PAPER *Special Issue on Special Section on Robust Speech Processing in Realistic Environments*

Canonicalization of Feature Parameters for Robust Speech Recognition Based on Distinctive Phonetic Feature (DPF) Vectors

Mohammad NURUL HUDA[†a)], **Muhammad GHULAM**[†], **Takashi FUKUDA**[††], **Kouichi KATSURADA**[†], ***Nonmembers***, ***and*** **Tsuneo NITTA**[†], ***Member***

SUMMARY Acoustic models of an hidden Markov model (HMM)-based classifier include various types of hidden factors such as speaker-specific characteristics, coarticulation, and an acoustic environment, etc. If there exists a canonicalization process that recovers the degraded margin of acoustic likelihoods between correct phonemes and other ones caused by hidden factors, the robustness of an automatic speech recognition (ASR) system can be enhanced. In this paper, we propose a novel method for canonicalizing feature parameters to realize robust ASR. The proposed canonicalization method is composed of multiple distinctive phonetic feature (DPF) extractors corresponding to the hidden factors, and a DPF selector which selects an optimum DPF as a canonicalized DPF. A noise reduction procedure based on Wiener filter is also introduced to the canonicalization process to eliminate noise factor. In the experiment on Japanese version AURORA2 database (AURORA-2J), the proposed method achieves 60.87% and 35.48% relative improvement in comparison with a standard ASR system with mel frequency cepstral coefficient (MFCC) parameters under clean and multicondition training , respectively. Moreover, the proposed method provides the high recognition rate with a reduced, approximately one-third in our experiment, mixture-set of HMMs and less memory requirements.

key words: *automatic speech recognition, feature extraction, canonicalization, distinctive phonetic feature, hidden factor*

1. Introduction

A major drawback on current automatic speech recognition (ASR) systems is their lack of robustness in practical conditions. Many approaches have been proposed to aim at a robust ASR system, however, the ASR system with enough performance in a practical environment has not been realized. One of the reasons is that acoustic models (AMs) of an Hidden Markov Model (HMM)-based classifier include many unspecific hidden factors such as, speaker-specific characteristics with a gender type and speaking style, an acoustic environment with ambient noise and/or channel characteristics, etc. To overcome this difficulty, an approach of parallel decoding with multiple HMMs corresponding to the hidden factors has been proposed as illustrated in Fig. 1 [1], [2]. Multi-path acoustic modeling was also proposed to represent the hidden factors with several paths in the same AM instead of applying multiple HMMs [3], [4]. Those methods have the possibility to achieve robust ASR systems in various conditions, however, they need a large amount of memory and computation time. If there exists a canonicalization process of feature parameters that reduces the influence of the hidden factors on HMM-based classifiers, the robust ASR system can be realized at low cost.

This paper proposes the canonicalization of feature parameters. Here, the canonicalization of feature parameters is defined as a process to recover the degraded margins of acoustic likelihoods between correct phonemes and other ones resulted from the above mentioned hidden factors. Fig. 2 shows an overview of the canonicalization process. The canonicalization is realized by introducing multiple distinctive phonetic feature (DPF) extractors [5] corresponding to hidden

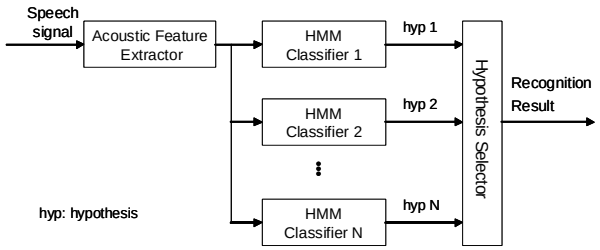


Fig. 1 A single feature extractor and multiple HMM classifiers.

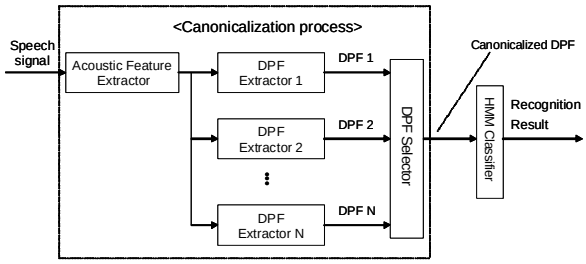


Fig. 2 Multiple DPF extractors and a single HMM classifier.

Manuscript received June 30, 2007.
Manuscript revised June 30, 2007.
Final manuscript received June 30, 2007.

[†]The authors are with the Graduate School of Engineering, Toyohashi University of Technology, Toyohashi, 441-8580, Japan

[††]Presently, the author is with Tokyo Research Laboratory, IBM Japan Ltd.

a) E-mail: huda@vox.tutkie.tut.ac.jp

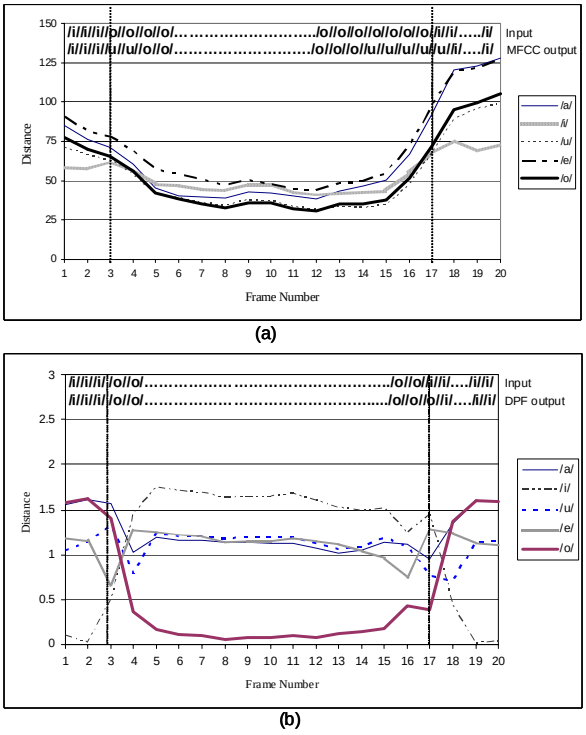


Fig. 3 Time vs. phoneme distance for input vectors of utterance /ioi/ with a) MFCC based system, and b) DPF using MLN.

factors between an acoustic feature extractor and an HMM classifier. Figure 3 shows time vs. phoneme distance of five Japanese vowels for utterance /ioi/ to illustrate the importance of DPF extractor. In an mel frequency cepstral coefficient (MFCC) based traditional system, the distance of a phoneme from a given input vector of 38 dimensions (12 MFCC,12 Δt, 12 $\Delta\Delta t$, ΔP, $\Delta\Delta P$) is calculated by using Maharanobis distance. In the MFCC based system of Fig. 3(a), a single mixture Guassian mixture model (GMM) for each vowel is used for the experiment. In distinctive phonetic feature (DPF) using multilayer neural network (MLN) of Fig. 3(b), the distance of a phoneme from a given vector of 15 dimensions that was generated by MLN is calculated by Euclidean distance. Each phoneme in the MFCC based system is closer to each other, hence noise from the environment can easily misclassify phoneme. On the other hand, each phoneme in DPF using MLN is far from each other, and so noise can not affect much. Besides, MFCC based system generates more misclassification of phoneme at boundary between two phonemes (See Fig. 3(a)), while DPF using MLN can detect phoneme boundary with less number of errors. Therefore, the concept of DPF extractor in the canonicalization process makes the ASR more robust in realistic environments.

In this paper, canonicalization process is applied to target gender factor as well as noise and speaker specific characteristics. Gender factor is resolved by using male, female and gender-independent DPF extractors, while speaker variability are determined by the DPF extractor designed by a clustering technique. On the other hand, the likelihood degradation caused by some hidden factors such as an ambient noise factor should be recovered by not only DPF extractors, but also the other appropriate mechanism according to the target factor. Therefore, a noise reduction technique based on a two-stage Wiener filtering process proposed by European Telecommunication Standards Institute (ETSI) [6] is embedded to the canonicalization process for the target factor of ambient noise. Experiments are conducted to evaluate recognition accuracy using speech data with the hidden factors of a gender type and ambient noise.

This paper is organized as follows. Section 2 outlines the implementation of a DPF extractor, and Sect. 3 explains the canonicalization process targeting gender factor. Section 4 extends the canonicalization process to deal with both hidden factors of gender type and ambient noise, and describes experimental result on an AURORA-2J [7] task. Finally, Sect. 5 finishes with some conclusions.

2. Overview of DPF Extractor

The configuration of the DPF extractor is illustrated in Fig. 4. At an acoustic feature extraction stage, firstly, input speech is converted into local features (LFs) that represent a variation in spectrum along time and frequency axes [8]. LFs are then entered into an MLN with four layers, including two hidden layers, after combining a current frame x_t with the other two frames that are 3-points before and after the current frame (x_{t-3}, x_{t+3}). Our previous work showed that LF is superior to MFCC for the input to MLNs [5]. The MLN has 45 output units (15x3) corresponding to a set of triphones, or context-dependent DPF vector that consists of three DPF vectors (a preceding context DPF, a current DPF, and a following context DPF) with 15 dimensions each. The two hidden layers consist of 256 and 96 units from the input layer. Fifteen DPF elements {mora, <high, low, an intermediate expression of high and low>, <anterior, back, an intermediate expression of anterior and back>, coronal, plosive, affricative, continuent, voiced, unvoiced, nasal, semi-vowel} are used to represent the balance of phoneme configuration in a DPF space. Each phoneme can be classified by a unique set of above mentioned DPF elements with binary value. The MLN is trained by using a standard back-propagation algorithm.

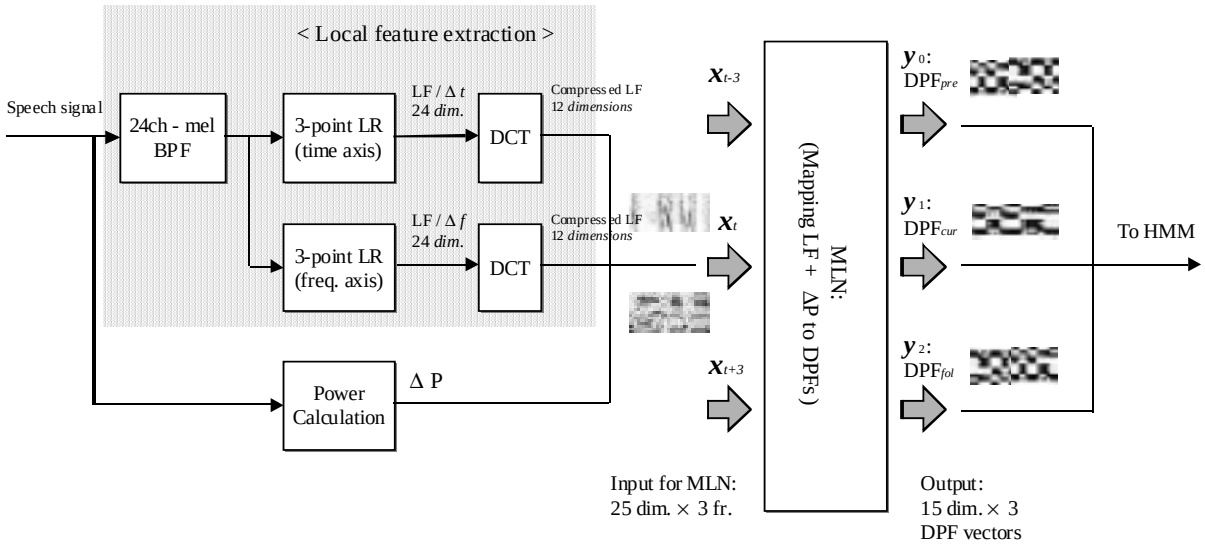


Fig. 4 DPF extractor.

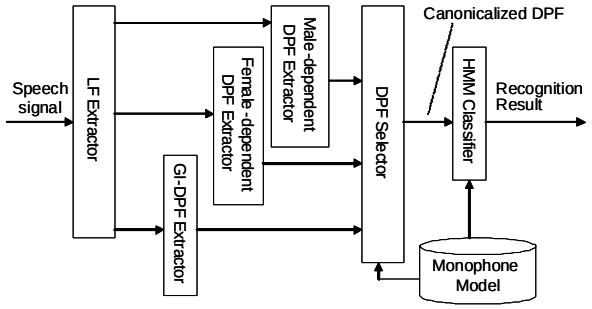


Fig. 5 Block diagram of the canonicalization process focused on gender factor.

3. Canonicalization process targeting gender factor

3.1 Configuration of canonicalization process

This section describes the configuration of the canonicalization process focused on the gender factor [9]. Figure 5 shows the canonicalization process. Male- and female-dependent DPF extractors are firstly used to map LFs onto two DPF spaces corresponding to the gender type. Two DPF vectors extracted by each DPF extractor are called DPF_{male} and DPF_{female}, respectively. These DPF extractors are trained individually with a male speech and a female speech data set. In addition, a gender-independent (GI) DPF extractor is used to compensate errors of a DPF selector as described in Sect. 3.3. GI-DPF extractor is trained with both the male and the female speech data set. The male- and female-dependent DPF extractor are intended to generate DPF vector with less distortion for the input speech signal uttered by the same-sex speaker. Therefore, if the desired type of DPF vector corresponding to the speaker's gender can be selected correctly, the influence of the gender factor is expected to be reduced. In this section, the desired type of DPF vector corresponding to the gender type is defined as the canonicalized DPF vector.

3.2 DPF selector

There are several strategies such as a distortion measure and an information criterion to decide the canonicalized DPF. In this paper, we propose a method based on a distance between DPF vectors and AMs. In the proposed method, firstly, DPF-based AMs are trained as follows.

(A) Extracts DPF_{male} vectors from male speech after designing the DPF extractor with the male speech data set.
(B) Extracts DPF_{female} vectors from female speech after designing DPF extractor with the female speech data set.
(C) Trains the DPF-based AMs using the DPF vector sets extracted in the above mentioned process.

Because the DPF-based AM represents the distribution better for the canonicalized DPF vectors than for the mismatched DPF vectors, such as DPF_{male} and DPF_{female} vectors obtained from female speech and male speech, respectively. It is expected that the canonicalized DPF vector is assigned after comparing the distance between two DPF vectors (DPF_{male} and DPF_{female}) and AMs. Here, Maharanobis distance is used as a distance measure between DPF vector and AMs. Further details are described below.

The DPF selector firstly accumulates minimal distances after comparing the Maharanobis distance D_M between a candidate DPF vector of DPF_{male} and AMs of all phonemes as follows.

$$d = \frac{1}{T_C} \sum_{t=1}^{T} \min_i D_M \Big\{ \mathbf{X}(t), \theta_i \Big\} \quad (1)$$

$$D_M \Big\{ \mathbf{X}(t), \theta_i \Big\} = \Big[log\{\mathbf{X}(t)\} - \mu_i \Big]^T \mathbf{\Sigma}_i^{-1} \Big[log\{\mathbf{X}(t)\} - \mu_i \Big] \quad (2)$$

Where $\mathbf{X}(t)$ is the input DPF_{male} of the t-th frame, θ_i is the distribution parameter set in the middle state of i-th HMM, and D_M is Maharanobis distance between $\mathbf{X}$(t) and θ_i. T and T_c are the total number of frames in an utterance and that in a vowel interval, respectively. μ and Σ are mean vectors and covariance matrices in the model, respectively. Here, logarithmic computation is involved in Eq.(2) because DPF distribution is approximated with lognormal distribution in HMM, that is, DPF distributions are represented using lognormal distributions instead of standard normal distributions in HMM [10]. The DPF selector then replaces the accumulated minimal distance obtained by Eq.(1) and (2) with a discriminant score d for assigning the canonicalized DPF vector. The discriminant score d is calculated between the HMM with a single mixture and the input DPF vectors only within vowel intervals, because the DPF elements in vowel intervals have higher reliability than that in consonantal parts. The interval in which the value of a DPF element of "mora" is bigger than 0.5 is recognized as a vowel interval. Next, the discriminant score d concerning another candidate DPF vector of DPF_{female} is also calculated using the same procedure as DPF_{male}. Then, the DPF selector decides about the canonicalized DPF vector by comparing the discriminant scores of DPF_{male} and DPF_{female} at the end of the utterance. DPF-based AM is retrained using the canonicalized DPF vector. These iterative training, or redesigning DPF-based AM using canonicalized DPF vector, is continued until the value of the HMM likelihood converges. At a recognition stage, the canonicalized DPF vector assigned by the DPF selector is used as an input to the HMM classifier.

3.3 Application of gender-independent DPF extractor

In the DPF selector, selection errors might be occurred even if in a noiseless condition. Thus, a neutral type of DPF extractor, or a GI-DPF extractor, is introduced to the canonicalization process in which intermediate vocal characteristics between male and female speech are represented and contribute to compensate the selection error. When the discriminant score d of DPF_{male} is close to that of DPF_{female}, for example, the DPF selector can assign the third candidate DPF vector of DPF_{GI} as the canonicalized DPF vector. In the following experiments, DPF_{GI} is applied when the discriminant score difference between the two is within γ%. The value of γ was set to 25.

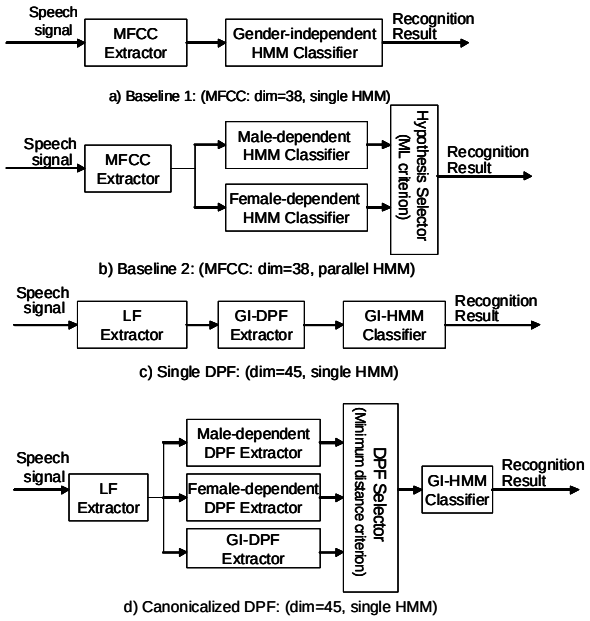


Fig. 6 ASR systems evaluated in experiments.

3.4 Experiments

3.4.1 Speech Database

The following two data sets were used.

(1) Training data set (D1)

A subset of ASJ (Acoustic Society of Japan) Continuous Speech Database consisting of 9003 sentences uttered by 30 male and 30 female speakers, respectively (16 kHz, 16-bit) [11]. This set is composed of clean speech.

(2) Test data set (D2)

A subset of Tohoku University and Matsushita Spoken Word Database consisting of 100 words uttered by 10 unknown male and female speakers each [12]. This set is also composed of clean speech. The sampling rate is converted from 24 kHz to 16 kHz.

3.4.2 Experimental setup

An input speech is sampled at 16 kHz and a 512-point Fast Fourier Transform (FFT) of the 25 ms Hamming-windowed speech segment is applied every 10 ms. The resultant FFT power spectrum is then integrated into 24-channel band pass filters (BPFs) output with mel-scaled center frequencies. At the acoustic feature ex-

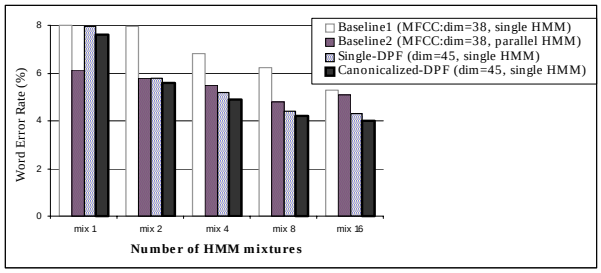


Fig. 7 Performance comparison among ASR systems.

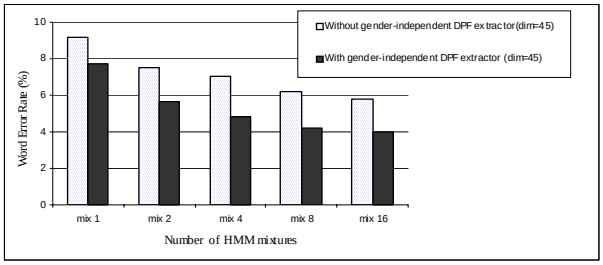


Fig. 8 Difference of canonicalization units.

traction stage, an output of BPF bank is converted into LF, and then LF is mapped to DPF_{male}, DPF_{female}, and DPF_{GI}, each. After assigning the canonicalized DPF vector from three candidate DPFs in the DPF selector, the canonicalized DPF vector is input into the HMM classifier. In the experiment, 43 Japanese monophone HMMs with three states are used. Output probabilities are represented in normal distribution for a baseline system based on an MFCC parameter set and lognormal distribution with only negative skewness for DPF vectors [10].

Speaker-independent isolated spoken-word recognition tests were carried out with the D2 data set. In the experiments, four types of ASR systems illustrated in Fig. 6 were investigated. Baseline 1 in Fig. 6(a) shows a standard ASR system in which an MFCC-based feature parameter set with dynamic features (Δt, $\Delta\Delta t$), ΔP and $\Delta\Delta P$ after conducting Cepstrum Mean Normalization (CMN) is input into a single HMM classifier, or GI-HMM classifier. The single HMM classifier is trained with D1 data set. Parallel HMMs of baseline 2 in Fig. 6(b) are followed by the hypothesis selector that recognizes an HMM with the target gender factor based on a maximum likelihood (ML) criterion. The male speech and female speech in D1 data set is used to train each HMM individually in the parallel HMM. A single DPF in Fig. 6(c) indicates a method of using a single DPF extractor which is designed by both male and female speech in D1 data set. Fig. 6(d) shows the proposed canonicalization process followed by a single HMM classifier.

3.4.3 Experimental results and discussion

Figure 7 shows the experimental results. The canonicalized DPF based system outperformed both the baseline systems except the case of a single mixture. The DPF vectors calculated using the single DPF extractor in Fig. 6(c) also yielded better performance than the baseline systems at four mixtures of HMM and above. This gain in performance is considered to be as follows. The DPF vectors are originally specified to classify phonemes even if the hidden factor of gender is included and, as a result, the canonicalized DPF vectors can form a feature vector that is independent of the targeting hidden factor of the gender type.

The proposed canonicalization process has three types of DPF extractors including two DPF extractors for male and female voice, and one DPF extractor for GI voice. Here, an effect of adding the GI-DPF extractor is investigated. Figure 8 shows the recognition result with and without using the GI-DPF extractor in the canonicalization process. The canonicalization process with the GI-DPF extractor is superior to that without it. This fact suggests that the GI-DPF extractor minimizes the selection error of the DPF selector.

4. Extension of the canonicalization process

4.1 Configuration of improved canonicalization process

In this section, the canonicalization process is extended to deal with not only the gender factor but also the other hidden factors [13].

Figure 9 shows a block diagram of the extended canonicalization process. In this process, an input speech is firstly denoised with a two-stage Wiener filter implemented in the ETSI standard advanced distributed speech recognition (DSR) front-end ES202, named WI008 [6]. The noise reduction process used in the WI008 front end is very powerful, however, gains in ASR performance depend on speakers [7]. Therefore, a succeeding feature extraction stage is expected to enhance such speech by reducing the effect of speaker variability. In Sect. 3, enhancement of features was realized in the form of canonicalization for gender factor. In this section, we propose a method to canonicalize a variability between speakers by designing multiple DPF extractors with more precise target than gender type. In the extended canonicalization process, secondly, the denoised speech is processed at the BPF bank and the LF extractor, and then passed to N number of DPF extractors. Each DPF extractor is designed by using a clustering algorithm as described in the next section. The output $\{\mathbf{X}_1(t), \mathbf{X}_2(t), \cdots, \mathbf{X}_N(t)\}$ of the DPF extractors representing speaker-specific characteristics are finally input to the DPF selector which assigns the

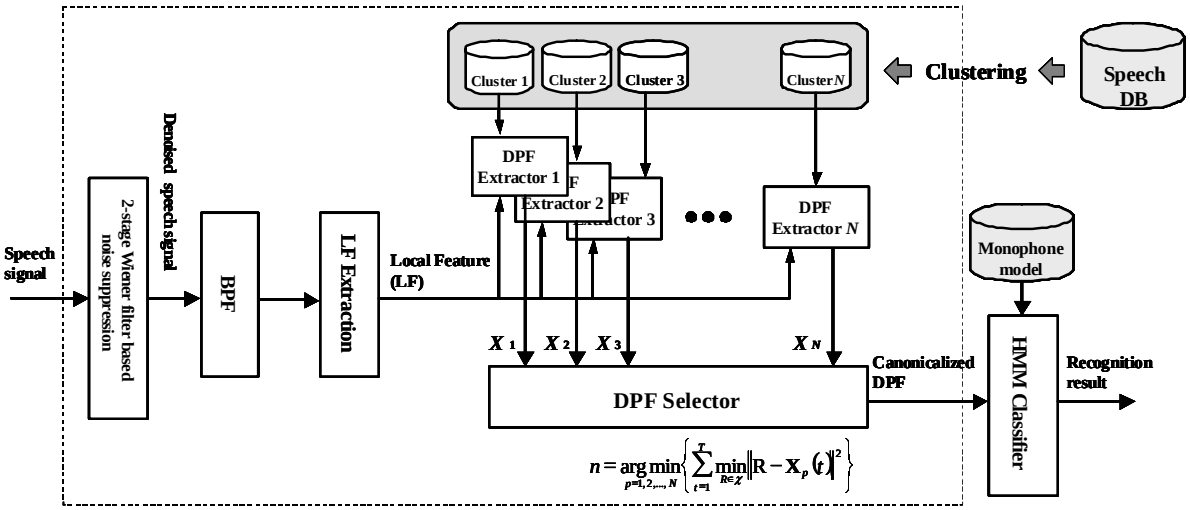


Fig. 9 Block diagram of the extended canonicalization process.

desired canonicalized DPF vector.

At the same time, the DPF selector is also modified to isolate its design from HMM classifier design because the isolation is important idea for DSR. The DPF selector design described in the previous section was coupled with HMM classifier design. In the modified DPF selector, the vector of n-th DPF extractor which satisfies the following equation is regarded as the canonicalized DPF.

$$n = \underset{p=1,2,\cdots,N}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T} \min_{\mathbf{R}\in\chi} \|\mathbf{R} - \mathbf{X}_p(t)\|^2 \right\} \tag{3}$$

Here, χ , $\mathbf{R}$, and T are a set of DPF vectors, a DPF vector of a phoneme, and the number of frames in an utterance , respectively. The number of phonemes is 38. In the case of phoneme /a/, for example, R is set to (1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0). The Eq.(3) accumulates minimal distances along a speech interval after comparing the distances between a DPF vector $\mathbf{X}_p(t)$ and DPF vectors of all the phoneme ($\mathbf{R}$). The n-th DPF vector that has minimal distance among extractors is considered as the canonicalized DPF vector.

4.2 DPF extractor design based on clustering technique

In this section, a clustering procedure to divide spoken sentences into groups that are expected to represent some hidden factors is introduced. The basic concept of the clustering is based on an algorithm derived by [14]. The spoken sentence data which contains all the vowels of /a, i, u, e, o/ and nasal-sound /m, n, N/ is used for clustering. The reason for including nasal sounds is that the individual difference of speech often appears in those sounds which are affected by many cavities of vocal organs. After clustering, a cluster to which each spoken sentence belongs is decided. The clustering procedure based on a modified K-means algorithm [15] is done as follows.

Step1: Calculate mean vectors on cepstrum space along each interval of all the same phoneme patterns in each spoken sentence data. After that, mean vectors of vowels and at least one nasal sound in /m, n, N/ are used for clustering.

Step2: Select two initial centroids that have maximum distance. The maximum distance is calculated by comparing distances between 6-8 phoneme pairs of two spoken sentence data.

Step3: Decide a cluster to which each spoken sentence data belongs by comparing the distances of 6-8 phoneme pairs.

Step4: Calculate mean vector of each phoneme using all the spoken sentences in each cluster and then replaces the centroids by them.

Step5: Repeat (3)- (4) until category of spoken sentence data is fixed.

Step6: Repeat (1)-(5) until the number of clusters reaches to a preset value N_c. When dividing, the cluster with maximum variance is selected by normalizing accumulated distance of each cluster and comparing the normalized distances of all the clusters.

4.3 Experiments

4.3.1 Speech Database

The extended canonicalization process was evaluated using the AURORA-2J database [7]. The sampling rate is 8 kHz, and the utterances are composed of connected Japanese digit strings. The database contains clean data as well as various types of noise-corrupted data. Different types of noise, for example, subway,

Table 1 Word accuracy [%] of the WI008 front end as baseline (mixture=20)[Clean Training].

Clean Training (%Acc)														
	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.43	98.37	98.57	98.40	98.44	98.43	98.37	98.57	98.40	98.44	98.62	98.64	98.63	98.48
20 dB	97.61	98.43	98.69	97.72	98.11	96.04	96.43	97.94	98.03	97.11	97.64	96.67	97.16	97.52
15 dB	94.17	96.07	97.94	95.80	96.00	92.17	94.23	96.00	96.17	94.64	93.80	93.65	93.73	95.00
10 dB	86.18	89.18	95.17	90.34	90.22	81.46	88.27	89.02	89.36	87.03	85.69	86.91	86.30	88.16
5 dB	66.53	69.20	83.21	73.19	73.03	59.90	72.46	73.22	77.78	70.84	62.57	68.83	65.70	70.69
0 dB	36.97	31.92	48.05	37.70	38.66	23.43	44.01	40.98	49.24	39.42	32.70	40.24	36.47	38.52
-5 dB	9.06	-2.15	12.65	6.26	6.46	-6.91	13.42	8.56	11.97	6.76	9.40	16.17	12.79	7.84
Average	76.29	76.96	84.61	78.95	79.20	70.60	79.08	79.43	82.12	77.81	74.48	77.26	75.87	77.98

Table 2 Word accuracy [%] of the proposed canonicalized DPF (No. of cluster: N=4; mixture=8) [Clean Training].

Clean Training (%Acc)														
	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.29	99.37	99.49	99.44	99.40	99.26	99.37	99.49	99.44	99.39	99.51	99.43	99.47	99.41
20 dB	98.31	97.57	99.01	97.93	98.21	90.97	97.91	96.11	98.29	95.82	98.59	97.91	98.25	97.26
15 dB	95.79	96.12	98.50	96.04	96.61	88.97	95.77	95.87	96.03	94.16	95.85	96.22	96.04	95.52
10 dB	89.21	90.15	96.27	91.99	91.91	83.17	89.03	90.01	89.91	88.03	87.53	90.08	88.81	89.74
5 dB	67.52	69.89	77.50	75.24	72.54	66.74	70.80	71.34	79.20	72.02	64.25	69.74	67.00	71.22
0 dB	38.97	36.76	50.27	39.85	41.46	37.53	40.11	41.52	51.82	42.75	33.75	39.03	36.39	40.96
-5 dB	15.38	15.05	15.33	14.93	15.17	15.79	16.84	18.61	16.69	16.98	18.34	17.20	17.77	16.42
Average	77.96	78.10	84.31	80.21	80.14	73.48	78.72	78.97	83.05	78.56	75.99	78.60	77.30	78.94

Table 3 Overall average word accuracy[%] of the proposed method for different number of clusters.

	N (Number of clusters)			
	1	2	4	8
%Acc	75.92	77.71	78.94	78.04

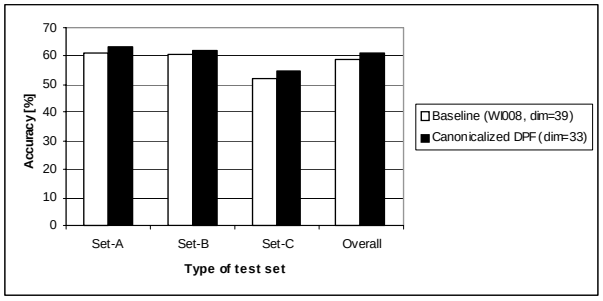


Fig. 10 Relative performances of the methods comparing to the WI007 in clean training.

babble, car, exhibition, restaurant, street, airport, station noise are added/convolved in them. In the baseline system [7], there are thirteen recognition units: eleven-digits HMMs with sixteen states and twenty Gaussian mixtures; one silence HMM with three states, and one short-pause HMM with one state. Thirty-six Gaussian mixtures are used for silence and short pause. For the experiment in this section, the training of the HMM classifier was performed using clean and multicondition data. For the multicondition training dataset, four types of noise (Subway, Babble, Car, Exhibition) are added to the clean speech at five values of signal-to-noise ratio (SNR) [SNR=clean, 20 dB, 15 dB, 10 dB, 5 dB].

4.3.2 Experimental setup

Twenty-five ms Hamming-windowed input speech segments are applied at every 10 ms. In the WI008 front end, 12 Mel-cepstral parameters together with log power and their delta, and delta-delta parameters (so, a total of 39 dimensions) are used as feature vector. In the proposed method, the canonicalized DPF vectors, selected by the DPF selector, is orthogonalized by using Kurhonen Loeve Transform (KLT) and Gram-Schmidt orthogonalization procedure [5]. The resultant orthogonalized DPF vectors have 33 dimensions and show normal distribution. In this section, normal distributions are used as an output probability representation. In the experiment, the number of clusters N is varied from 1 to 8. Eleven-digit HMMs with sixteen states and eight Gaussian mixtures are prepared as back-end system together with a silent model with three states and a pause model with one state.

4.3.3 Experimental results

The experimental results of the WI008 and the proposed canonicalized DPF are shown in Table 1 and 2, respectively, for clean training. In the experiment, the number of cluster N was fixed to four. From Tables 1 and 2, we can observe that the proposed method performed better than the WI008 in clean and almost all noisy conditions. For example, in clean condition,

Table 4 Word accuracy [%] of the WI008 front end as baseline (mixture=20)[Multicondition Training].

Multicondition Training (%Acc)														
	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.79	99.73	99.79	99.75	99.77	99.79	99.73	99.79	99.75	99.77	99.88	99.70	99.79	99.77
20 dB	99.63	99.58	99.76	99.38	99.59	99.32	99.43	99.31	99.38	99.36	99.60	99.58	99.59	99.50
15 dB	99.26	99.40	99.61	99.01	99.32	98.56	98.94	98.39	98.40	98.57	99.39	99.06	99.23	99.00
10 dB	98.28	98.31	98.30	97.13	98.01	94.53	96.58	94.90	95.25	95.32	97.97	96.95	97.46	96.82
5 dB	94.14	92.14	94.72	92.01	93.25	82.16	88.85	86.22	88.77	86.50	91.80	88.63	90.22	89.94
0 dB	78.94	68.77	80.14	76.09	75.99	52.13	69.07	67.97	71.34	65.13	70.13	63.18	66.66	69.78
-5 dB	44.15	25.67	43.13	43.20	39.04	8.35	33.74	28.30	37.67	27.02	33.80	29.29	31.55	32.73
Average	94.05	91.64	94.51	92.72	93.23	85.34	90.57	89.36	90.63	88.98	91.78	89.48	90.63	91.01

Table 5 Word accuracy [%] of the proposed canonicalized DPF(No. of cluster: N=4; mixture=8) [Multicondition Training].

Multicondition Training (%Acc)														
	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.24	99.47	99.14	99.56	99.35	99.24	98.49	99.14	99.16	99.01	99.01	99.62	99.32	99.21
20 dB	99.23	98.97	99.79	98.97	99.24	97.75	99.22	99.12	98.50	98.65	98.87	98.46	98.67	98.89
15 dB	99.02	99.08	99.14	98.43	98.92	97.35	99.15	98.14	97.42	98.02	98.83	97.44	98.14	98.40
10 dB	98.44	97.91	98.09	97.03	97.87	90.63	96.87	94.56	95.33	94.35	97.67	95.84	96.76	96.24
5 dB	94.29	90.51	96.71	94.22	93.93	89.75	88.95	86.42	84.90	87.51	92.83	88.44	90.64	90.70
0 dB	77.73	68.54	75.75	76.23	74.56	58.63	70.98	68.01	69.51	66.78	72.66	65.84	69.25	70.39
-5 dB	43.58	34.92	36.18	44.22	39.73	17.57	33.43	32.11	36.21	29.83	33.89	33.57	33.73	34.57
Average	93.74	91.00	93.90	92.98	92.90	86.82	91.03	89.25	89.13	89.06	92.17	89.20	90.69	90.92

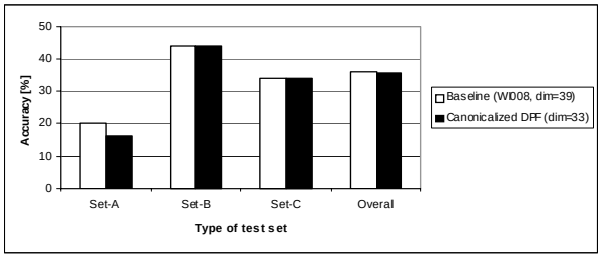


Fig. 11 Relative performances of the methods comparing to the WI007 in multicondition training.

the proposed method achieved 99.41% word accuracy, while the WI008 had 98.48%. In noisy condition with SNR=0 dB, the proposed method yielded word accuracy of 40.96%, while the WI008 had 38.52%. Figure 10 shows relative performance of the WI008 and the proposed method comparing to WI007, which is a previous version of the WI008 front end and does not include Wiener filter based noise reduction procedure, in clean training. The proposed canonicalized DPF achieved 60.87% relative improvement over WI007, while the WI008 gained it 59.09%. Between three data sets of A, B, and C, the C-set is open data concerning channel characteristics. The proposed method achieved the most gain at the C-set (54.86%) compared to the WI008 (51.84%). These results indicate that the WI008 is not effective for channel distortion.

Table 3 shows the performance of the proposed method for all different number of clusters. With four clusters, it showed the optimal performance. From the clustering point of view, we also checked the contents of resultant clusters for designing DPF extractors. It showed that each cluster was composed of almost same gender speakers and about 10% opposite gender speakers.

Table 4 and 5 show multicondition training results for the WI008 and the proposed method, respectively. Average accuracies using the proposed method for set-B and set-C are 89.06% and 90.69%, respectively that are slightly better than the corresponding accuracies of 88.98% and 90.63% generated by the WI008. Besides, overall average accuracies using the proposed method for 0 dB and -5 dB signals are 70.39% and 34.57%, respectively that are superior to the corresponding accuracies 69.78% and 32.73% generated by the baseline WI008 system.

Relative performances of the baseline WI008 and the proposed method in multicondition training are given in Fig. 11. The proposed canonicalized DPF method achieved 35.48% relative improvement over WI007, while the WI008 gained it 36.08%. However, the proposed method achieved this comparable gain with one-third Gaussian mixtures needed for the WI008. The WI008 had overall accuracy of 91.01% using twenty Gaussian mixtures of HMM, while the proposed method showed 90.92% accuracy using only eight Gaussian mixtures.

It is claimed that the proposed method requires less computation than the other existing methods because higher recognition accuracy is achieved at lower number of mixture of HMM. The WI008 requires additional number of mixtures to get approximately equal recognition rate. It may be mentioned that additional mixture(s) requires more computation cost. These findings justify the use of canonicalized DPF for a low com-

putational ASR system.

We can also find that the proposed improved canonicalization needs less memory than the other existing methods that deal with hidden factors. One of the existing methods, parallel decoding with multiple HMMs needs more storage than the proposed improved method. It also needs more computation to select optimum hypothesis from a large amount of data, while the proposed improved method used a single HMM classifier which needs less storage.

5. Conclusion

In this paper we proposed the canonicalization method of acoustic features based on the DPF vectors, and also addressed the design procedure of multiple DPF extractors by using the clustering technique. The concept of canonicalization is important especially for robust DSR. In the experiment on the AURORA-2J task, the effectiveness of the canonicalization was justified. We can summarize the findings as follows:

(a) In the experiment on the isolated spoken word recognition task, the proposed method could reduce the influence of gender factor as hidden factors.

(b) The proposed method achieved significant improvements on the AURORA-2J task when combining the canonicalization process with the noise reduction technique based on two-stage Wiener filtering.

(c) Compared to the ETSI-based front end system, it showed superior recognition accuracy in clean training, but comparable recognition accuracy at multicondition training. However, the proposed method required only one-third Gaussian mixtures comparing to that of ETSI-based system.

It is known that recurrent neural network (RNN) can resolve one of the hidden factors, co-articulation, better than MLN. We would like to implement the DPF extractors of the proposed method by using RNN for its capability of handing a longer context window, in future.

References

[1] S. Matsuda, T. Jitsuhiro, K. Markov and S. Nakamura, "Speech Recognition System Robust to Noise and Speaking Styles," Proc. ICSLP 04, Vol. IV, pp. 2817-2820, 2004.

[2] T. Shinozaki and S. Furui, "Spontaneous Speech Recognition Using a Massively Parallel Decoder," Proc. ICSLP 04, Vol.III, pp. 2817-2820, 2004.

[3] A. Lee, Y. Mera, K. Shikano and H. Saruwatari, "Selective multi-path acoustic model based on database likelihoods," Proc. ICSLP02, pp.2661-2664, 2002.

[4] M. Ida and S. Nakamura, "Rapid Environment Adaptation Method Based on HMM Composition with Prior Noise GMM and Multi-SNR Models for Noisy Speech Recognition," The Institute of Electronics, Information and Communication Engineers (IEICE) Transaction on Information and Systems. , Vol. J86-D-II, No.2, pp. 195-203, 2003.

[5] T. Fukuda and T. Nitta, "Orthogonalized Distinctive Phonetic feature Extraction for Noise-Robust Automatic Speech Recognition," The Institute of Electronics, Information and Communication Engineers (IEICE) Transactions on Information and Systems, Vol. E87-D, No.5, pp. 1110-1118, 2004.

[6] ETSI ES 202 050 v1.1.1, "Distributed speech recognition; advanced front end feature extraction algorithm; compression algorithm," 2004.

[7] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, et al., "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition," IEICE Transaction on Information and Systems, Vol.E88-D, No.3 pp.535-544, 2005.

[8] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. ICASSP99, pp.421-424, 1999.

[9] T. Fukuda and T. Nitta, "Canonicalization of Feature Parameters for Automatic Speech Recognition," Proc. ICSLP04, Vol. IV, pp.2537-2540, 2004.

[10] T. Fukuda and T. Nitta, "Noise-robust ASR by Using Distinctive Phonetic Features Approximated with Logarithomic Normal Distribution of HMM," Proc. Eurospeech 03, Vol.III, pp.2185-2188, 2003.

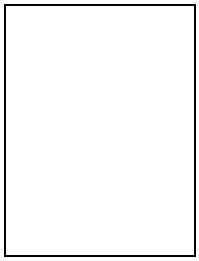
[11] T. Kobayashi, S. Itahashi, S. Hayamizu and T. Takezawa, "ASJ Continuous Speech Corpus for Research," Acoustic Society of Japan Trans. Vol.48, No.12, pp.888-893, 1992.

[12] S. Makino, N. Niyada, Y. Mafune and K. Kido, "Tohoku University and Matsushita isolated spoken word database," Acoustic Society of Japan (ASJ) Trans. Vol.48, No.12, pp.899-905, 1992.

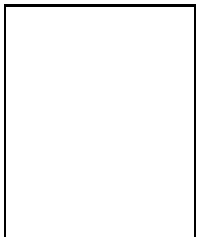
[13] T. Fukuda and T. Nitta, "Designing Multiple Distinctive Phonetic Feature Extractors for Canonicalization by Using Clustering Technique," Proc. Eurospeech05, pp.3141-3144, 2005.

[14] M. Samejima, A. Lee, H. Saruwatari, and K. Shikano, "Evaluation of Acoustic Models and Adaptation Methods Based on Collection of Spontaneous Speech for Child Speech recognition," IPSJ SIG Technical Reports, 2004-SLP-54, pp.199-204, 2004.

[15] J. G. Wilpon and L. R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," IEEE ASSP-33, 3, pp.587-594, 1985.

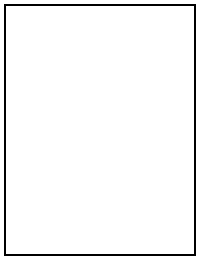


Mohammad Nurul Huda was born in 1973. received his B. Sc and M. Sc. in Computer Science and Engineering degree from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh in 1997 and 2004, respectively. Now he is a Ph.D. student of Toyohashi University of Technology, Japan. His research field includes Automatic Speech Recognition. He is a student member of the Acoustic Society of Japan (ASJ) and the International Speech Communication Association (ISCA).

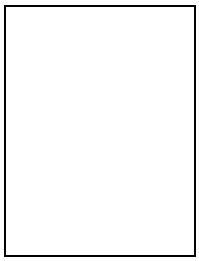


Muhammad Ghulam received his Bachelor degree in Computer Science and

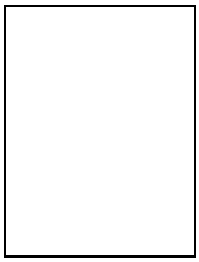
Engineering in 1997 from Bangladesh University of Engineering and Technology, and M.E. and Ph.D. degree in 2003 and 2006, respectively, from Toyohashi University of Technology, Japan. He is currently a researcher in the same university. His research interest includes automatic speech recognition and human-computer interface. He is a member of the Acoustic Society of Japan (ASJ), the IEEE, and the ISCA.



Takashi Fukuda received his Ph. D. degree in the Department of Electronics and Information Engineering in 2005 from Toyohashi University of Technology, Japan. He is currently engaged as researcher at IBM research, Tokyo Research Laboratory. His Research field includes automatic speech recognition. He is a member of ASJ, IPSJ, and ISCA.



Kouichi Katsurada received his Ph.D. degree from Osaka University in 2000. He has been a research associate at the department of knowledge-based information engineering, Toyohashi University of Technology since 2000. His research interests include multimodal interaction, knowledge processing and semantic web. He is a member of AAAI, IPSJ, JSAI, ASJ, NLP and HIS.



Tsuneo Nitta was born in 1946. He received his B.E.E. degree in 1969 and his Dr. Eng. Degree in 1988, both from Tohoku University, Japan. After engaging in research and development at the R&D Center of Toshiba Corporation and Multimedia Engineering Laboratory, where he was a chief Research Scientist, since 1998 he has been a Professor at the Graduate School of Engineering, Toyohashi University of Technology. His current research interests include speech recognition, multi-modal interaction, and acquisition of language and concepts. He received the Best Paper Award from the Institute of Electronics, Information and communications Engineers (IEICE), Japan, in 1988. He is a member of the Information Processing Society of Japan (IPSJ), the Acoustic Society of Japan (ASJ), the Japanese Society for Artificial Intelligence and the IEEE.