Journal of KIIT. Vol. 22, No. 4, pp. 91-97, Apr. 30, 2024. pISSN 1598-8619, eISSN 2093-7571 91
http://dx.doi.org/10.14801/jkiit.2024.22.4.91

Vocabulary Recognition Rate Enhancement using Clustering Model and Non-parametric Correlation Coefficient

Sang Yeob Oh*

요 약

음성인식 기술은 언어를 기계적으로 이해하고 처리하는 기술이며, 최근 인공 지능 기반 음성 처리로 인해 음성에 대한 인식률은 향상되었지만 사용 환경에서의 잡음으로 인해 원래 음성 본연의 정확한 음성을 인식할 수 없다. 또한 인공지능 환경에서 많은 모델 파라미터들로 인한 데이터 부족 문제가 수반되며 모델별 훈련 데이터의 수가 균일하지 않은 경우가 일반적이므로 이를 해결하기 위해 개선된 군집화 모델링과 선택적 음성 특징 추출에서 비모수 상관계수를 이용하여 청각 구성의 상관 관계를 구성하였으며, 제안된 방법의 실험에서 평균 왜곡율이 0.36dB 감소되어 인식률이 개선된 것을 확인하였다.

Abstract

Vocabulary recognition technology is a technology that mechanically understands and processes language. Although the speech recognition rate has recently improved due to artificial intelligence-based speech processing, the original voice cannot be accurately recognized due to noise in the usage environment. In addition, in the artificial intelligence environment, there is a data shortage problem due to many model parameters, and the number of training data for each model is usually uneven. To solve this, improved clustering modeling and selective voice feature extraction are used. A correlation map of auditory elements was constructed using the Non-parametric Correlation Coefficient , and in an experiment of the proposed method, it was confirmed that the result average distortion of separation was reduced by 0.36dB, improving the recognition rate.

Keywords
vocabulary recognition, clustering modeling, feature extraction, correlation coefficient

* Professor in the Department of Computer Engineering, Gachon University
- ORCID: ttps://orcid.org//0000-0002-8002-9588

· Received: Mar. 14, 2024, Revised: Apr. 11, 2024, Accepted: Apr. 14, 2024

· Corresponding Author: Sang Yeob Oh
Dept. of Computer Engineering, Gachon University, Korea
Tel.: + 82-31-750-5798, Email: syoh@gachon.ac.kr

Ⅰ. Introduction

Due to the advancement of AI and mobile hardware, the voice recognition system supports a variety of voice recognition in wireless IP (Internet Protocol) networks, and removes noise signals generated from voice signals that are basically processed in voice recognition to restore the original voice signal. Voice recognition must be performed accurately[1]-[4][10]. Speech recognition has been significantly improved with the application of recent AI based technologies such as DNN(Deep Neural Network), RNN(Recurrent Neural Network), CNN(Convolutional Neural Network), TDNN(Time Delay Neural Network), and Kaldi, but noise signals are reduced to existing voice recognition. It can have similar voice signal characteristics by adding it to the signal, and the characteristics of the voice signal change over time with respect to the existing voice signal, creating unstable noise that can generate noise, and the noise of the voice signal can be accurately When not classified, voice recognition is performed with reduced accuracy due to the noise in the voice signal having an uneven threshold. Therefore, noise removal technology is required to increase the speech recognition rate when processing speech with environmental noise, and noise removal and feature extraction for model estimation technology must be utilized in this process.

This paper uses improved cluster modeling to improve the precision of model noise by re-estimating it by increasing the number of mixing elements in each state. In addition, in order to improve the recognition rate of the voice recognition system, we propose feature extraction techniques that extract only the output voice from the input voice by mixing several voices and noise. For feature extraction for selective voice extraction, voice features were modeled by constructing a correlation map of auditory elements using continuity of the time and similarity between channels, and a method of extracting voice features using non -parametric correlation coefficient was used. The proposed method was analyzed and tested using the Aurora 2.0 database, and was composed of 20 area names and 20 subway station names in Seoul to measure the performance of improved clustering modeling and noise removal using Non-parametric Correlation Coefficient., In the performance evaluation of the proposed model, the results processed using cross-correlation coefficients to evaluate accuracy for speech were compared with the proposed method to evaluate the recognition rate, and it was confirmed that there was an improvement of 0.36dB in the difference in the average value of the correlation coefficients for them. did. By applying this method, the performance of AI-based voice recognition is supplemented and the precision of noise processing for voice recognition is improved. This paper is structured as follows: Chapter 1 is an introduction, Chapter 2 analyzes existing related studies, Chapter 3 explains the proposed method, Chapter 4 presents experiments and analysis, and finally Chapter 5 presents conclusions.

Ⅱ. Related work

2.1 Model sharing and clustering

The main problem in creating using shared structures is how to obtain a general model that represents the characteristics of the data with minimal complexity. By using a shared structure, the increasing number of parameters can be reduced, thereby increasing training efficiency, and by reducing the number of model parameters, the amount of computation required for recognition can be reduced. The purpose of shared state acoustic modeling is to effectively include the acoustic and phonetic knowledge of the language into the model according to a certain maximum probability technique. Fig. 1 shows a form of clustering by sharing states.

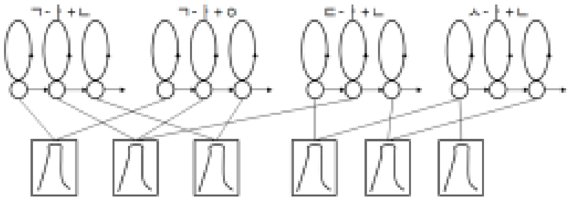

Fig. 1. Cluster tied-state

Let any S be a set of states. If $L(S)$ is the log likelihood of S and is generated from F, a set of frames of training data, under the assumption that all states in the set S are grouped, then the common average is $\mu(S)$ over dispersion $\Sigma(S)$ will be shared. Also, assuming that the alignment of frames for each state of the bound states does not change $L(S)$, An approximate expression such as the following equation (1) can be written[5].

$$L(S)=\sum_{f=F}\sum_{s=S}\log(\Pr(o_f;\mu(S)))^{*}G \quad (1)$$

$(Here, G=\gamma_s(o_f))$
G : $Posteriorprobabilitgeneratedbystates$
o_f : $observationframe$

Gaussian distributions with the central phoneme are collected to create one large pool, and the observation probability of each triphone uses Gaussians belonging to the common pool and varies the weights to reflect the acoustic characteristics of the triphone. The method of constructing a common Gaussian is to collect and cluster speech feature vectors with the same central phoneme.

2.2 Feature extraction

The feature extraction process is a process of obtaining the time axis periodicity feature of the filter bank frequency response coefficient and the frequency similarity feature between channels of the frequency axis based on the cochleargram obtained by modeling the auditory organ. The traditional method of finding periodicity features uses the autocorrelation coefficient and is expressed as equation (2)[6].

$$\begin{aligned} &ACF(c,m,\tau) \\ &\quad=\sum_{k=0}^{K-1} h(c,m-k)\,h(c,m-k-\tau)\,w(k) \end{aligned} \quad (2)$$

Time m, there is a T-F unit in lag τ of the first channel to window function $\omega(k)$ is performed by multiplying[7].

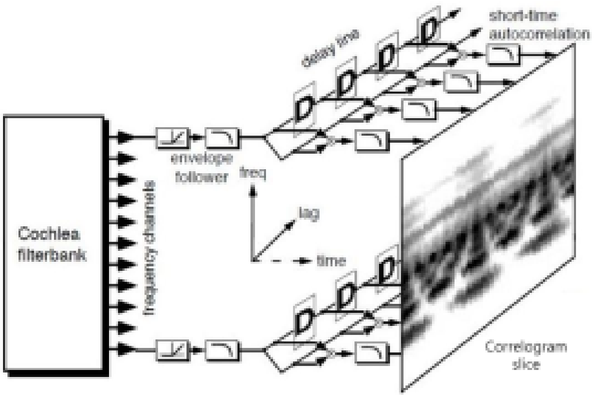

Fig. 2. Structure and scene of correlogram

The window function used in Frame uses a rectangular window function and is expressed as a correlogram expressing periodicity. Fig. 2 shows the correlogram system structure and scene using autocorrelation coefficient. Cross correlation coefficient is the input of autocorrelation and calculates the degree of correlation between channels to periodically determine similarity for patterns.

III. System model

For an improved clustering model, an initial set of model 3-state single phoneme models is created and trained. The state output distribution of a single phoneme is copied to initialize a set of trained triphone models using Baum-Welch re-estimation. The corresponding states are clustered for each set of triphones derived from the same phoneme. A representative state is selected from each result cluster, and the states within all clusters are grouped into the representative state. The precision of the model is improved by re-estimation by increasing the number of mixing elements in each state. In clustering modeling, an initial set of 3-state vocabulary models for the vocabulary is created and trained, and the state output distribution of the vocabulary is used to initialize a set of triphone models trained using Baum-Welch re-estimation. The corresponding states are clustered for each set of triphones derived from the same vocabulary, and a representative state is selected from each resulting clustering, and all states in the clustering are grouped into the representative state. The precision of the model is improved by re-estimation by increasing the number of mixing elements in each state.

The improved clustering model uses the input vocabulary to model the distribution density of the sample data set as a probability density function, and the Gaussian mixture model is a density estimation method that models the distribution density of the sample data set as a single probability density function, it is a method of modeling distribution. It is expressed as equation (3), which linearly combines the Gaussian probability density functions. In equation (3) $p(x|\omega_i, \theta_i)$ is the input data x about vocabulary about Expressing the second component parameter, Expresses a probability density function consisting of θ_i, $P(\omega_i)$ is processed by expressing it as a mixed weight.

$$p(x|\theta) = \sum_{i=1}^{M} p(x|\omega_i, \theta_i) P(\omega_i) \tag{3}$$

Additionally, a correlation map of auditory elements was constructed using continuity of the time and similarity between channels as feature extraction for selective voice extraction. The voice feature was modeled from the correlogram according to the constructed auditory elements, and the voice feature was extracted using the non-parametric correlation coefficient. For selective voice feature extraction, continuity of the time used an autocorrelation coefficient using the time axis periodicity feature of the filter bank frequency response coefficient based on the cochleargram, and a correlogram was constructed after extracting features based on an existing research method. Additionally, as a method to determine similarity between channels, features were extracted using Kendall's Tau, a non-parametric correlation, and a correlogram was constructed.

Kendall's Tau is calculated using the compatibility of two variables. It was used to check the compatibility of the increase or decrease in the value of other variables when the value of one variable increases or decreases for all possible combinations of two observed values, and is expressed as equation (4).

$$\tau = \frac{\sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}} \tag{4}$$

$$T_0 = n(n-1)/2$$
$$T_1 = \sum t_i (t_i - 1)/2$$
$$T_2 = \sum u_i (u_i - 1)/2$$

Where n represents the observed value of variables, t_i is the same at a given rank X. Indicates the observed value of u_i is the same at a given rank Indicates the observed value of Y. This sgn represents $sgn(z) = 1$ is $z > 0$, if $sgn(z) = 0$ is represent $z = 0$, if $sgn(z) = -1$ is represent $z < 0$. The calculated coefficient has a value between 0 and 1 through normalization. If the channel 's periodic pattern compatibility is high, a value close to 1 will be displayed, and in the opposite case, a value close to 0 will be displayed.

In the process of creating voice signal information through non-parametric correlation coefficient prediction, the process of calculating the feature extraction probability from the currently input voice signal information and updating the standard feature based on this is performed. Therefore, if the voice feature stops for several frames, the object is recognized as a voice feature. Conversely, there may also be an object that initially does not move and is recognized as noise, but then begins to move at some point. After completing the secondary matching, the remaining object that has not been matched to the previous frame is an object that has no movement and is recognized as noise in the current frame, and is an object that is not matched to the current frame and moves at some moment. In the former case, the information of the previous object is applied to the current frame, and in the latter case, it is registered as a new object. When matching for all objects is completed, the objects of the previous frame are exactly one-to-one matched with the objects of the current frame. Find speed, acceleration, and azimuth between each object with one-to-one matching. Because the time between successive frames is almost constant, speed becomes an element representing the distance an object has moved, and acceleration is a numerical value indicating how much that distance has changed.

Ⅳ. Performance experiment

The proposed method was analyzed and tested using the Aurora 2.0 database, and improved clustering modeling and Non-parametric Correlation Coefficient were used. The Aurora 2 database artificially adds additional noise to the voice signal and consists of noise voice for channel distortion. It is internationally recognized and is one of the most widely used voice data. The voice signal used an 8kHz sampling rate and 16bit to support real-time processing of noise signals, and the FFT size was 256 samples. A 1/2 overlapping section was used to eliminate voice short-circuiting, and a Hamming window was used to reduce signal distortion. For comparative analysis of speech signals, a Warner filter was used to process noise for speech, and for speech recognition experiments, 20 area names and 20 subway station names in Seoul were used. In the recognition experiment, the speaker who participated in the experiment pronounced the voice recognition list 5 times and applied a total of 100 words. The correlation coefficient calculation process for each channel for the voice was performed and the resulting data were compared. To evaluate the accuracy of the voice, the results processed using the cross-correlation coefficient for each voice are compared with the proposed method and are shown in the following Table 1, and the evaluation of the voice uses the SDR (Signal to Distotion Ratio)[9]. This was performed, and the formula for this is as follows.

$$SDR(dB) = 10\log_{10}\frac{\sum_{n=1}^{N}\left[x(n)-\hat{x}(n)\right]^2}{\sum_{n=1}^{N}x^2(n)} \quad (5)$$

In this equation, $x(n)$ represents the voice without noise reflected, $\hat{x}(n)$ refers to a voice signal with noise added, n is the time index, and shows the degree of distortion using equation (5).

Table 1. Compare with the results

Separate voice	Cross correlation	Proposed methode	Distortion ratio
1	2.11	2.10	0.21
2	2.31	1.31	0.73
3	5.36	1.79	2.87
4	6.04	1.67	3.61
5	1.78	1.39	-0.31
6	2.37	2.38	-0.67
7	1.37	1.93	0.27
8	6.01	5.17	0.67
9	2.09	2.06	-0.37
10	2.16	2.51	-0.23
11	1.31	1.59	-0.17
12	2.37	2.27	0.14
13	2.61	2.53	-0.03
14	2.11	2.14	-0.06
15	2.17	2.19	0.11
16	2.51	2.71	-0.06
17	2.11	2.11	0.27
18	2.80	2.51	0.11
19	2.47	2.31	0.05
20	2.17	2.14	-0.03
Average	2.71	2.24	0.36

A small value for distortion ratio means that the similarity of the compared voices is high. In Table 1, the difference between the correlation coefficient for the voice signal and the method applying the proposed method is high for the 3rd and 4th voice signals, but the proposed method can be seen that the correlations are evenly displayed, and only the 8th voice signal is high. As a result of evaluating the recognition rate of the proposed method, it was confirmed that the average difference in correlation coefficients for these was improved by 0.36dB. And this result shows a significant difference in performance results compared to the paper in [9], proving that research on improved methods to voice distortion is necessary.

V. Conclusions

In order to efficiently extract the features of the voice signal for each voice frame, improved clustering modeling was used, and the Non-parametric Correlation Coefficient was used in the feature extraction method for selective voice extraction. The noise removal method used was applied. For performance evaluation measurements, the proposed method was analyzed and tested using the Aurora 2.0 database, and in the performance evaluation of the proposed model, the results processed using the cross-correlation coefficient to evaluate accuracy for speech were compared with the proposed method. As a result of evaluating the recognition rate, it was confirmed that the average difference in correlation coefficients was improved by 0.36dB.

References

[1] Sang Yeob Oh, "Speech Recognition Performance Improvement using a convergence of GMM Phoneme Unit parameter and Vocabulary Clustering", Journal of Convergence for Information Technology, Vol. 10, No. 8, pp. 35-39, Aug. 2020. https://doi.org/10.22156/CS4SMB.2020.10.08.035.

[2] Sang Yeob Oh, "DNN based Robust Speech Feature Extraction and Signal Noise Removal Method Using Improved Average Prediction LMS Filter for Speech Recognition", Journal of Convergence for Information Technology, Vol. 11, No. 6, pp. 1-6, Jun. 2021. https://doi.org/10.22156/CS4SMB.2021.11.06.001.

[3] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning", IEEE/ACM Transactionson Audio, Speech and Language Processing, Aug. 2020. https://doi.org/10.48550/arXiv.2008.03648.

[4] A. Arango, J. P´erez, and B. Poblete, "Hate Speech Detection is Not as Easy as You May Think : A Closer Look at Model Validation", Information Systems, Vol. 105, pp. 101584, Mar. 2022. https://doi.org/10.1016/j.is.2020.101584.

[5] A. S. Manos and V. W. Zue, "A study on out-of-vocabulary word modeling for a segment-based keyword spotting system", Master Thesis, MIT, 1996.

[6] T. T. Pham, J. Y. Kim, S. Y. Na, and S. T. Hwang, "Robust Eye Localization for Lip Reading in Mobile Environment", Proc. of SCIS&ISIS in Japan, pp. 385-388, 2008. https://doi.org/10.14864/softscis.2008.0.385.0.

[7] T. T. Pham, M. G. Song, J. Y. Kim, S. Y. Na, and S. T. Hwang, "A Robust Lip Center Detection in Cell Phone Environment", Proceedings of IEEE Symposium on Signal Processing and Information Technology, Sarajevo, Bosnia and Herzegovina, pp. 390-395, Dec. 2008. https://doi.org/10.1109/ISSPIT.2008.4775724.

[8] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", IEEE Transactions on Neural Networks, Vol. 15, No. 5, pp. 1135-1150,

Sep. 2004. https://doi.org/10.1109/TNN.2004.832812.

[9] Sang Yeob Oh, "Noise Elimination Using Improved MFCC and Gaussian Noise Deviation Estimation", Journal of The Korea Society of Computer and Information, Vol. 28 No. 1, pp. 87-92, Jan. 2023. https://doi.org/10.9708/jksci.2023.28.01.087.

[10] S. G. Lee and S. Lee, "Data Augmentation for DNN-based Speech Enhancement", Journal of Korea Multimedia Society Vol. 22, No. 7, pp. 749-758, Jul. 2019. https://doi.org/10.9717/kmms.2019.22.7.749.

Authors

Sang Yeob Oh

1985 ~ 1989 : B.S. degrees in Computer engineering from Gachon University

1989 ~ 1999 : M.S. and Ph. D. degrees in Computer engineer-ing from KwangWoon University

1992 ~ present : Professor in the Department of Computer Engineering, Gachon University

Research interests : voice recognition & feature extract-ion, noise detect, multimedia data communication