Article

Classification between Elderly Voices and Young Voices Using an Efficient Combination of Deep Learning Classifiers and Various Parameters

Ji-Yeoun Lee

Department of Biomedical Engineering, Jungwon University, 85 Munmu-ro, Goesan-eup, Goesan-gun 28024, Korea; jylee@jwu.ac.kr; Tel.: +82-43-830-8607

Abstract: The objective of this research was to develop deep learning classifiers and various parameters that provide an accurate and objective system for classifying elderly and young voice signals. This work focused on deep learning methods, such as feedforward neural network (FNN) and convolutional neural network (CNN), for the detection of elderly voice signals using mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstrum coefficients (LPCCs), skewness, as well as kurtosis parameters. In total, 126 subjects (63 elderly and 63 young) were obtained from the Saarbruecken voice database. The highest performance of 93.75% appeared when the skewness was added to the MFCC and MFCC delta parameters, although the fusion of the skewness and kurtosis parameters had a positive effect on the overall accuracy of the classification. The results of this study also revealed that the performance of FNN was higher than that of CNN. Most parameters estimated from male data samples demonstrated good performance in terms of gender. Rather than using mixed female and male data, this work recommends the development of separate systems that represent the best performance through each optimized parameter using data from independent male and female samples.

Keywords: elderly voice analysis; deep learning classifier; skewness; kurtosis; medical IT

Citation: Lee, J.-Y. Classification between Elderly Voices and Young Voices Using an Efficient Combination of Deep Learning Classifiers and Various Parameters. *Appl. Sci.* **2021**, *11*, 9836. <https://doi.org/10.3390/app11219836>

Academic Editor: Francesco Bianconi

Received: 4 September 2021

Accepted: 19 October 2021

Published: 21 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the author. 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

The human voice represents a complex biological signal resulting from the dynamic interaction between adduction/vibration of the vocal folds and pulmonary air emission and flow through the resonant structures [1]. Physiologic aging leads to specific changes in the anatomy and physiology of all structures involved in the production and modulation of the human voice [2–4]. The aging of laryngeal tissue changes the movement of the vocal cords, their vibration, and their opening and closing processes [5]. Hence, a possible approach to evaluate the effect of physiological aging in humans could include the analysis of voice [1]. Voice characteristics are measured by the frequency of vocal cord oscillations per second, that is, the fundamental frequency (F0), jitter, shimmer, excitation source component, etc. [6–9]. The elderly voice can also be distinguished from the voices of children, adolescents, middle-aged people, and seniors that have distinct acoustic features measured as different voice characteristics [5]. Therefore, in order to create a system for recognizing the voice of the elderly, it is necessary to understand the characteristics of changes in vocal cord tissue due to anatomical or physiological aging [10], and various welfare systems using only the voice database of the elderly should be implemented.

Previous studies that analyzed the characteristics of elderly voices can be divided into studies that recognized elderly voices through machine learning methods [11–18], and acoustic phonetic studies that recorded and analyzed the voices of subjects by gender and age [13,19–25]. In particular, deep learning studies on elderly voices have focused on the extraction of voice features calculated using computational functionals (e.g., mean, quartiles, percentiles, position of maximum and minimum, and linear regression) over

acoustic low-level descriptors (LLDs), including those related to the energy, spectrum, and cepstrum of the signal [11,12,16,18], and including the mel-frequency cepstral coefficients (MFCCs) [14,15], jitter, shimmer, sound quality descriptors, and prosodic features using acoustic analysis in sustained vowel phonation tasks [10,13,17,19].

More recent studies have widely used ensemble-based machine learning classifiers, such as classification and regression tree (CART), random forest, XGBoost, and the gradient boosting algorithm (GBM) to objectively distinguish presbyphonia from neurological voice disorders [10]. The key finding of another study was that cepstral peak prominence (CPP), one of the cepstral indices, was the most important predictor for identifying hypokinetic dysarthria in the developed prediction model, which included sociodemographic factors, health behavior, and acoustic–phonetic indices [26]. In other words, it had the advantage of analyzing elderly voices, reflecting the speech characteristics of actual subjects acoustically and sustained vowel phonation tasks for analyzing voice characteristics [27]. Lee et al. proposed higher-order statistical (HOS) parameters for the classification of young and elderly voice signals and identifying gender- and age-related differences through HOS analysis [5]. The acoustic parameters were extracted from young and elderly voice signals using Praat and a time–frequency analysis program (TF32). Additionally, their study investigated the gender- and age-related differences in HOS parameters. Young and elderly voice signals differed significantly in normalized skewness ($p = 0.005$) in females and normalized kurtosis ($p = 0.011$) in males. Therefore, normalized skewness is a useful parameter for distinguishing between young and elderly female voices, and normalized kurtosis is essential for distinguishing between young and elderly male voices.

Until now, most acoustic studies that analyzed elderly voices using deep learning methods and acoustic parameters have evaluated the sustained phonation of single vowels [13,19–25], and only a few studies have examined the recognition of elderly voices [11–18]. Furthermore, in terms of methodology, even the previous studies [10] that analyzed elderly voices using deep learning approaches merely identified the differences between elderly voices and hypokinetic dysarthria based on acoustic–phonetic analysis using both the cepstrum and spectrum by collecting the voice data of subjects.

On the other hand, deep learning methods, such as the feedforward neural network (FNN) and convolutional neural network (CNN), have been widely used as data mining algorithms for developing disease prediction models in recent years [1,28]. No study has explored the classification of elderly versus young voices using deep learning algorithms applied to the Saarbruecken voice database (SVD). Therefore, the ultimate goal of this research was to develop deep learning classifiers and various parameters that provide an accurate and objective system for classifying elderly and young voices. This work focused on deep learning methods, such as FNN and CNN, for the detection of elderly voice signals using mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstrum coefficients (LPCCs), skewness, and kurtosis parameters. Voice data were obtained from the publicly available SVD. Finally, this work investigated voice performance using deep learning methods with various combinations of parameters. Therefore, the originality of this work can be found in its proposal of a new parameter and a novel deep learning method that combines HOSs, MFCCs, and LPCCs in the /a/ vowel voice signals of elderly and young individuals. The contribution of this paper can be summarized with the following points:

- This paper introduces an intelligent elderly voice classification system based on deep learning classifiers and various parameters.
- The suggested combinations of various parameters and deep learning methods can effectively distinguish young from elderly voices.
- A lot of experimental tests are performed to confirm the effectiveness of the elderly voice classification system using the Saarbruecken voice database.
- The experimental results emphasize the superiority of the proposed elderly voice classification system integrating deep learning methods and various parameters to monitor an elderly voice for an effective and reliable system.

2. Materials and Methods

2.1. Database

This work used the Saarbruecken voice database (SVD) recorded by the Phonetics Research Institute at Saarland University, Germany [29]. This database is a collection of voice recordings from more than 2000 persons, consisting of the vowels /a/, /i/ and /u/ produced at normal, high, low, and low-high-low pitch and the sentence “Guten Morgen, wie geht es Ihnen?”. In addition, the electroglottogram (EGG) signal is also stored for each case in a separate file. The length of the files with sustained vowels is between 1 and 3 s. All recordings are sampled at 50 kHz and their resolution is 16-bit.

Table 1 shows descriptive statistics of voice samples used in the study. The study used recordings of the sustained vowel sound /a/ from 126 normal speakers (63 elderly and 63 young) at neutral pitch, including 36 voice samples of “vox senilis”. The voice signals were classified into two different groups, namely those from young and elderly subjects, based on a recent publication [2,5]. This paper also analyzed voice differences by comparing young and elderly subjects according to gender. Thus, this work subdivided the voice signals into four subcategories. Group 1 consisted of 24 young men between the ages of 22 and 59 years (mean age = 39.13). Group 2 consisted of 24 elderly men between the ages of 60 and 89 years (mean age = 72.2). Group 3 included 39 young women between the ages of 20 and 58 years (mean age = 39.12). Finally, Group 4 consisted of 39 elderly women between the ages of 60 and 87 years (mean age = 70.7). According to the gender innovation analysis methods [19,20], in order to match the number of elderly and young voice signals, the experiments were conducted using the dataset extracted from the SVD. The number of elderly voices used this study was the maximum number of elderly voices available in SVD. Although the author admits that the amount of data is small to use deep learning algorithms, there are also other papers [9,30,31] experimenting with deep learning algorithms using a small amount of data.

Table 1. Descriptive statistics of voice samples used in the study.

	Number of Samples	Mean Age	Standard Deviation Age
Men (age)			
22–59	24	39.13	12.76
60–89	24	72.2	7.28
Women (age)			
20–58	39	39.12	12.57
60–87	39	70.7	8.27

2.2. Feature Parameters

The mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstrum coefficients (LPCCs) are well known in speech signal processing, and thus, they will not be discussed in detail here [32,33]. More information about their application and derivation can be found in a number of previous studies [14,15,28]. In this work, 20-dimensional MFCCs and 20-dimensional LPCCs were extracted from a 40 ms window signal using a 20 ms frameshift [28].

Among the various higher-order statistics (HOS), the third- and fourth-order cumulants were used as characteristic parameters in this study. These parameters are called normalized skewness (γ_3) and normalized kurtosis (γ_4) and they are defined as shown in Equation (1). In this paper, kurtosis is, to be exact, excess kurtosis. Skewness and kurtosis in the time domain have shown massive potential as classification indices for pathological signals [34,35]. They worked well in the classification between elderly and young voices, as shown in a previous study [5].

$$\gamma_3 = \frac{\sum_{n=1}^N (x_n - \mu)^3}{(N-1)\sigma^3}, \quad \gamma_4 = \frac{\sum_{n=1}^N (x_n - \mu)^4}{(N-1)\sigma^4} - 3 \quad (1)$$

where x_n is the n^{th} sample value, N is the number of samples, and μ and σ represent the mean and standard deviation, respectively.

Skewness indicates the symmetry of the probability density function (PDF) of the amplitude of a time series. A normal distribution has a skewness of zero, which indicates a time series with an equal number of large and small amplitude values. If the skewness value is between -0.5 and 0.5 , the data are fairly symmetrical. If the skewness value is between -1 and -0.5 or between 0.5 and 1 , the data are moderately skewed. If the skewness value is positive, it indicates a time series with many small values and few large values that is positively skewed (right tail). On the contrary, a time series with many large values and few small values is negatively skewed (left tail), and the skewness values are negative. Kurtosis measures the tailedness of the PDF of a time series. The kurtosis value is often compared with a normal distribution, which is equal to zero in Equation (1). If the kurtosis value is greater than zero, then the time series has relatively sharp peaks and heavier tails than a normal distribution. On the contrary, if the kurtosis is less than zero, then the time series has relatively flat peaks and lighter tails than a normal distribution [28,36].

In this study, the parameters of the 20-dimensional MFCCs and the 20-dimensional LPCCs were extracted from a 40 ms window signal using a 20 ms frameshift. The same settings have been applied in many previous studies [28,37]. Parameters such as MFCC + skewness and LPCC + kurtosis were created by appending one parameter to the original 20-dimensional MFCCs and LPCCs, thus producing 21 dimensions. MFCC + skewness + kurtosis and LPCC + skewness + kurtosis had 22 dimensions. Finally, the MFCC + delta + skewness + kurtosis parameter, which is made up of the 20-dimensional MFCCs, the 20-dimensional MFCC deltas, skewness, and kurtosis, had 42 dimensions. Therefore, the total number of dimensions ranged from 20 to 42. All voice data were grouped into training (70% of the data) and test (30%) sets to implement all methods. For a five-fold cross validation scheme, each set was randomly selected from the subset [35,38,39]. The whole process was repeated 10 times and the results were averaged. The results are shown as the mean and standard deviation.

2.3. Deep Learning Classifiers

FNN and CNN are well-known deep learning classifiers and, thus, will not be discussed in detail here. More information about their application and derivation can be found in previous studies [18,28]. This work used the FNN with two feed-forward layers to solve the binary classification problem of the classification of healthy young and elderly voices. The first layer was followed by rectified linear unit (ReLU) activation, and the last layer was followed by softmax activation. The selected parameters of FNN values are shown in Table 2.

Table 2. Parameters of FNN.

Parameter	Value
Loss function	Tanh
Optimization algorithm	SGD + momentum
Regularization	L2
Mismatch propagation	BPTT
Minibatch size	100 samples
Learning rate	0.001 or 0.00001 exponential attenuation
Loss function	Cross-entropy
Weights for samples	Yes

For the CNN, this work used two convolutional layers and two feed-forward layers. Dropout with a probability of 0.05 and batch normalization were applied to all layers except the feed-forward layers. Max-pooling and average pooling operations performed between the convolutional layers to downsample the intermediate representations in time and to add some time invariance in the process. This study also used a filter size of eight frames and a stride length of one in every convolutional layer with max-pooling at a pool

size of eight and a stride length of eight between the first and second convolutional layers, and average pooling at a pool size of eight and a stride length of two between the other layers. The output of the last fully connected layer with a size of 64 was fed to a two-way softmax classifier, which classifies a patch into one of the two classes. The parameter details used in CNN are shown in Table 3.

Table 3. Parameters of CNN.

Parameter	Value
Activation function	ReLU
Kernel size	(2, 2)
Optimizer	Adam
Epoch	100
Loss function	Cross-entropy
Dropout	0.05
Pooling window	Max-pooling (2, 2)
Neurons at the dense layer	512

3. Experimental Results and Discussion

A histogram is an approximate representation of the distribution of numerical data. To configure a histogram, the first step is to create the value range of a “bin”. That is, it divides the entire range of values into a series of intervals and then calculate the number of values contained in each interval [36]. A histogram can be thought of as simple kernel density estimates that smooth the frequency of bins by using kernels. This usually results in a smoother probability density function that more accurately reflects the distribution of the underlying variable. The density estimate can be displayed instead of a histogram, and is usually drawn as a curve rather than a set of boxes [36].

Figure 1 shows the histograms and curves of MFCC data extracted from elderly and young voices to provide a better visualization of elderly and young voice signals from men and women for the /a/ vowel. In Figure 1a, overall similar tendencies are observed in both older and younger voice signals in both the histograms and curves of MFCC for female /a/ vowel samples. As shown in Figure 1b,c, the histograms and curves of young voice signals are more concentrated near zero than older voice signals for male and mixed /a/ vowel samples. Figure 2 shows the histograms and curves of LPCC data extracted from elderly and young voice signals of men and women for the /a/ vowel. In Figure 2a, the histogram and curve estimated for elderly voice signals tend to show a similar pattern compared to those extracted from young voice signals. In contrast, the histogram and curve extracted from elderly voice signals tend to be slightly more distributed near zero than those extracted from young voice signals, as shown in Figure 2b. A similar tendency towards symmetry is observed for both older and younger voice signals in both the histograms and curves of LPCC for mixed female and male /a/ samples. In Figures 1 and 2, the distribution of MFCC and LPCC data for female, male, and mixed sampled have similar patterns. Figure 3 shows the histograms and curves of skewness extracted from elderly and young voice signals by men and women for the /a/ vowel. In female /a/ vowel samples (Figure 3a), the skewness extracted from elderly voice signals tends to be distributed above zero. There is a tendency to be positively skewed with moderate skewness. However, the skewness extracted from young voice signals tends to be distributed near and below zero compared with elderly voice signals. That is, there seems to be a fairly symmetrical and negatively skew with a left tail. In male /a/ vowel samples (Figure 3b), similar tendencies can be observed for both elderly and young voice signals in both the histograms and curves, but the histogram and curve of skewness extracted from young voice signals have a longer distribution and larger values. There seems to be a fairly symmetrical and moderate skew in both elderly and young voice signals. Overall, in Figure 3a,b, it can be seen that the skewness extracted from female voice signals is distributed above zero compared with the skewness extracted from male voice signals. Clearly, the skewness distribution in the two figures

shows a different pattern. In the mixed female and male /a/ vowel samples (Figure 3c), the histogram and curve patterns also have a distinctly different pattern compared to female and male /a/ vowel samples. In female /a/ vowel samples (Figure 4a), the histograms and curves of young voice signals tend to have a longer kurtosis distribution, and the kurtosis estimated for elderly voice signals tends to be shorter and wider than that of young voice signals. However, in terms of kurtosis, the histograms and curves for both voice signals tend to have relatively flat peaks and lighter tails than a normal distribution. On the other hand, in case of male /a/ vowel samples (Figure 4b), the kurtosis estimated in elderly voice signals tends to be longer and wider than that of young voice signals. Because the histograms of the kurtosis estimated for elderly voice signals tends to be less than zero, the tendency has relatively flat peaks and lighter tails than a normal distribution. In the case of the histograms of the kurtosis extracted from young voice signals, some of them tend to have values greater than zero. Therefore, it can be seen that some of them have sharp peaks and heavier tails than a normal distribution, and most of them have relatively flat peaks and lighter tails. In the mixed female and male /a/ vowel samples (Figure 4c), the histogram and curve have a different pattern to those in Figure 4a for both elderly and young voice signals.

Table 4 shows the classification results obtained by the MFCC, skewness, and kurtosis parameters and by the deep learning methods (FNN and CNN). In Table 4, for the FNN classifier, when using male /a/ vowel samples, the combination of MFCC, skewness, and kurtosis showed the best performance (87.50%) for classifying elderly and young voices. In the mixed data samples from women and men, an average accuracy of 87.50% was also obtained by using MFCC + skewness. In addition, the average performance was 75.00% for female data samples with the MFCC parameter. For the CNN classifier, the best accuracy (87.50%) for classifying elderly and young voices was obtained with the combination of MFCC and skewness in the case of male data samples; in mixed and female data samples, 78.95% and 75.00% accuracy were obtained with the MFCC + skewness + kurtosis and MFCC + skewness parameters, respectively.

Figure 1. Histograms of the 20-dimensional MFCCs extracted from elderly and young voices.

Figure 2. Histograms of the 20-dimensional LPCCs extracted from elderly and young voices.

Figure 3. Histograms of skewness extracted from elderly and young voices.

Figure 4. Histograms of kurtosis extracted from elderly and young voices.

Table 4. Classification results obtained by MFCC, skewness, and kurtosis parameters and the two different deep learning methods.

		MFCC	MFCC + S	MFCC + K	MFCC + SK
FNN					
/a/	Women and men	81.58 ± 3.30	87.50 ± 2.31	78.95 ± 2.07	81.58 ± 3.72
	Women	75.00 ± 4.12	75.00 ± 3.67	75.00 ± 2.95	75.00 ± 3.61
	Men	81.25 ± 3.02	87.50 ± 3.23	87.50 ± 3.28	87.50 ± 3.95
CNN					
/a/	Women and men	78.95 ± 3.73	73.68 ± 2.22	73.68 ± 2.71	78.95 ± 6.35
	Women	70.83 ± 3.54	75.00 ± 3.95	70.83 ± 4.39	75.00 ± 2.81
	Men	81.25 ± 3.55	87.50 ± 6.87	81.25 ± 7.33	75.00 ± 5.14

S: skewness; K: kurtosis; SK: skewness and kurtosis.

Table 5 shows the classification results obtained by the MFCC delta, skewness, and kurtosis parameters for FNN and CNN. These results were based on the MFCC delta parameter. In Table 5, utilization of the FNN classifier with the MFCC + delta + skewness parameters obtained the highest accuracy (93.75%) in male /a/ vowel samples compared with the other classifier and parameter combinations. Additionally, 78.95% and 79.17% accuracy were obtained, respectively, by the MFCC + delta + skewness + kurtosis and MFCC + delta parameters with the FNN classifier for mixed and female data samples. For the CNN classifier, 78.95%, 75.00%, and 81.25% accuracy was achieved with MFCC + delta, MFCC + delta + kurtosis, and MFCC + delta + skewness for the mixed, female, and male data samples, respectively.

Table 5. Classification results obtained by MFCC delta, skewness, and kurtosis parameters and the two different deep learning methods.

		MFCC + Delta	MFCC + Delta + S	MFCC + Delta + K	MFCC + Delta + SK
FNN					
/a/	Women and men	76.32 ± 3.47	78.95 ± 1.58	76.32 ± 1.94	78.95 ± 3.54
	Women	79.17 ± 2.95	75.00 ± 2.95	75.00 ± 3.61	75.00 ± 2.94
	Men	87.50 ± 7.10	93.75 ± 6.59	93.75 ± 5.15	93.75 ± 4.37
CNN					
/a/	Women and men	78.95 ± 3.46	76.32 ± 4.07	78.95 ± 2.82	73.68 ± 2.61
	Women	75.00 ± 2.91	75.00 ± 3.95	75.00 ± 4.91	70.83 ± 2.81
	Men	68.75 ± 5.27	81.25 ± 8.04	81.25 ± 4.88	75.00 ± 3.29

S: skewness; K: kurtosis; SK: skewness and kurtosis.

Table 6 shows the classification results obtained by the LPCC, skewness, and kurtosis parameters for FNN and CNN. These results were based on the LPCC delta parameter. For the FNN classifier, good accuracy (83.33%) was achieved for classifying elderly and young voices with the LPCC parameter in the case of female data samples. In mixed and male data samples, 78.95% and 81.25% accuracy were achieved by the LPCC and LPCC + kurtosis parameters, respectively. For the CNN classifier, similar results (75.00%) were shown for female and data samples with LPCC + kurtosis and LPCC + skewness + kurtosis, respectively; 63.16% was achieved with LPCC + kurtosis in mixed female and male data samples.

Table 6. Classification results obtained by LPCC, skewness, and kurtosis parameters and the two different deep learning methods.

		LPCC	LPCC + S	LPCC + K	LPCC + SK
FNN					
/a/	Women and men	78.95 ± 3.73	78.95 ± 2.54	73.68 ± 3.08	76.32 ± 3.55
	Women	83.33 ± 6.24	79.17 ± 2.95	79.17 ± 3.03	79.17 ± 3.02
	Men	81.25 ± 3.23	81.25 ± 4.37	81.25 ± 6.07	81.25 ± 4.22
CNN					
/a/	Women and men	63.16 ± 3.29	63.16 ± 3.88	63.16 ± 4.47	68.42 ± 4.26
	Women	66.67 ± 2.81	70.83 ± 2.91	75.00 ± 4.41	70.83 ± 2.36
	Men	68.75 ± 4.37	68.75 ± 4.22	68.75 ± 5.47	75.00 ± 4.93

S: skewness; K: kurtosis; SK: skewness and kurtosis.

For male /a/ vowel data samples, utilization of the FNN classifier, MFCC + delta + skewness, and MFCC + delta + kurtosis parameters obtained the highest accuracy (93.75%) compared with the combination of CNN, FNN, and various parameters. The best accuracy (87.50% and 83.33%) was obtained with MFCC + skewness and the LPCC parameters for the FNN classifier in mixed and female data samples, respectively. Another finding is that the fusion of the skewness and kurtosis parameters had a positive effect on the overall accuracy of the classification. Most results were better with FNN than CNN in all mixed, female, and male data samples.

Figure 5 summarizes the performance trends in terms of network, gender, and various parameters. In this figure, it can be seen that FNN performed well in most cases. In terms of gender, the use of male data samples demonstrated good performance with most parameters. Therefore, rather than experimenting with mixed female and male data samples, the author recommends experimenting with and constructing a system that divides the data into female and male individuals. Figure 5c shows that the highest performance was 93.75%, obtained by using the parameters that added skewness and kurtosis to the MFCC and MFCC delta parameters. Of these, the addition of skewness showed the best performance. The figure also shows the better performance of using MFCC

rather than LPCC. In conclusion, it can be seen that it is appropriate to use the MFCC and skewness parameters based on FNNs to achieve good classification of elderly and young voice signals.

Figure 5. Comparison of performances in terms of networks, gender, and various parameters (S: skewness; K: kurtosis; SK: skewness and kurtosis; W&M: mixed female and male data; M: male data; W: female data). Gray bars indicate FNN.

4. Conclusions

It is important to objectively distinguish between elderly and young voices because physiological aging leads to specific changes in the anatomy and physiology of all structures involved in the production and modulation of the human voice [2,4–6,10,13]. Currently, most systems use all data at once, from children's voices to elderly voices, so when older people use a voice-activated system, their voices are often not recognized well [40–42]. Therefore, in order to develop a system for recognizing the voice of the elderly, it is necessary to understand the characteristics of changes in vocal cord tissue due to anatomical or physiological aging [3,15,16,18].

This study developed a classifier that distinguishes elderly voices from young voices by using deep learning algorithms and MFCCs, LPCCs, skewness, and kurtosis parameters. This study also compared the prediction performance of FNN and CNN based on the combination of various parameters to find out which combinations are best for classifying elderly and young voices. The results of this study revealed that the performance of FNN was higher than that of CNN. Therefore, future studies are necessary to compare the predictive performance estimated by diverse evaluation methods using big data of elderly and young voice signals, such as sensitivity, specificity, etc., suitable for achieving the analysis's objective, rather than using one performance index such as accuracy.

Another key finding of this study was that skewness, one of the higher-order statistics (HOS), was the most important predictor for identifying elderly voice signals in the deep learning models (FNN and CNN). Although the fusion of the skewness and kurtosis parameters had a positive effect on the overall accuracy of the classification, the highest performance (93.75%) appeared when the skewness was added to the MFCC and MFCC delta parameters. Specifically, the study found that skewness was more effective in combination with MFCC and MFCC delta parameters, and kurtosis was more effective in combination with LPCC parameters.

The last important discovery of this study is as follows. Most parameters estimated from male data samples demonstrated good performance. Rather than using mixed female and male data, this study recommends the development of separate systems that represents the best performance obtained for each optimized parameter by using independent data for males and females.

In conclusion, the purpose of this study was to provide HOS parameters and deep learning methods for the classification of young and elderly voice signals; clarify the network-, gender-, and parameter-related differences regarding classification performance; and find the best combination for achieving good classification of elderly and young voice signals. The results of this study will be the basis for preparing a voice interface optimized for the elderly, such as a smart mirror system and a conversation recognition system for the elderly. Future studies will be based on the spread of elderly speech processing, along with the parameters and deep learning algorithms that reflect important information about elderly voice signals to achieve high-quality classification performance of younger and older voice signals using large datasets.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2017R1A2B4011373). The sponsor had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Asci, F.; Costantini, G.; Di Leo, P.; Zampogna, A.; Ruoppolo, G.; Berardelli, A.; Saggio, G.; Suppa, A. Machine-Learning Analysis of Voice Samples Recorded through Smartphones: The Combined Effect of Ageing and Gender. *Sensors* **2020**, *20*, 5022. [CrossRef]
2. Mezzedimi, C.; Di Francesco, M.; Livi, W.; Spinosi, M.C.; De Felice, C. Objective Evaluation of Presbyphonia: Spectroacoustic Study on 142 Patients with Praat. *J. Voice* **2017**, *31*, 257.e25–257.e32. [CrossRef]
3. Gonçalves, T.M.; Dos Santos, D.C.; Pessin, A.B.B.; Martins, R.H.G. Scanning Electron Microscopy of the Presbylarynx. *Otolaryngol. Head Neck Surg.* **2016**, *154*, 1073–1078. [CrossRef]
4. Hammond, T.H.; Gray, S.D.; Butler, J.; Zhou, R.; Hammond, E. Age- and gender-related elastin distribution changes in human vocal folds. *Otolaryngol. Head Neck Surg.* **1998**, *119*, 314–322. [CrossRef]
5. Choi, H.J.; Lee, J.Y. Comparative Study between Healthy Young and Elderly Subjects: Higher-Order Statistical Parameters as Indices of Vocal Aging and Sex. *Appl. Sci.* **2021**, *11*, 6966. [CrossRef]
6. Kahane, J.C. Anatomic and physiologic changes in the aging peripheral speech mechanism. In *Aging: Communications Processes and Disorders*; Beasley, D.S., Davis, A., Eds.; Grune and Stratton: New York, NY, USA, 1981; pp. 21–45.
7. Lee, S.Y. The Overall Speaking Rate and Articulation Rate of Normal Elderly People. Graduate Program in Speech and Language Pathology. Master's Thesis, Yonsei University, Seoul, Korea, 2011.
8. Kadiri, S.R.; Gangamohan, P.; Gangashetty, S.V.; Alku, P.; Yegnanarayana, B. Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference. *Circuits Syst Signal Process* **2020**, *39*, 4459–4481. [CrossRef]
9. Kadiri, S.R.; Gangamohan, P.; Yegnanarayana, B. Discriminating Neutral and Emotional Speech Using Neural Networks. *ICON 2014*. Available online: <https://aclanthology.org/W14-5132> (accessed on 13 May 2018).
10. Byeon, H. Comparing Ensemble-Based Machine Learning Classifiers Developed for Distinguishing Hypokinetic Dysarthria from Presbyphonia. *Appl. Sci.* **2021**, *11*, 2235. [CrossRef]
11. Fukuda, M.; Nishimura, R.; Nishizaki, H.; Iribe, Y.; Kitaoka, N. A New Corpus of Elderly Japanese Speech for Acoustic Modeling, and a Preliminary Investigation of Dialect-Dependent Speech Recognition. In Proceedings of the 2019 22nd Conference of the Oriental COCOSA International Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSA), Cebu, Philippines, 25–27 October 2019; pp. 1–6.

12. Werner, L.; Huang, G.; Pitts, B.J. Automated Speech Recognition Systems and Older Adults: A Literature Review and Synthesis. *Proc. Hum. Factors Ergon. Soc. Annu. Meeting*. **2019**, *63*, 42–46. [\[CrossRef\]](#)
13. Ambrosini, E.; Caielli, M.; Milis, M.; Loizou, C.; Azzolino, D.; Damanti, S.; Bertagnoli, L.; Cesari, M.; Moccia, S.; Cid, M.; et al. Automatic speech analysis to early detect functional cognitive decline in elderly population. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019; pp. 212–216. [\[CrossRef\]](#)
14. Forero Mendoza, L.A.; Cataldo, E.; Vellasco, M.; Silva, M. Classification of Voice Aging Using Parameters Extracted from the Glottal Signal. In *Artificial Neural Networks—ICANN 2010*; Diamantaras, K., Duch, W., Iliadis, L.S., Eds.; ICANN 2010; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6354.
15. Collado-Villaverde, A.; R-Moreno, M.D.; Barrero, D.F.; Rodriguez, D. Machine Learning Approach to Detect Falls on Elderly People Using Sound. In *Advances in Artificial Intelligence: From Theory to Practice*; Benferhat, S., Tabia, K., Ali, M., Eds.; IEA/AIE 2017; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10350.
16. Vacher, M.; Aman, F.; Rossato, S.; Portet, F. Development of Automatic Speech Recognition Techniques for Elderly Home Support: Applications and Challenges. In *Human Aspects of IT for the Aged Population*; Zhou, J., Salvendy, G., Eds.; Design for Everyday Life ITAP 2015; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9194.
17. Solano, A.; López, G.; Guerrero, L.; Quesada, L. User Experience Evaluation of Voice Interfaces: A Preliminary Study of Games for Seniors and the Elderly. *Proceedings* **2019**, *31*, 65. [\[CrossRef\]](#)
18. Yang, G.; Yang, J.; Sheng, W.; Junior, F.E.F.; Li, S. Convolutional Neural Network-Based Embarrassing Situation Detection under Camera for Social Robot in Smart Homes. *Sensors* **2018**, *18*, 1530. [\[CrossRef\]](#)
19. Lee, J.Y. Gender analysis in elderly speech signal processing. *J. Digital Converg.* **2018**, *16*, 351–356. [\[CrossRef\]](#)
20. Lee, J.Y. Elderly speech signal processing: A systematic review for analysis of gender innovation. *J. Converg. Inf. Technol.* **2019**, *9*, 148–154. [\[CrossRef\]](#)
21. Ferrand, C.T. Harmonics-to-noise ratio: An index of vocal aging. *J. Voice* **2002**, *16*, 480–487. [\[CrossRef\]](#)
22. da Silva, P.T.; Master, S.; Andreoni, S.; Pontes, P.; Ramos, L.R. Acoustic and long-term average spectrum measures to detect vocal aging in women. *J. Voice* **2011**, *25*, 411–419. [\[CrossRef\]](#)
23. Maslan, J.; Leng, X.; Rees, C.; Blalock, D.; Butler, S.G. Maximum phonation time in healthy older adults. *J. Voice* **2011**, *25*, 709–713. [\[CrossRef\]](#)
24. Schaeffer, N.; Knudsen, M.; Small, A. Multidimensional Voice Data on Participants with Perceptually Normal Voices from Ages 60 to 80: A Preliminary Acoustic Reference for the Elderly Population. *J. Voice* **2015**, *29*, 631–637. [\[CrossRef\]](#)
25. Linville, S.E.; Rens, J. Vocal Tract Resonance Analysis of Aging Voice Using Long-Term Average Spectra. *J. Voice* **2001**, *15*, 323–330. [\[CrossRef\]](#)
26. Peterson, E.A.; Roy, N.; Awan, S.N.; Merrill, R.Y.; Banks, R.; Tanner, K. Toward validation of the cepstral spectral index of dysphonia (CSID) as an objective treatment outcomes measure. *J. Voice* **2013**, *27*, 401–410. [\[CrossRef\]](#)
27. Byeon, H.; Jin, H.; Cho, S. Characteristics of hypokinetic dysarthria patients' speech based on sustained vowel phonation and connected speech. *Int. J. Serv. Sci. Technol.* **2016**, *9*, 417–422. [\[CrossRef\]](#)
28. Lee, J.-Y. Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database. *Appl. Sci.* **2021**, *11*, 7149. [\[CrossRef\]](#)
29. William, J.B.; Manfred, P. Saarbrücken Voice Database. Institute of Phonetics, Univ. of Saarland. Available online: <http://www.stimmdatenbank.coli.uni-saarland.de/2007> (accessed on 13 May 2018).
30. Oh, S.L.; Vichresh, J.; Ciaccio, E.J.; Yuvaraj, R.; Acharya, U.R. Deep Convolutional Neural Network Model for Automated Diagnosis of Schizophrenia Using EEG Signals. *Appl. Sci.* **2019**, *9*, 2870. [\[CrossRef\]](#)
31. Phang, C.R.; Ting, C.M.; Noman, F.; Ombao, H. Classification of EEG-Based Brain Connectivity Networks in Schizophrenia Using a Multi-Domain Connectome Convolutional Neural Network. *arXiv* **2019**, arXiv:1903.08858.
32. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [\[CrossRef\]](#)
33. Thomas, D.W. *Linear Prediction of Speech*; Markel, J.D., Gray, A.H., Eds.; Springer: Berlin/Heidelberg, Germany; New York, NY, USA, 1976; p. xii+288 Cloth; ISBN 3-540-07563-1.
34. Nemer, E.; Goubran, R.; Mahmoud, S. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 217–231. [\[CrossRef\]](#)
35. Lee, J.-Y.; Jeong, S.; Hahn, M. Pathological Voice Detection Using Efficient Combination of Heterogeneous Features. *IEICE Trans. Inf. Syst.* **2008**, *E91-D*, 367–370. [\[CrossRef\]](#)
36. Wikipedia. Histogram. Available online: <https://en.wikipedia.org/wiki/Histogram> (accessed on 9 June 2021).
37. Adam, P.; Sam, G.; Soumith, C.; Gregory, C. Automatic differentiation in PyTorch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
38. Lee, J.Y.; Hahn, M. Automatic Assessment of Pathological Voice Quality Using Higher-Order Statistics in the LPC Residual Domain. *EURASIP J. Adv. Signal Process.* **2010**, *2009*, 1–8. [\[CrossRef\]](#)
39. Wang, X.; Zhang, J.; Yan, Y. Discrimination between Pathological and Normal Voices Using GMM-SVM Approach. *J. Voice* **2011**, *25*, 38–43. [\[CrossRef\]](#)

-
40. Song, Y.-K. Prevalence of Voice Disorders and Characteristics of Korean Voice Handicap Index in the Elderly. *Phon. Speech Sci.* **2012**, *4*, 151–159. [[CrossRef](#)]
 41. Lee, S.J.; Kwon, S.I. Elderly speech analysis for improving elderly speech recognition. *J. Korean Inst. Inf. Sci. Eng.* **2014**, *32*, 16–20.
 42. Jeong, J.H.; Jang, J.H.; Moon, M. Development of AI Speaker with Active Interaction Customized for the Elderly. *J. Korean Inst. Electron. Commun. Sci.* **2020**, *15*, 1223–1230.