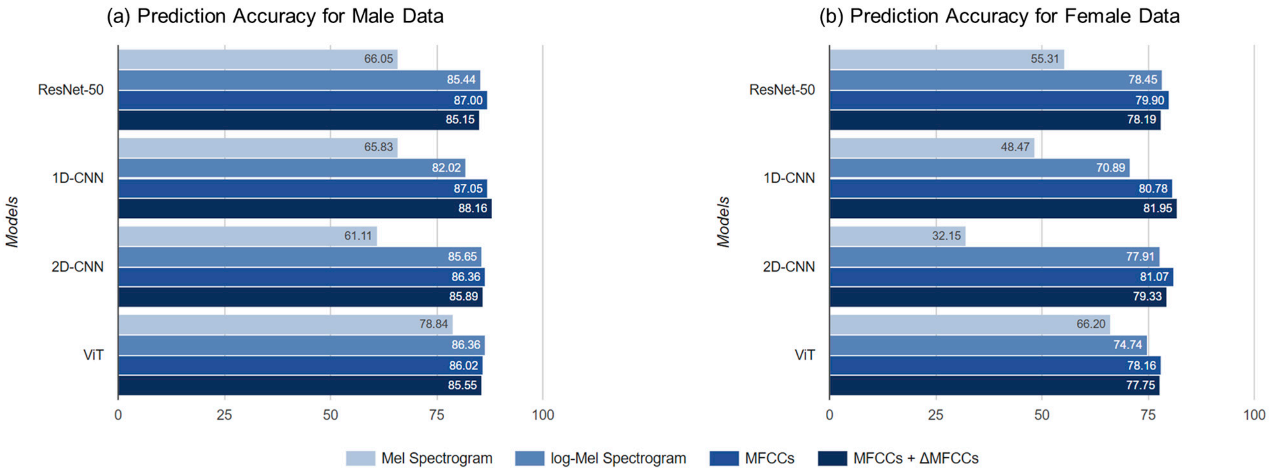


Figure 13. Comparison of age-prediction accuracy across different input features by model for (**a**) male and (**b**) female data.

Although ViT incorporates positional embeddings and self-attention to learn the global context, it may overlook the fine-grained details that are essential for tasks that rely on localized patterns. Consequently, while ViT exhibited robust overall accuracy, the highest accuracy was achieved by the 1D-CNN model with MFCCs + ΔMFCCs, which excelled in capturing intricate age-related features specific to Korean speech data. This finding suggests that, while transformers are beneficial for tasks requiring broad contextual analysis, CNNs are particularly effective when nuanced, localized feature extraction is crucial for accuracy.

In summary, the MFCCs + ΔMFCCs and 1D-CNN combination proved to be the most effective for predicting the age of Korean speakers. In addition, the ViT model demonstrated high performance, even in cases where feature identification was more difficult, demonstrating its potential for application in speech recognition tasks.

5. Conclusions

In this study, four different speech feature extraction methods and four neural network architectures were employed to predict the ages of Korean male and female speakers aged between their teens and 50s, with the goal of identifying the optimal age-prediction combination specific to Korean speech. While many studies have explored predicting age from speech, few have focused on Korean speech, and most either lack specific age ranges or do not segment the data into distinct age groups. To address these limitations, this study used a large Korean dataset to predict ten segmented age groups.

The speech features used were the Mel spectrogram, log-Mel spectrogram, MFCCs, and MFCCs + ΔMFCCs, while the deep learning models comprised ResNet-50, 1D-CNN, 2D-CNN, and ViT. Among the four speech features, the model that achieved the highest average accuracy was ViT. For all combinations, the age prediction performance for males surpassed that for females. For both males and females, the highest accuracy was obtained when MFCCs + ΔMFCCs was trained on 1D-CNN, with accuracies of 88.16% and 81.95%, respectively. This combination represented the optimal pairing of speech features and models specialized for the Korean speech dataset presented in this study.

While this study employed established deep learning architectures, there remains significant potential for developing novel architectures that are specifically optimized for speech age prediction. Future research could focus on specialized attention mechanisms that more effectively capture age-specific vocal characteristics. Furthermore, exploring hybrid architectures that integrate the strengths of CNNs and transformers may enhance model performance, whereas investigating self-supervised learning approaches could leverage unlabeled speech data more efficiently. For female speakers, further analysis is recommended to identify the underlying factors that make distinguishing voices of those in their late 30s from those in their early 40s more challenging. The model could also be extended to perform multitask learning, allowing simultaneous recognition of age, emotions, place of origin, and other characteristics.

Nevertheless, our current framework achieves state-of-the-art performance for fine-grained Korean speech age prediction while maintaining practical applicability, demonstrating significant improvements over existing approaches. The achievement of 88.16% accuracy across ten distinct age groups represents a substantial advance in the field, particularly given the challenging nature of distinguishing between narrowly separated age ranges. The results of this study can be applied to personalized services designed for Korean speech or services sensitive to narrow age ranges. The model trained on the optimal combination demonstrated strong performance in predicting 10 segmented age groups by using short speech samples. This suggests the feasibility of automated age prediction for large, unspecified populations with potential applications in various fields. Although applying speech data from different languages or cultural backgrounds may not guarantee the same level of performance, the preprocessing techniques and framework presented in this study offer a valuable reference for the development of more advanced models and applications in speech recognition and age prediction. Looking ahead, this approach presents promising potential to further improve automated speech systems across diverse linguistic and cultural contexts. Nonetheless, several open issues remain, such as further investigating age-related changes beyond the 50s, addressing the scarcity of high-quality labeled data in other dialects and languages, and exploring advanced techniques (e.g., self-supervised or domain-adaptive methods) to enhance model robustness. These directions could pave the way for more inclusive and accurate age prediction systems, extending the applicability of this research to a broader demographic scope and a wider range of linguistic and cultural environments.

Author Contributions: Conceptualization, H.K., J.L., M.L., S.K. and J.Y.; Methodology, H.K. and J.Y.; Software, H.K., M.L. and S.K.; Validation, J.Y.; Formal analysis, J.L.; Investigation, H.K. and J.Y.; Resources, J.Y.; Data curation, H.K., M.L. and S.K.; Writing—original draft, H.K., J.L. and J.Y.; Writing—review and editing, J.L. and J.Y.; Visualization, H.K. and J.L.; Supervision, J.L. and J.Y.; Project administration, H.K., J.L. and J.Y.; Funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HK23C0030). This research was also supported by a grant from the Korea Dementia Research Project through the Korea Dementia Research Center (KDRC), funded by the Ministry of Health and Welfare and Ministry of Science and ICT, Republic of Korea (grant number: RS-2024-00342071).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed in this study are available in the AI-Hub (https://www.aihub.or.kr, accessed on 1 December 2024) of the National Information Society Agency.

Conflicts of Interest: Author Jiho Lee was employed by the company Neopons Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Table A1. Hyperparameters of experiments using the male dataset.

Male				
Features **Parameter**	**Mel Spectrogram**	**log-Mel Spectrogram**	**MFCCs**	**MFCCs + ΔMFCCs**
Model 1: ResNet-50				
Learning rate	0.0001	0.0001	0.0007	0.0005
Batch size	128	64	128	128
Epochs	100	100	70	100
Optimizer	Adam	Adam	Adam	Adam
Model 2: 1D-CNN				
Learning rate	0.001	0.0001	0.0005	0.0005
Batch size	128	128	128	64
Epochs	100	70	50	50
Optimizer	Adam	Adam	Adam	Adam
Model 3: 2D-CNN				
Learning rate	0.0001	0.0001	0.0005	0.001
Batch size	128	64	64	128
Epochs	100	100	70	100
Optimizer	Adam	Adam	Adam	Adam
Model 4: ViT				
Learning rate	0.001	0.0007	0.0005	0.0005
Batch size	128	128	128	128
Epochs	100	100	150	150
Optimizer	Adam	Adam	Adam	Adam

Table A2. Hyperparameters of experiments using the female dataset.

Female				
Features / **Parameter**	**Mel spectrogram**	**log-Mel Spectrogram**	**MFCCs**	**MFCCs + ΔMFCCs**
Model 1: ResNet-50				
Learning rate	0.0001	0.0005	0.0007	0.002
Batch size	64	128	64	128
Epochs	100	70	100	50
Optimizer	Adam	Adam	Adam	Adam
Model 2: 1D-CNN				
Learning rate	0.001	0.0001	0.0007	0.0005
Batch size	64	128	64	128
Epochs	100	100	50	50
Optimizer	Adam	Adam	Adam	Adam
Model 3: 2D-CNN				
Learning rate	0.001	0.0001	0.0005	0.001
Batch size	128	128	64	128
Epochs	100	100	70	70
Optimizer	Adam	Adam	Adam	Adam
Model 4: ViT				
Learning rate	0.0007	0.001	0.001	0.0007
Batch size	128	128	128	128
Epochs	100	100	100	70
Optimizer	Adam	Adam	Adam	Adam

Appendix B

Table A3. Age prediction performance of different models by input feature in the male dataset.

Male					
Input Feature	**Deep Learning Model**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Mel spectrogram	ResNet-50	69.79	66.51	66.67	66.05
	1D-CNN	66.03	66.37	65.94	65.83
	2D-CNN	66.64	61.77	61.71	61.11
	ViT	**79.12**	**79.15**	**78.87**	**78.84**
log-Mel spectrogram	ResNet-50	85.98	85.98	85.78	85.73
	1D-CNN	82.23	82.32	81.92	82.02
	2D-CNN	85.90	85.65	85.46	85.44
	ViT	**86.50**	**86.56**	**86.32**	**86.36**
MFCCs	ResNet-50	87.09	87.19	87.11	87.00
	1D-CNN	**87.18**	**87.25**	**87.12**	**87.05**
	2D-CNN	87.01	87.00	86.76	86.80
	ViT	85.40	85.36	85.15	85.15
MFCCs + ΔMFCCs	ResNet-50	87.31	87.43	87.19	87.23
	1D-CNN	**88.32**	**88.32**	**88.30**	**88.16**
	2D-CNN	86.26	86.10	85.87	85.89
	ViT	85.78	85.77	85.51	85.55

Table A4. Age prediction performance of different models by input feature in the female dataset.

Female					
Input Feature	**Deep Learning Model**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Mel spectrogram	ResNet-50	56.28	55.57	55.24	55.31
	1D-CNN	48.18	49.19	47.66	48.47
	2D-CNN	39.51	32.68	30.18	32.15
	ViT	**66.22**	**66.56**	**66.09**	**66.20**
log-Mel spectrogram	**ResNet-50**	**79.64**	**78.63**	**78.68**	**78.45**
	1D-CNN	71.30	71.20	70.86	70.89
	2D-CNN	78.50	78.11	78.08	77.91
	ViT	74.97	75.16	74.74	74.81
MFCCs	ResNet-50	80.34	79.90	79.87	79.64
	1D-CNN	80.84	81.02	80.84	80.78
	2D-CNN	**81.35**	**81.26**	**81.18**	**81.07**
	ViT	78.53	78.67	78.41	78.43
MFCCs + ΔMFCCs	ResNet-50	78.70	78.42	78.39	78.19
	1D-CNN	**81.92**	**82.17**	**81.85**	**81.95**
	2D-CNN	79.92	79.60	79.35	79.33
	ViT	77.74	78.05	77.74	77.75

Table A5. Age prediction performance of different input features by model in the male dataset.

Male					
Deep Learning Model	**Input Feature**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
ResNet-50	Mel spectrogram	69.79	66.51	66.67	66.05
	log-Mel spectrogram	85.98	85.98	85.78	85.73
	MFCCs	87.09	87.19	87.11	87.00
	MFCCs + ΔMFCCs	**87.31**	**87.43**	**87.19**	**87.23**
1D-CNN	Mel spectrogram	66.03	66.37	65.94	65.83
	log-Mel spectrogram	82.23	82.32	81.92	82.02
	MFCCs	87.18	87.25	87.12	87.05
	MFCCs + ΔMFCCs	**88.32**	**88.30**	**88.18**	**88.16**
2D-CNN	Mel spectrogram	66.64	61.77	61.71	61.11
	log-Mel spectrogram	85.90	85.65	85.46	85.44
	MFCCs	**87.01**	**87.00**	**86.76**	**86.80**
	MFCCs + ΔMFCCs	86.26	86.10	85.87	85.89
ViT	Mel spectrogram	79.12	79.15	78.87	78.84
	log-Mel spectrogram	**86.50**	**86.56**	**86.32**	**86.36**
	MFCCs	85.40	85.36	85.15	85.15
	MFCCs + ΔMFCCs	85.78	85.77	85.51	85.55

Table A6. Age prediction performance of different input features by model in the female dataset.

Female					
Deep Learning Model	**Input Feature**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
ResNet-50	Mel spectrogram	56.28	55.57	55.24	55.31
	log-Mel spectrogram	79.64	78.63	78.68	78.45
	MFCCs	**80.34**	**79.90**	**79.87**	**79.64**
	MFCCs + ΔMFCCs	78.70	78.42	78.39	78.19
1D-CNN	Mel spectrogram	48.18	49.19	47.66	48.47
	log-Mel spectrogram	71.30	71.20	70.86	70.89
	MFCCs	80.84	81.02	80.84	80.78

	MFCCs + ΔMFCCs	**81.92**	**82.17**	**81.85**	**81.95**
2D-CNN	Mel spectrogram	39.51	32.68	30.18	32.15
	log-Mel spectrogram	78.50	78.11	78.08	77.91
	MFCCs	**81.35**	**81.26**	**81.18**	**81.07**
	MFCCs + ΔMFCCs	79.92	79.60	79.35	79.33
ViT	Mel spectrogram	66.22	66.56	66.09	66.20
	log-Mel spectrogram	74.97	75.16	74.74	74.81
	MFCCs	**78.53**	**78.67**	**78.41**	**78.43**
	MFCCs + ΔMFCCs	77.74	78.05	77.74	77.75

References

1. Grzybowska, J.; Kacprzak, S. Speaker Age Classification and Regression Using i-Vectors. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016.
2. Ilyas, M.; Othmani, A.; Nait-Ali, A. Auditory perception based system for age classification and estimation using dynamic frequency sound. *Multimed. Tools Appl.* **2020**, *79*, 21603–21626.
3. Tursunov, A.; Mustaqeem; Choeh, J.Y.; Kwon, S. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors* **2021**, *21*, 5892.
4. Yücesoy, E. Two-level classification in determining the age and gender group of a speaker. *Int. Arab J. Inf. Technol.* **2021**, *18*, 663–670.
5. Bahari, M.H.; McLaren, M.; van Leeuwen, D.A. Speaker age estimation using i-vectors. *Eng. Appl. Artif. Intell.* **2014**, *34*, 99–108.
6. Osman, M.M.; Büyük, O.; Tangel, A. Effect of number and position of frames in speaker age estimation. *Sigma J. Eng. Nat. Sci.* **2023**, *41*, 243–255.
7. Badr, A.A.; Abdul-Hassan, A.K. Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks. *Eng. Technol. J.* **2021**, *39*, 129–140.
8. Han, K.J.; Narayanan, S. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Comput. Speech Lang.* **2013**, *27*, 151–167.
9. Avikal, S.; Sharma, K.; Barthwal, A.; Kumar, K.N.; Badhotiya, G.K. Estimation of age from speech using excitation source features. *Mater. Today Proc.* **2021**, *46*, 11046–11049.
10. So, S.; You, S.M.; Kim, J.Y.; An, H.J.; Cho, B.H.; Yook, S.; Kim, I.Y. Development of Age Classification Deep Learning Algorithm Using Korean Speech. *J. Biomed. Eng. Res.* **2018**, *39*, 63–68.
11. Gupta, T.; Truong, T.D.; Anh, T.T.; Chng, E.S. Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model. *arXiv* **2022**, arXiv:2203.11774.
12. Burkhardt, F.; Wagner, J.; Wierstorf, H.; Eyben, F.; Schuller, B. Speech-based Age and Gender Prediction with Transformers. In Proceedings of the 15th ITG Conference, Aachen, Germany, 20–22 September 2023; VDE: Frankfurt am Main, Germany, 2023.
13. Truong, D.-T.; Anh, T.T.; Siong, C.E. Exploring Speaker Age Estimation on Different Self-Supervised Learning Models. In Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 7–10 November 2022; IEEE: Piscataway, NJ, USA, 2022
14. Kalluri, S.B.; Vijayasenan, D.; Ganapathy, S. Automatic speaker profiling from short duration speech data. *Speech Commun.* **2020**, *121*, 16–28.
15. Akinpelu, S.; Viriri, S.; Adegun, A. An enhanced speech emotion recognition using vision transformer. *Sci. Rep.* **2024**, *14*, 13126.
16. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end audiovisual speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018.
17. Daniel, E. Noise and hearing loss: A review. *J. Sch. Health* **2007**, *77*, 225–231.
18. Alemu, A.A.; Melese, M.D.; Salau, A.O. Towards audio-based identification of Ethio-Semitic languages using recurrent neural network. *Sci. Rep.* **2023**, *13*, 19346.
19. Koduru, A.; Valiveti, H.B.; Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. *Int. J. Speech Technol.* **2020**, *23*, 45–55.
20. Pyrovolakis, K.; Tzouveli, P.; Stamou, G. Multi-modal song mood detection with deep learning. *Sensors* **2022**, *22*, 1065.
21. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881.

22. Mu, W.; Yin, B.; Huang, X.; Xu, J.; Du, Z. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Sci. Rep.* **2021**, *11*, 21552.
23. Lalitha, S.; Geyasruti, D.; Narayanan, R. Emotion detection using MFCC and cepstrum features. *Procedia Comput. Sci.* **2015**, *70*, 29–35.
24. da Silva, A.C.M.; MCoelho, A.N.; Neto, R.F. A Music Classification model based on metric learning applied to MP3 audio files. *Expert Syst. Appl.* **2020**, *144*, 113071.
25. Venkataramanan, K.; Rajamohan, H.R. Emotion recognition from speech. *arXiv* **2019**, arXiv:1912.10458.
26. Akpudo, U.E.; Hur, J.-W. A cost-efficient MFCC-based fault detection and isolation technology for electromagnetic pumps. *Electronics* **2021**, *10*, 439.
27. Singh, M.K. Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and classification techniques. *Multimed. Tools Appl.* **2024**, *83*, 77327–77345.
28. Chelali, F.Z.; Djeradi, A. Text dependant speaker recognition using MFCC, LPC and DWT. *Int. J. Speech Technol.* **2017**, *20*, 725–740.
29. Anvarjon, T.; Mustaqeem; Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors* **2020**, *20*, 5212.
30. Bai, Z.; Zhang, X.-L. Speaker recognition based on deep learning: An overview. *Neural Netw.* **2021**, *140*, 65–99.
31. Ghosh, S.; Sarkar, S.; Ghosh, S.; Zalkow, F.; Jana, N.D. Audio-visual speech synthesis using vision transformer–enhanced autoencoders with ensemble of loss functions. *Appl. Intell.* **2024**, *54*, 4507–4524.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
33. Flower, T.M.L.; Jaya, T. A novel concatenated 1D-CNN model for speech emotion recognition. *Biomed. Signal Process. Control* **2024**, *93*, 106201.
34. Allamy, S.; Koerich, A.L. 1D CNN architectures for music genre classification. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021; IEEE: Piscataway, NJ, USA, 2021.
35. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398.
36. Aftab, A.; Morsali, A.; Ghaemmaghami, S.; Champagne, B. LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022.
37. Tellai, M.; Gao, L.; Mao, Q. An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network. *Int. J. Speech Technol.* **2023**, *26*, 541–557.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
39. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110.
40. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Kumar, C.A.; Maharana, A.D.; Krishnan, S.M.; Hanuma, S.S.S.; Lal, G.J.; Ravi, V. Speech emotion recognition using CNN-LSTM and vision transformer. In *International Conference on Innovations in Bio-Inspired Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2022.
42. Liu, Z.; Jiang, H.; Zhang, F.; Ouyang, W.; Li, X.; Pan, X. Heart sound classification based on bispectrum features and Vision Transformer mode. *Alex. Eng. J.* **2023**, *85*, 49–59.
43. Al-Maashani, T.; Mendonça, I.; Aritsugi, M. Age Classification Based on Voice Using Mel-Spectrogram and MFCC. In Proceedings of the 2023 24th International Conference on Digital Signal Processing (DSP), Rodos, Greece, 11–13 June 2023; IEEE: Piscataway, NJ, USA, 2023.
44. Derdour, A.; Henni, F.; Boubchir, L. On the Use of Recurrent Neural Network based on LSTM Model for Voice-based Age Estimation. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; IEEE: Piscataway, NJ, USA, 2024; IEEE: Piscataway, NJ, USA, 2024.
45. Hunter, E.J.; Tanner, K.; Smith, M.E. Gender differences affecting vocal health of women in vocally demanding careers. *Logop. Phoniatr. Vocology* **2011**, *36*, 128–136.

46. Markova, D.; Richer, L.; Pangelinan, M.; Schwartz, D.H.; Leonard, G.; Perron, M.; Pike, G.; Veillette, S.; Chakravarty, M.M.; Pausova, Z.; et al. Age-and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. *Horm. Behav.* **2016**, *81*, 84–96.
47. Lã, F.M.; Ardura, D. What voice-related metrics change with menopause? A systematic review and meta-analysis study. *J. Voice* **2022**, *36*, 438.e1–438.e17.
48. Meurer, E.M.; Wender, M.C.O.; Corleta, H.v.E.; Capp, E. Phono-articulatory variations of women in reproductive age and postmenopausal. *J. Voice* **2004**, *18*, 369–374.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.