Received: 11 May 2018 | Revised: 3 August 2018 | Accepted: 5 September 2018

DOI: 10.4218/etrij.2018-0189

ORIGINAL ARTICLE

WILEY ETRI Journal

Rank-weighted reconstruction feature for a robust deep neural network-based acoustic model

Hoon Chung | **Jeon Gue Park** | **Ho-Young Jung**

SW Contents Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea.

Correspondence
Hoon Chung, SW Contents Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea.
Email: hchung@etri.re.kr

Funding information
This work was supported by the ICT R&D program of MSIT/IITP [2015-0-00187, Core Technology Development of Spontaneous Speech Dialogue Processing for Language Learning]. This work was supported by an Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government [18ZS1100, Core Technology Research for Self-Improving Artificial Intelligence System].

In this paper, we propose a rank-weighted reconstruction feature to improve the robustness of a feed-forward deep neural network (FFDNN)-based acoustic model. In the FFDNN-based acoustic model, an input feature is constructed by vectorizing a submatrix that is created by slicing the feature vectors of frames within a context window. In this type of feature construction, the appropriate context window size is important because it determines the amount of trivial or discriminative information, such as redundancy, or temporal context of the input features. However, we ascertained whether a single parameter is sufficiently able to control the quantity of information. Therefore, we investigated the input feature construction from the perspectives of rank and nullity, and proposed a rank-weighted reconstruction feature herein, that allows for the retention of speech information components and the reduction in trivial components. The proposed method was evaluated in the TIMIT phone recognition and Wall Street Journal (WSJ) domains. The proposed method reduced the phone error rate of the TIMIT domain from 18.4% to 18.0%, and the word error rate of the WSJ domain from 4.70% to 4.43%.

KEYWORDS
deep neural network, rank limitation, speech recognition

1 | INTRODUCTION

Recently, the use of a feed-forward deep neural network (FFDNN)-based acoustic model has a significantly improved recognition accuracy in large vocabulary continuous speech recognition [1–3]. An FFDNN is a neural network composed of a set of affine transformations and nonlinear activation functions. The input feature for the FFDNN-based acoustic model is constructed by vectorizing a submatrix that is created by slicing the feature vectors of frames within a context window [2–5].

Various studies have been conducted to improve the performance of the FFDNN-based acoustic model. The objectives of these studies have been as follows: to optimize the network structure [6–8], develop loss functions [9] and optimization techniques [10], augment data [11], and optimize the hyperparameters.

This work focuses on input feature construction. In a pioneering work on an FFDNN-based acoustic model [2], the construction of an input feature was introduced by stacking the feature vectors of five preceding frames, one central frame, and five successor frames. In a previous study [4], the importance of the context window size was investigated empirically. This work showed that substantial gains of an FFDNN are attributable to input features concatenated from several consecutive speech frames within a relatively long context window. In the paper [5], it was also shown that frame-level metrics are further improved through the use of larger context windows. These works proved that the context window size should neither be too small nor be too large.

However, there has been little discussion regarding why the context window should be a certain size. Instead, there has been a tendency to expect black-boxed FFDNNs to be sufficiently representative of a feature's structure.

In this work, we propose a rank-weighted reconstruction method after investigating the input feature construction from the perspectives of rank and nullity. The proposed method factorizes the independent and null components of a sliced submatrix using singular value decomposition (SVD), and reconstructs the submatrix by weighting the null components to suppress trivial components and retain informative components.

The rest of this paper is organized as follows. In Section 2, we briefly describe statistical speech recognition using the FFDNN-based acoustic model. In Section 3, the input feature construction is reviewed. In Section 4, we present our approach in detail. Section 5 describes the experimental results, and Section 6 concludes the paper.

2 | STATISTICAL SPEECH RECOGNITION USING FFDNN-BASED ACOUSTIC MODEL

In this section, we briefly describe statistical speech recognition using the FFDNN-based acoustic model.

2.1 | Statistical Speech Recognition

Statistical speech recognition is a process that finds a word sequence $\mathbf{W}^*$ that maximizes the likelihood for a given input feature vector sequence $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{T}$, as follows:

$$\mathbf{W}^* = \mathrm{argmax}_{\mathbf{w}} \mathrm{max}_{\mathbf{s}} P(\mathbf{X}|\mathbf{S}) P(\mathbf{S}|\mathbf{W}) P(\mathbf{W}), \tag{1}$$

where $P(\mathbf{X}|\mathbf{S})$ is an acoustic model, $P(\mathbf{S}|\mathbf{W})$ is a pronunciation model, $P(\mathbf{W})$ is a language model, and $\mathrm{argmax}_{\mathbf{w}}$ is the best path search operation.

2.2 | FFDNN-based acoustic model

An FFDNN is a neural network that has more than one hidden nonlinear layer. For an input vector $\mathbf{x}_t$, each hidden layer transforms its input vector from the layer below to the layer above by applying an affine transform and nonlinear mapping, as follows:

$$\mathbf{z}^{(0)} = \mathbf{x}_t, \tag{2}$$

$$\mathbf{y}^{(l)} = \mathbf{W}^{(l)} \mathbf{z}^{(l-1)}, \tag{3}$$

$$\mathbf{z}^{(l)} = \sigma\left(\mathbf{y}^{(l)}\right), \tag{4}$$

where $\mathbf{W}^{(l)}$ is a weight matrix of the lth layer, and $\sigma(x)$ is a nonlinear activation function, such as sigmoid(x) or ReLU(x) [12]. In the last layer, softmax is used to obtain the probability of the ith class s_i for an input feature vector $\mathbf{x}_t$, as follows:

$$p(s_i|\mathbf{x}_t) = \mathrm{softmax}\left(\left\langle \mathbf{w}_i^{(L)}, \mathbf{z}^{(L-1)} \right\rangle\right), \tag{5}$$

where L is the number of hidden layers, and $\left\langle \mathbf{w}_i^{(L)}, \mathbf{z}^{(L-1)} \right\rangle$ are the inner product of the ith row vector of an output layer matrix and the output of the $(L-1)$th layer. To summarize, the FFDNN model parameter θ is defined using weight matrices, as follows:

$$\theta = \left\{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}\right\}. \tag{6}$$

3 | INPUT FEATURE CONSTRUCTION

In this section, we describe conventional input feature construction and discuss the issue of the context window size.

3.1 | Background

In speech recognition, a sequence of short-time speech frames $S = [s_1, s_2, \ldots, s_t, \ldots, s_T]$ are assumed to be a realization of the corresponding phoneme sequence $P = [p_1, p_2, \ldots, p_n, \ldots, p_N]$, where s_t is the t-th speech frame and p_n is the n-th phoneme. In general, the phoneme sequence length N is much smaller than that of speech frame sequence T. Therefore, there is average of T/N redundant speech frames, or there is high correlation among the adjacent speech frames. In addition, short-time speech signal s_t is assumed to be contaminated by background noise n_t for a clean speech frame c_t, as $s_t = c_t + n_t$. There is low cross-correlation between speech components and noise components, and the power of the noise signal is generally lower than that of the speech signal.

Therefore, to improve the performance of speech recognition, reducing the influence of background noise by reducing uncorrelated low-power components and containing adjacent phoneme information by minimizing temporal dependency in the feature extraction step should be considered. There are various approaches to remove the influence of background noise. Some of the most representative approaches are speech reconstruction using SVD [13–15] or a Wiener filter [16–18]. However, there is little research on considering the temporal dependency among adjacent speech frames.

3.2 | Conventional input feature construction

In the FFDNN-based acoustic model, for an input speech signal s_t, studies have been conducted on the use of raw

waveform signals as the direct input vector [19–21]. However, Mel frequency cepstral coefficients, perceptual linear prediction, or a Mel filter bank are more commonly used feature representations of a speech frame. For a sequence of these types of feature vectors, an input feature for the FFDNN-based acoustic model is constructed by stacking features within a context window. For example, assuming that a matrix of $\mathbf{O} \in \mathbb{R}^{n \times T}$, representing the feature vectors of an utterance, is given as follows:

$$\mathbf{O} = \begin{pmatrix} 0.1 & 0.1 & 0.3 & \cdots & 0.2 \\ 0.1 & 0.1 & 0.3 & \cdots & 0.2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & 0.2 & \ldots & 0.2 \end{pmatrix} \tag{7}$$

where o_t is the feature of the t-th speech frame, n is the dimension of o_t, and T is the number of frames in an utterance; in addition, a submatrix of $\mathbf{O} \in \mathbb{R}^{n \times (2c+1)}$ is constructed by slicing the left and right c column vectors from the t-th column, as follows:

$$\mathbf{O}_t = [o_{t-c} \ldots o_t \ldots o_{t+c}] \tag{8}$$

The input feature x_t is then constructed using the vectorization operation, as follows:

$$x_t = \mathrm{vec}(\mathbf{O}_t), \tag{9}$$

where vec $(\cdot)$ is a column stacking operator.

3.3 | Context window size issue

Conventional input feature construction is straightforward. The only aspect is to set the context window size as an appropriate value that is neither too small nor too large, allowing the FFDNN-based acoustic model to achieve the best performance that it possibly can. As described previously, a constructed input feature x_t is expected to be composed of high-power uncorrelated components, which are related to current and adjacent phoneme information without low-power background noise components. However, due to the difference between the phoneme rate and speech frame rate, it is difficult to guarantee that a fixed size context window contains the same amount of phoneme information at each time. Upon review of the submatrix construction, it can be noted that the context window size affects the dimension of the submatrix $\mathbf{O}_t$, where the dimension of the submatrix is the sum of the rank and nullity, as follows:

$$\dim(\mathbf{O}_t) = \mathrm{rank}(\mathbf{O}_t) + \mathrm{nul}(O_t). \tag{10}$$

In other words, it can be assumed that the rank and nullity of the submatrix may vary depending on the context window size, wherein the rank and nullity represent the number of informative and trivial components, respectively. However, the problem is that it is not known how many number of rank or nullity increase actually. As the context window changes, the rank or nullity may increase.

4 | RANK-WEIGHTED RECONSTRUCTION FEATURE

The primary objective of this work is to factorize the submatrix $\mathbf{O}_t$ with two components, as follows:

$$\mathbf{O}_t = \mathbf{R}_t + \mathbf{N}_t, \tag{11}$$

where $\mathbf{R}_t$ denotes the informative components, and $\mathbf{N}_t$ denotes trivial components, and the submatrix $\mathbf{O}_t$ is then reconstructed by controlling the contribution of trivial components, as follows:

$$\mathbf{O}_t \approx \mathbf{R}_t + \gamma \mathbf{N}_t, \tag{12}$$

where γ controls the contribution of the relatively trivial components. In this framework, problems exist in terms of how to define the more informative components and how to factorize $\mathbf{R}_t$ and $\mathbf{N}_t$ from $\mathbf{O}_t$ based on this criterion.

4.1 | Single value decomposition

In this work, we use SVD-based matrix factorization [22]. A given matrix $\mathbf{U}_t \in \mathbb{R}^{n \times (2c+1)}$ can be decomposed into three matrices, $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$, as follows:

$$\mathbf{O}_t = \mathbf{U}\mathbf{\Sigma}V^{\mathrm{T}}, \tag{13}$$

where the left and right singular matrices $\mathbf{U} \in \mathbb{R}^{n \times (2c+1)}$ and $\mathbf{V} \in \mathbb{R}^{n \times (2c+1)}$ are orthogonal, and the matrix $\mathbf{\Sigma} \in \mathbb{R}^{(2c+1) \times (2c+1)}$ has a diagonal form, as follows:

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{2c+1}). \tag{14}$$

The diagonal elements $\{\sigma_i\}$ of $\mathbf{\Sigma}$ are the singular values of $\mathbf{O}_t$, and are ordered such that

$$\sigma_1 \geq \sigma_2 \geq \cdots, \sigma_{2c+1}. \tag{15}$$

A submatrix is then reconstructed, as follows:

$$\mathbf{O}_t = \sum_{i=1}^{2c+1} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}. \tag{16}$$

It is generally understood that smaller singular values and their corresponding singular vectors should not significantly contribute to the matrix, and thus, the original matrix should be accurately reconstructed by ignoring the smaller singular values along with the singular vectors $\mathbf{U}$ and $\mathbf{V}$.

4.2 | Rank-weight matrix reconstruction

In this work, we consider a matrix reconstructed by top-k singular vectors to be $\mathbf{R}_t$, and a matrix reconstructed by

lower $2c + 1 - k$ singular vectors to be $\mathbf{N}_t$. In the proposed approach, a submatrix is therefore represented as the sum of two matrices, as follows:

$$\bar{\mathbf{O}}_t \approx \underbrace{\sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}}_{\mathbf{R}_t} + \underbrace{\gamma \sum_{i=k+1}^{2c+1} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}}_{\mathbf{N}_t} \tag{17}$$

where the two variables k and γ are introduced to control the amount of information more precisely. If γ is setas 1.0, it is the same as the baseline system, and if γ is set as 0.0, it is the same as a truncated SVD. For a reconstructed submatrix, the input feature at time t, $\mathbf{x}_t$ is constructed through vectorization as follows:

$$\mathbf{x}_t = \mathrm{vec}\left(\bar{\mathbf{O}}_t\right). \tag{18}$$

Table 1 summarizes the parameters needed to construct the input features for both the conventional and proposed approaches. In the conventional approach, the context window size c is the only controllable parameter. Other parameters, such as rank k and weight γ, are fixed at $2c + 1$ and 1.0, respectively. However, all of these parameters are controllable in the proposed approach.

5 | EXPERIMENTS AND RESULTS

The effectiveness of the proposed approach was evaluated for the TIMIT phone recognition and WSJ domains using the Kaldi toolkit [23].

5.1 | TIMIT domain

This training set contains 3696 sentences uttered by 462 speakers. Five percent of the training data is randomly selected as part of the validation set. The 24-speaker standard test set is used for evaluation. Each frame was represented by 40-dimensional speaker adapted feature space maximum likelihood linear regression (fMLLR) features. We used the Kaldi TIMITs5 recipe to build the phoneme recognizer. The recognizer includes a bigram phoneme language model that was created from the training set. The 61 phonemes were mapped into 48 phonemes for training, and 39 phonemes for testing.

TABLE 1 Parameters used for input vector construction

Parameter	Conventional	Proposed
Context window	c	c
Rank	$2c + 1$	k
Weight	1.0	γ

5.2 | WSJ domain

The WSJ corpus consists of read speech and text data from the Wall Street Journal. The speech data were recorded from many speakers reading subsets of the text data. We used the si284 training set with 81 h of speech data.

We also used Dev93 as the dev set and Eval92 as the test set [24]. The WSJ speech recognition task converts speech audio into a sequence of words. We used the Kaldi WSJ s5 recipe (eg, "local/online/run_nnet2 baseline.sh") to build the speech recognizer. The acoustic model was an FFDNNHMM hybrid. Each frame was represented using 13-dimensional fMLLR features.

5.3 | Experimental results

In this work, we conducted experiments by varying the weight parameter γ from 1.0 to 0.0 in intervals of 0.1, and the rank parameter k from 1 to $2c + 1$ for each context window size c.

Tables 2 and 3 show the experimental results for the TIMIT domain, and Tables 4 and 5 show the experimental results for the WSJ domain. It shows the best performance and parameter settings for different context window sizes. For the TIMIT domain, at up to two context window sizes, the baseline system and the proposed method achieved the same best performance levels by using input features that had been reconstructed with full ranks. However, for larger context window sizes, the proposed method reduced the WER more than that of the baseline by using input features that had been reconstructed using the top-k rank instead. As shown in Table 2, for the TIMIT development set, the baseline system achieved the lowest WER by 17.3%, with a context window size of 3 and a corresponding rank of 7, and it achieved WER by 19.1% for the test set with the same context window size. The proposed method further reduced the WER to 18.6% by reconstructing top-6 ranks, and weighting other features by 0.2 for the same context window size of 3.

TABLE 2 WER (%) of baseline on different context window sizes for TIMIT domain

c	Dev (%)	Test (%)
0	19.8	21.6
1	18.4	19.7
2	17.7	19.2
3	17.3	19.1
4	17.5	18.5
5	17.6	18.4
6	17.6	18.9

TABLE 3 WER (%) of proposed method on different context window sizes for TIMIT domain

c	k	γ	Dev (%)	Test (%)
0	1	1.0	19.8	21.6
1	3	1.0	18.4	19.7
2	5	1.0	17.7	19.2
3	6	0.2	17.4	18.6
4	7	0.1	17.6	18.0
5	8	0.1	17.5	18.0
6	9	0.0	17.7	18.4

TABLE 4 WER (%) of baseline on different context window sizes for WSJ domain

c	Dev (%)	Test (%)
0	25.10	25.82
1	9.61	6.43
2	8.37	5.12
3	7.86	4.84
4	7.93	4.74
5	7.52	4.75
6	7.65	4.87
7	7.74	4.70

TABLE 5 WER (%) of proposed method on different context window sizes for WSJ domain

c	k	γ	Dev (%)	Test (%)
0	1	1.0	25.10	25.82
1	3	1.0	9.61	6.43
2	5	1.0	8.37	5.12
3	7	1.0	7.86	4.84
4	7	0.2	7.80	4.68
5	10	0.1	7.43	4.50
6	12	0.2	7.64	4.43
7	12	0.0	7.60	4.54

For the TIMIT domain, it should be noted that state-of-the-art systems, including the Kaldi toolkit, report the baseline performance using a five context window size from the aspect of the test set, even if the best context window size is 3 [24,25]. Therefore, from the aspect of the best performance for the test set, the baseline systems achieved the lowest WER by 18.4% with a context window size of 5 and a corresponding rank of 11. However, the proposed method further reduced the WER by 18.0% using reconstruction features with top-8 ranks, and other features weighted by 0.1 for the same context window size of 5.

For the case of the WSJ domain, for up to three context window sizes, both the baseline system and the proposed method achieved the same best performance levels using input features that had been reconstructed with full ranks. However, for context window sizes >3, the proposed rank-weighted reconstruction features showed a better performance. For example, the baseline achieved the lowest WER by 7.52% for the development set and 4.75% for the test set using a context window size of 5. However, the proposed approach reduced the WER to 7.43% for the development set and 4.50% for the test set. For the case considering the test set, the baseline achieved the lowest WER by 4.70% when it used a context window size of 7, with a full rank of 15. However, the proposed approach further reduced WER to 4.54% for the same context window size of 7, with a lower rank of 12. In addition, the proposed approach achieved the lowest WER by 4.43%, with a context window size of 6 and a rank of 12.

6 | CONCLUSIONS

In this paper, we proposed the rank-weighted reconstruction method after investigating an input feature construction from the perspective of rank and nullity. The proposed method factorizes independent and null components of the sliced submatrix using SVD, and reconstructs a submatrix by weighting the null components to suppress trivial components and retain informative components. When compared to the conventional approach, the proposed method provides a more sophisticated strategy for constructing the input features by introducing two controllable parameters, such as the rank and weighting factor. For the TIMIT and WSJ domains, the proposed approach reduced the WERs from 18.4% to 18.0%, and from 4.70% to 4.43%, respectively, by using reconstructed features with weighted low rank components. In conclusion, the proposed input feature construction method introduces two additional parameters outside of the context window size to control the rank of a submatrix and the contribution of trivial components. In addition, the method shows that a high dimension combined with a low rank improves the performance of the FFDNN-based acoustic model.

ACKNOWLEDGMENTS

This work was supported by the ICT R&D program of MSIT/IITP [2015-0-00187, Core Technology Development of Spontaneous Speech Dialogue Processing for Language Learning]. This work was supported by an Electronics and

Telecommunications Research Institute (ETRI) grant funded by the Korean Government [18ZS1100, Core Technology Research for Self-Improving Artificial Intelligence System].

ORCID

Hoon Chung https://orcid.org/0000-0003-2551-1352

REFERENCES

1. G. Hinton et al., *Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups*, Signal Process Mag. **29** (2012), no. 6, 82–97.
2. G. E. Dahl et al., *Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition*, IEEE Trans. Audio Speech Language Process. **20** (2012), no. 1, 30–42.
3. L. Deng et al., Recent advances in deep learning for speech research at Microsoft, in *IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 26–31, 2013, pp. 8604–8608.
4. J. Pan et al., Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: why DNN surpasses GMMS in acoustic modeling, in *IEEE Int. Symp. Chinese Spoken Language Process (ISCSLP)*, Kowloon, China, Dec. 2012, pp. 301–305.
5. A. L. Maas et al., *Building DNN acoustic models for large vocabulary speech recognition*, Comput. Speech Lang. **41** (2017), pp. 195–213.
6. T. N. Sainath et al., Deep convolutional neural networks for LVCSR, in *IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 8614–8618.
7. H. Sak, A. Senior, and F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in *Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, Sept. 14–18, 2014, pp. 338–342.
8. T. N. Sainath et al., Convolutional, long short-term memory, fully connected deep neural networks, in *IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 19–24, 2015, pp. 4580–4584.
9. Y. Shinohara, Adversarial multi-task learning of deep neural networks for robust speech recognition, in *INTERSPEECH*, San Francisco, CA, USA, Sept. 8–12, 2016, pp. 2369–2372.
10. D. Povey, X. Zhang, and S. Khudanpur, *Parallel training of deep neural networks with natural gradient and parameter averaging*, arXiv preprint, 2014.
11. X. Cui, V. Goel, and B. Kingsbury, *Data augmentation for deep neural network acoustic modeling*, IEEE/ACM Trans. Audio Speech Language Process. **23** (2015), no. 9, 1469–1477.
12. V. Nair, and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *Proc. Int. Conf. Mach. Learn. (ICML-10)*, Haifa, Israel, June 21–24, 2010, pp. 807–814.
13. K. Hermus, and P. Wambacq, *A review of signal subspace speech enhancement and its application to noise robust speech recognition*, EURASIP J. Appl. Signal Process. **2007** (2007), 1–15.
14. K. Hermus et al., Fully adaptive SVD-based noise removal for robust speech recognition, in *Eur. Conf. Speech Commun. Technol.*, Budapest, Hungary, Sept. 5–9, 1999, pp. 1–4.
15. T. Schanze, *Compression and noise reduction of biomedical signals by singular value decomposition*, IFAC-PapersOnLine **51** (2018), no. 2, 361–366.
16. S. Chirtmay, and M. Tahernezhadi, *Speech enhancement using wiener filtering*, Acoustics lett. **21**, (1997), 110–115.
17. J. Chen et al., *New insights into the noise reduction wiener filter*, IEEE Trans. Audio Speech Language Process. **14** (2006), no. 4, 1218–1234.
18. S. Lee et al., *Statistical model-based noise reduction approach for car interior applications to speech recognition*, ETRI J. **32** (2010), no. 5, 801–809.
19. D. Palaz et al., Analysis of CNN-based speech recognition system using raw speech as input, in *INTERSPEECH*, Dresden, Germany, Sept. 6–10, 2015, pp. 11–15.
20. P. Golik et al., Convolutional neural networks for acoustic modeling of raw time signal in LVCSR, in *INTERSPEECH*, Dresden, Germany, Sept. 6–10, 2015, pp. 26–30.
21. T. N. Sainath et al., Learning the speech front-end with raw waveform CLDNNs, in *INTERSPEECH*, Dresden, Germany, Sept. 6–10, 2015, pp. 1–5.
22. G. H. Golub, C. Reinsch, *Singular value decomposition and least squares solutions*, Numerische Mathematik **14** (1970), no. 5, 403–420.
23. D. Povey et al., The Kaldi speech recognition toolkit, in *IEEE Workshop Automatic Speech Recogn. Understanding*, Waikoloa, HI, USA, Dec. 11–15, 2011, no. EPFL-CONF192584.
24. D. B. Paul, J. M. Baker, The design for the wall street journal-based CSR corpus, in *Proc. Workshop Speech Natural Language*, Harriman, NY, USA, Feb. 23–26, 1992, pp. 357–362.
25. C. Lopes, F. Perdigao, Phoneme recognition on the TIMIT database, in *Speech Technologies*, InTech, 2011.

AUTHOR BIOGRAPHIES

Hoon Chung received his BS, MS, and PhD degrees in Electronics Engineering from Kangwon National University, Chuncheon, Korea, in 1994, 1996, and 2007 respectively. He joined the Electronics and Telecommunication Research Institute (ETRI) in 2004, and is currently a research member at the Automatic Speech Translation and Artificial Intelligence Research Center. His current research interests include fast decoding, robust speech recognition, and large vocabulary speech recognition systems.

Jeon Gue Park received his PhD degree in information and communication engineering from Paichai University, Daejeon, Rep. of Korea, in 2010. He has worked for ETRI, Daejeon, Rep. of Korea as a senior researcher since 1991, Lernout and Hauspie Korea, Seoul, Rep. of Korea as a director and division head

since 2000, and Donga Seetech Inc. Seoul, Rep. of Korea as a director and CTO since 2002. He rejoined ETRI since 2004 and is currently leading a spoken dialog processing research project. His current research interests include artificial intelligence, computer-assisted language learning, spoken dialog system, and cognitive systems.

Ho-Young Jung received his BS degree in electronics engineering from Kyungpook National University, Daegu, Korea, in 1993, and his MS and PhD degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1995 and 1999, respectively. His PhD dissertation was on robust speech recognition. He joined the Electronics and Telecommunications Research Institute, Daejeon, Korea, in 1999, as a senior researcher in the field of speech/language systems.