Parts-Based Feature Extraction of Spectrum of Speech Signal Using Non-Negative Matrix Factorization

Jeong-Won Park, Chang-Keun Kim, Kwang-Seok Lee, Si-Young Koh and **Kang-In Hur**, *Member, Kimics*

Abstract—In this paper, we proposed new speech feature parameter through parts-based feature extraction of speech spectrum using Non-Negative Matrix Factorization (NMF). NMF can effectively reduce dimension for multi-dimensional data through matrix factorization under the non-negativity constraints, and dimensionally reduced data should be presented parts-based features of input data. For speech feature extraction, we applied Mel-scaled filter bank outputs to inputs of NMF, than used outputs of NMF for inputs of speech recognizer. From recognition experiment result, we could confirm that proposed feature parameter is superior in recognition performance than mel frequency cepstral coefficient (MFCC) that is used generally.

Index Terms—Non-Negative Matrix Factorization, Parts-based Feature Extraction, Mel-scaled Filter Bank Output.

I. INTRODUCTION

Consideration in the speech recognition performance effectively selects the recognition algorithm and the feature parameter. In general, most speech recognition systems used to select MFCC modeling human's vocal tract as the feature parameter of the speech signal.[8][9][11] Because it includes the whole characteristic of vocal tract, it has overlapped information between speech signals that have distinctive meaning. In this paper, we proposed new feature parameter that characteristic overlap is lower than MFCC using NMF that is parts-based feature extraction algorithm [Daniel D. Lee, H. Sebastian Seung, 1999].

Human's brain recognizes a whole object through parts-based feature.[1] Because weighted sums of parts-based features is presented to feature of the whole object, they can be obtained through iterative train using NMF under the non-negativity constraints.

In this experiment, we applied NMF algorithm to mel filter bank output of speech spectrum and used MFCC and parts-based feature obtained by NMF as inputs of recognizer. As a result of recognition experiment, we could verify effectiveness of this algorithm.

Manuscript received November 27, 2003.

Jeong-Won Park (phone: +82-51-200-6961, email: jwpark@donga.ac.kr), Chang-Keun Kim (email: chkkim@donga.ac.kr) and Kang-In Hur (email: kihur@mail.donga.ac.kr) are with the Department of Electronic Engineering, Dong-A University, Busan, Korea.

Kwang-Seok Lee (phone: +82-55-751-3333, email: kslee@jinju.ac.kr) is with the Department of Electronic Engineering, Jinju National University, Jinju, Korea.

Si-Young Koh (phone: +82-53-850-7164, email: kohsy@kiu.ac.kr) is with the School of Electronic Information and Communicatio n Engineering, Kyungil University, Daegu, Korea.

The organization of this paper is as follows. In Section II, we introduce basic idea and training rules of NMF algorithm and explain proposed feature extraction procedure in Section III. In Section IV, we discuss recognition result for two feature parameters and conclude in Section V.

II. NON-NEGATIVE MATRIX FACTORIZATION

A. Basic Idea

NMF is to find non-negative matrix W and H whose multiplication is approximately presented to V using matrix factorization under the non-negativity constraints as the following formula (1).

$$\begin{aligned} &V \approx WH \\ &V_{iu} \approx (WH)_{iu} = \sum_{a=1}^{r} W_{ia} H_{au} \\ &\text{all elements of } (V, W, H) \geq 0 \end{aligned} \tag{1}$$

V is input data matrix of $n \times m$ dimension and W is basis matrix of $n \times r$ dimension that has characteristic of weight, H is dimensionally reduced matrix for the input data matrix of $r \times m$ dimension. Where, n is dimension of the input data and m is numbers of the input data set, r is dimension of reduced input data, r is generally chosen so that $(n + m)r < nm$. Only, all elements of matrixes must be those of non-negativity.[1]

B. Training Rules

For training procedure of NMF algorithm, W and H is iteratively updated using unsupervised learning rule until objective function(F) of formula(2) is converged to local minimum.[2]

$$F = \sum_{i=1}^{n} \sum_{u=1}^{m} [V_{iu} \log(WH)_{iu} - (WH)_{iu}] \tag{2}$$

Training Rule for W and H uses methods of minimizing Euclidian Distance and Kullback-Leibler Divergence below rules.

· **Training Rule 1**

- Minimize the Euclidian Distance $\|V - WH\|$

$$\begin{aligned} H_{au} &\leftarrow H_{au} \frac{(W^T V)_{au}}{(W^T WH)_{au}} \\ W_{ia} &\leftarrow W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \end{aligned} \tag{3}$$

· **Training Rule 2**
- Minimize Kullback-Leibler Divergence $D(V \| WH)$

$$H_{au} \leftarrow H_{au} \frac{\sum_u H_{au} V_{iu} / (WH)_{iu}}{\sum_k W_{ka}}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_u H_{au} V_{iu} / (WH)_{iu}}{\sum_v H_{av}} \quad (4)$$

The Euclidian Distance $\|V - WH\|$ and the Kullback-Leibler Divergence $D(V \| WH)$ are nonincreasing under the training rules and objective function, F, is always converged to local minimum.[2] In this paper, we trained by using Training Rule 2 that is to minimize Kullback-Leibler Divergence.

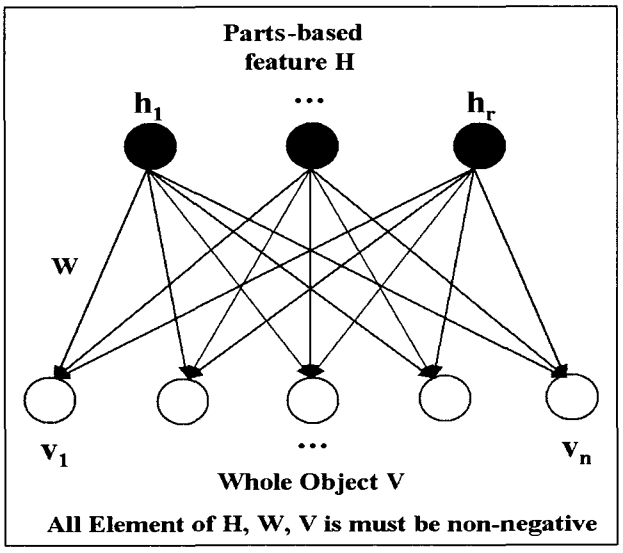

Fig. 1 Non-Negative Matrix Factorization

Fig. 1 diagram theory of NMF algorithm that represent characteristic of whole object V through selective weighted sum of parts-based feature H.[1] Weight matrix W that is trained with non-negativity means actual axis of the whole feature from feature space and this non-negativity lead to parts-based representations H because it allows only additive. Accordingly, dimensionally reduced matrix H means significant parts-based feature vectors about features of whole object and H and W is sparse matrix.

III. PROPOSED FEATURE EXTRACTION PROCEDURE

A. Proposed Idea

If speech spectrum is made by addition of each frequency component for resonance of vocal tract based on fundamental frequency originated from vocal cord, we should consider parts-based feature H as information about each position of vocal tract including a similar shape in same speech. The case of MFCC include overlapped information because it estimate whole characteristic of vocal tract, but in case of parts-based feature using NMF algorithm, we consider that characteristic overlap and intraspeaker variation of it are lower than those of MFCC.

In this paper, we applied mel filter bank output of speech spectrum to input of NMF algorithm and made characteristic robust under the limited speech train data using parameter of parts-based feature of vocal tract.

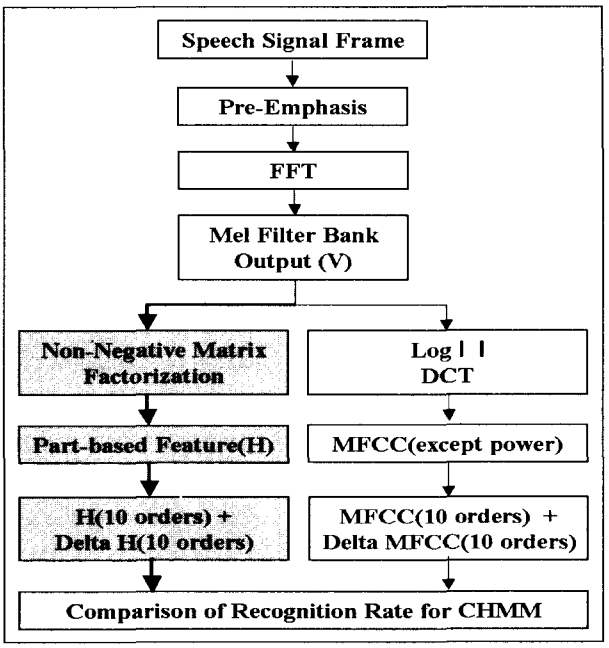

Fig. 2 Proposed Feature Extraction Procedure

Proposed feature extraction procedure of Fig. 2 is as follows.

1. Segmentation of input speech signal using frame analysis.
2. Implementation of pre-emphasis and Fast Fourier Transform for each speech frame.
3. Generation of outputs of 20 dimensions through mel filter bank analysis for speech spectrum
4. Application of mel filter bank output to input of NMF algorithm and study by training rule 2.
5. Generation of feature of 20 dimension added 1st delta component to 10 dimensional NMF output.

Only, MFCC of 10 dimension is generated by implementing Discrete Cosine Transform to the logarithm of mel filter bank output. Also, first delta components are added to MFCC.

IV. EXPERIMENT RESULT

In experiment speech DB is ETRI (Electronics and Telecommunications Research Institute) Samdori DB composed of 800 Korean digit speech data from 20 speakers.

For verification of recognition performance under the limited train speech data, we used 100 speech data from 10 speakers as train data (10 speech data per each digit speech). And we used test data that is 800 speech data from 20 speakers (80 speech data per each digit speech).

The analysis condition for feature extraction is as follows.

Table 1 Analysis Conditions

Speech Format	PCM RAW(16KHz, 16bit)
Frame Length	320 samples(20ms)
Frame Overlap Size	160 samples(10ms)
FFT Size	512(zero padding)
Mel Filter Bank Number	20
Feature Order(NMF/MFCC)	20/20(included 10 order Delta Component)

We implemented recognition experiment by CHMM (continuous hidden markov model) using two feature parameters that is proposed feature parameter (NMF feature) and MFCC. Experiment results for each feature parameter are showed in Table 2.

Table 2 Experiment Result

Feature	Recognition data	Recognition result
MFCC	Trained Person	99.25%
	Non-trained Person	93.25%
	Total	96.25%
Proposed Feature	Trained Person	99.72%
	Non-trained Person	95.25%
	Total	97.50%

In experiment result, proposed feature parameter is superior in recognition performance than MFCC in both Trained Person and Non-trained Person. This result shows 0.5% performance improvement in Trained Person and 2% performance improvement in Non-trained Person. Totally, performance improvement of proposed feature is 1.25 %.

In the results we considered that characteristic overlap and intraspeaker variation of proposed feature is lower than those of MFCC.

Fig. 3 and 4, scattering 1st and 2nd order coefficients for other digit speeches('3, sam' and '4, sa') of same speaker show well that characteristic overlap for other digit speeches of proposed feature(NMF) is lower than that of MFCC. Also, Fig. 5 and 6, scattering 1st and 2nd order coefficients for same digit speech('5, o') of other speakers show well that intraspeaker variation for same digit speech of proposed feature is lower than that of MFCC.

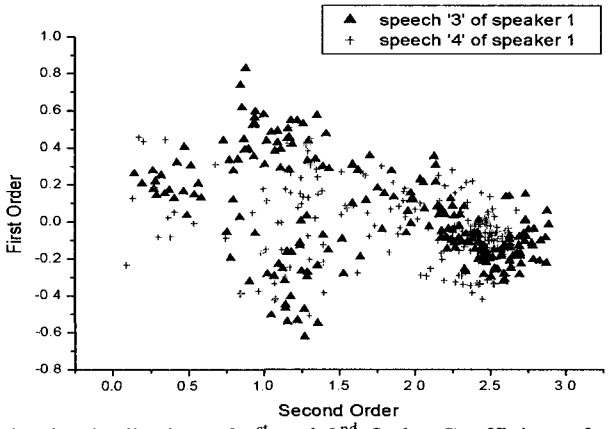

Fig. 3 Distribution of 1st and 2nd Order Coefficients for other digit speeches of same speaker (MFCC)

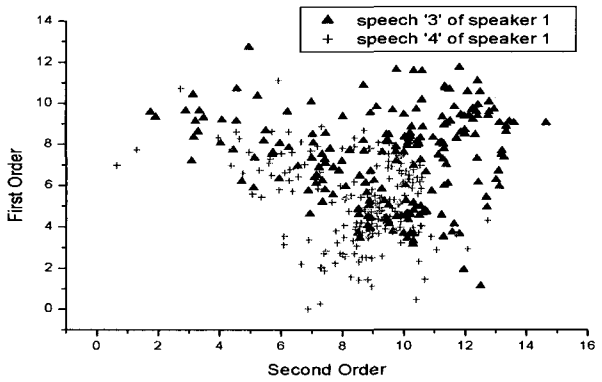

Fig. 4 Distribution of 1st and 2nd Order Coefficients for other digit speeches of same speaker (NMF)

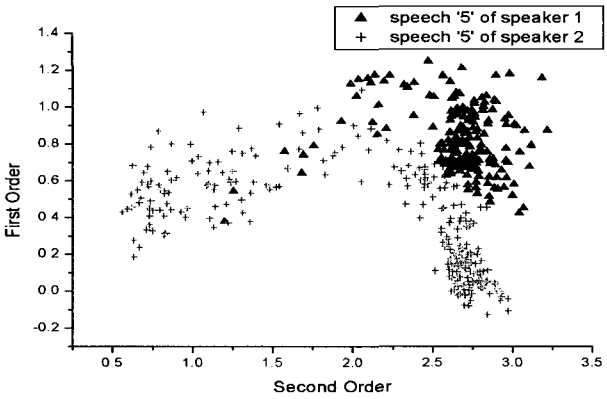

Fig. 5 Distribution of 1st and 2nd Order Coefficients for same digit speech of other speakers (MFCC)

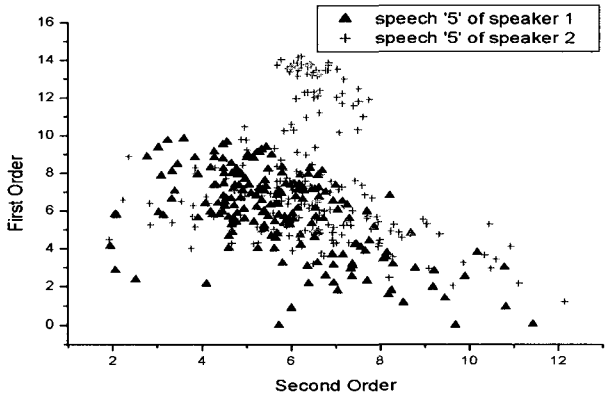

Fig. 6 Distribution of 1st and 2nd Order Coefficients for same digit speech of other speakers (NMF)

V. CONCLUSION

In this paper, we proposed new speech parameter through representation of characteristic of vocal tract using Non-Negative Matrix Factorization. In experiment result, we could verify that proposed feature is superior in recognition performance than MFCC that is used generally and high performance under the limited train speech data. For an upper fact, it showed that proposed feature extraction is sufficiently applicable to the continuous speech recognition system. But it still remained the problem how to reduce elapsed time for train process.

In the future, we are going to study data dependency through many-sided experiments using various speech databases. Also, we will verify effectiveness for proposed feature parameter with applying to various speech signal processing field and real-time continuous speech recognition system through modification and supplement about upper problem.

REFERENCES

[1] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature* vol. 401, Oct. 21, 1999, pp-788-791.

[2] Daniel D. Lee, H. Sebastian Seung, "Algorithms for Non-Negative Matrix Factorization", *in Advances in Neural Information Processing System 13, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds.*, 2001.

[3] H. Y. Choi, S. J. Choi, " Learning the Sparse Codes of Speeches via Non-Negative Matrix Factorization, *CVPR 2002.*

[4] Sven Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization", *IJCNN'03*, vol. 4, Oct. 14, 2003, pp. 2758-2763.

[5] Hoyer. P. O, "Non-Negative Sparse Coding", *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, 2002, pp. 557-565,

[6] S. Tsuge, M. Shishibori, S. Kurojwa, K. Kita, "Dimensionally Reduction Using Non-Negative Matrix Factorization for Information Retrieval", *Systems, Man, and Cybermetics, 2001 IEEE International Conference on*, vol. 2, 2001, pp. 960-965.

[7] D. Guillamet, B. Schiele, J. Vitria, "Analyzing non-negative matrix factorization for image classification", *Pattern Recognition, 2002. Proceedings. 16th international Conference on*, vol. 2, Aug. 2002, pp. 116-119.

[8] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals", Prentice Hall, 1978.

[9] L. R. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[10] Simon Haykin, "Neural Networks a Comprehensive Foundation", Prentice Hall, 1999.

[11] J. W. Park, P. W. Kim, C. K. Kim, K. I. Hur, "Adoption of Support Vector Machine and Independent Component Analysis for Implementation of Speech Recognizer", *Summer Conference of IEEK*, vol. 26, no.1, July, 2003, pp. 2164-2167.

Jeong-won Park
Received his BS degree in electronic engineering from Dong-A University in Busan, Korea, in 2002. he is currently MS course in electronic engineering at Dong-A University. His research interests include speech analysis, speech recognition, speaker recognition, neural network, kernel machine and hardware implementation.

Chang-Keun Kim
Received his BS degree in electronic engineering from Dong-A University in Busan, Korea, in 1994 and the MS degree in electronic engineering from Dong-A University in Busan, Korea, in 1998. he is currently Ph.D. course in electronic engineering at Dong-A University. His research interests include speech recognition, digital signal processing and hardware implementation.

Kwang-Seok Lee
Received B.S. and M.S. degrees of electronic engineering in 1983 and 1985 respectively, from Dong-A University. And Ph.D. from Dong-A University, in 1992. In 1995, He joined the Jinju National University in Jinju, Korea, where is currently an Professor and Industrial-University Cooperation Director. His research interests are in Intelligent System, DSP, Speech Recognition, Synthesis and Biometrics.

Si-Young Koh
Received the B.S and M.S degrees of electronic engineering in 1979 and 1983 respectively, from Young Nam University. And Ph.D. from Dong-A University, in 1992. In 1986 he joined the Kyungil University in Gyeongsan-si, Korea, where is currently an Professor. His research interests are in Bio Signal Processing, Speech Signal Processing.

Kang-In Hur
Received the B.S and M.S degrees of electronic engineering in 1980 and 1982 respectively, from Dong-A University. And Ph.D. from Kyung Hee University, in 1990. In 1984 he joined the Dong-A University in Busan, Korea, where is currently an Professor. His research interests are in DSP, Speech Recognition, Synthesis and Neural Networks.