

Noise-robust Automatic Speech Recognition Using Orthogonalized Distinctive Phonetic Feature Vectors

Takashi Fukuda and Tsuneo Nitta

Graduate School of Engineering, Toyohashi University of Technology, Japan
fukuda@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

Abstract

With the aim of using an automatic speech recognition (ASR) system in practical environments, various approaches focused on noise-robustness such as noise adaptation and reduction techniques have been investigated. We have previously proposed a distinctive phonetic feature (DPF) parameter set for a noise-robust ASR system, which reduced the effect of high-level additive noise[1]. This paper describes an attempt to apply an orthogonalized DPF parameter set as an input of HMMs. In our proposed method, orthogonal bases are calculated using conventional DPF vectors that represent 38 Japanese phonemes, then the Karhunen-Loeve transform (KLT) is used to orthogonalize the DPFs, output from a multi-layer neural network (MLN), by using the orthogonal bases. In experiments, orthogonalized DPF parameters were firstly compared with original DPF parameters on an isolated spoken-word recognition task with clean speech. Noise robustness was then tested with four types of additive noise. The proposed orthogonalized DPFs can reduce the error rate in an isolated spoken-word recognition task both with clean speech and with speech contaminated by additive noise. Furthermore, we achieved significant improvements over a baseline system with MFCC and dynamic feature-set when combining the orthogonalized DPFs with conventional static MFCCs and ΔP .

1. Introduction

A current automatic speech recognition (ASR) system based on MFCC parameters can achieve high performance when speech signals, uttered clearly in noiseless environments, are input into the ASR. Recognition accuracy is, by contrast, degraded in practical environments by deformations of the log-spectrum envelope caused by various noises. As one of the approaches for a noise robust ASR, recently, the use of distinctive phonetic features (DPFs) has been again receiving attention[2, 3, 4]*. In [2], a set of lower-level multi-layer neural networks (MLNs) corresponding to five groups was used to map acoustic features into the DPFs. Each MLN was trained to extract a corresponding DPF in the group. The DPFs output from lower-level MLNs were input to a higher-level MLN which produced the acoustic likelihood of subword units. This work improved the recognition accuracy of spontaneous speech in addition to speech contaminated by additive noises. A different work[3] also used a set of MLNs corresponding to each BPF channel to extract the DPFs. The

output DPFs were then used in a higher-level MLN, the same as in [2].

In our previous work[1], we proposed a method to extract context-dependent DPFs by a single MLN, and applied them to a noise-robust ASR based on a standard HMM with diagonals in covariance matrices. In the DPF extraction stage, after converting a speech signal to acoustic features composed of local features (LFs) and ΔP , an MLN with 33 output units corresponding to context-dependent DPFs of 11 DPFs, 11 preceding context DPFs, and 11 following context DPFs maps the LFs to DPFs. The output DPFs could reduce the word error rate on an isolated spoken-word recognition task with additive noise.

The components of a DPF vector output from the MLN, however, correlate one another. Because the HMM with diagonals in covariance matrices does not consider the correlation between the components, it requires feature components with noncorrelation as an input. In this paper, we firstly describe a method to orthogonalize the DPFs output from the MLN. Orthogonalized DPF parameters are then evaluated in comparison with original DPF parameters and standard parameters of MFCCs and dynamic features in experiments using a clean speech database, and experiments are carried out to evaluate the accuracy when adding various noise to clean speech. A combined feature parameter set of orthogonalized DPFs and MFCCs is also applied.

This paper is organized as follows. Section 2 outlines the implementation of an orthogonal DPF extractor, then section 3 describes the experimental setup and results, and provides a discussion, and section 4 finishes with some conclusions.

2. Overview of an orthogonalized DPF extractor

The DPF extractor[1] is illustrated in Figure 1. At the acoustic feature extraction stage, firstly, an input speech is converted into LFs. They 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 N -points before and after the current frame (x_{t-N} , x_{t+N}). The MLN has 33 output units (11×3) corresponding to context-dependent DPFs that consist of 11 DPFs, 11 preceding context DPFs, and 11 following context DPFs. The hidden layer consists of 256 and 64 units from the input layer with 75 units. Eleven DPF elements of high, low, front, back, coronal, plosive, continuant, fricative, nasal, voiced and semi-vowel are used. The MLN is trained using a back-propagation algorithm to output the value of 1 for the corresponding DPF elements with an input phoneme and its adjacent phonemes. The number of training data of each tri-phoneme is limited to a maximum of 30 and the data is selected using nearest neighborhood clustering.

*Linguists have proposed DPFs that separates each phoneme by representing the manner of articulation and tongue position, etc[5]. The use of DPFs had been investigated previously in speech recognition[6, 7].

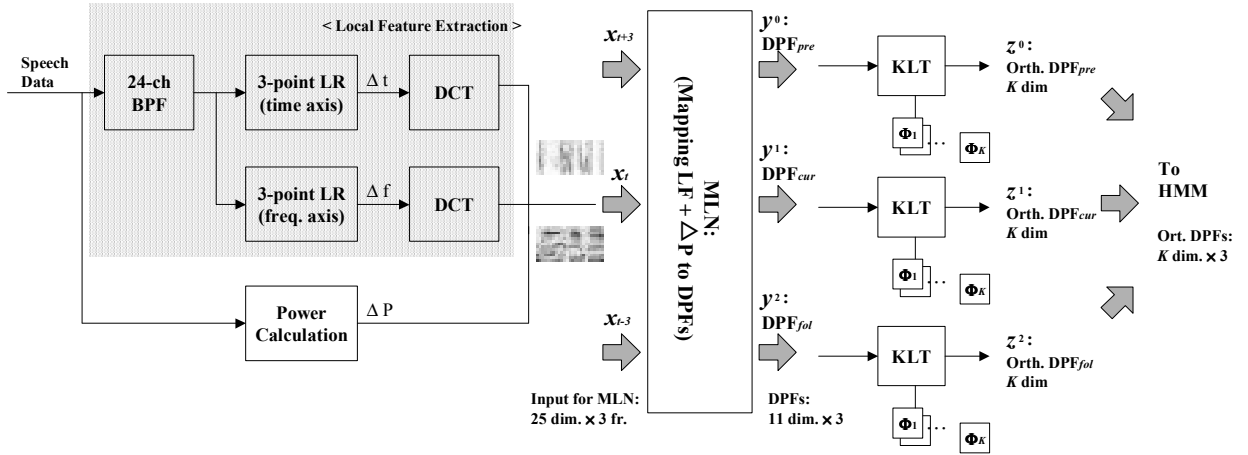


Fig. 1 Orthogonalized DPF feature extraction

The output DPFs are orthogonalized by Karhunen-Loeve transformation (KLT) as follows.

$$z_k^m = \mathbf{y}^m \cdot \Phi_k \quad (m = 0, 1, 2) \quad (1)$$

where \mathbf{y}^m is DPF vector before orthogonalization, z_k^m ($k = 1, 2, \dots, K$) is k -th dimensional orthogonalized DPF, and Φ_k is k -th primal component, or orthogonal basis vector. Additionally, $m=0$, $m=1$ and $m=2$ represent the preceding DPF (DPF_{pre}), the current DPF (DPF_{cur}), or DPF corresponding to centered phoneme, and the following DPF (DPF_{fol}), respectively. The orthogonal basis was extracted by solving eigenvalue problems after calculating a covariance matrix \mathbf{A} from conventional DPF vectors:

$$(\mathbf{A} - \lambda \mathbf{I})\Phi = \mathbf{0} \quad (2)$$

where λ is the eigenvalue corresponding to orthogonal basis vector and \mathbf{I} is the unit matrix. The covariance matrix \mathbf{A} is calculated as follows.

$$\mathbf{A} = \frac{1}{38} \sum_{\mathbf{x} \in \chi} (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \quad (3)$$

where χ is the set of 38 conventional DPF vectors, that represent 38 Japanese phonemes, with 11 dimensions and $\bar{\mathbf{x}}$ is the mean vector of 38 DPF vectors included in χ . The value obtained by subtracting $\bar{\mathbf{x}}$ from \mathbf{x} is normalized by its norm. Finally, the orthogonalized DPFs, which consist of orth. DPF_{pre} , orth. DPF_{cur} and orth. DPF_{fol} , are used as inputs to an HMM classifier as a sequence of DPF vectors.

3. Experiments

3.1. Speech and Noise Database

The following three data sets were used:

D1. Acoustic model design set with clean speech:

A subset of “ASJ (Acoustic Society of Japan) Continuous Speech Database”, consisting of 4,503 sentences uttered by 30 male speakers (16 kHz, 16-bit).

D2. Test data set with clean speech:

A subset of “Tohoku University and Matsushita Spoken Word Database”, consisting of 100 words uttered by 10 unknown

male speakers each. The sampling rate was converted from 24 kHz to 16 kHz.

D3. Additive noise data set:

A subset of “RWCP Sound Scene Database in Real Acoustic Environments”, consisting of the following three kinds of noise:

- Mobile Phone: the ring tone of a mobile phone
- Particles: the sound when particles fall onto a metal plate
- Whistle: the sound when a whistle is blown

In addition to these three types of noise, white noise is also applied. “Mobile Phone” and “Whistle” are consecutive sounds in a certain frequency band, while “Particles” will contaminate the clean speech in all frequency bands like white noise.

3.2. Experimental Setup

The D1 data set was used to design 43 Japanese monophone HMMs with five states and three loops. In the HMM, output probabilities are represented in the form of Gaussian mixtures, and diagonal matrices are used. Speaker-independent isolated spoken-word recognition tests were carried out with the D2 data set.

3.3. Experimental results and discussion

3.3.1. Effect of orthogonalization and basic performance

(A) Comparison between orth. DPF and non-orth. DPF

Figure 2 shows the experimental results. The orthogonal DPF vector improves the word error rate in comparison with the original DPF vector with correlation between its components. The proposed feature vector achieved the best performance at $K=6$, but degraded the performance at $K=4$.

(B) Comparison with MFCC parameter

Figure 3 illustrates the recognition result. In the baseline system, the input of HMM is the conventional acoustic feature set with 38 dimensions which consists of MFCC with CMN, dynamic features (Δ_i , $\Delta_i \Delta_i$), ΔP and $\Delta \Delta P$. The orthogonalized DPF outperformed the baseline system.

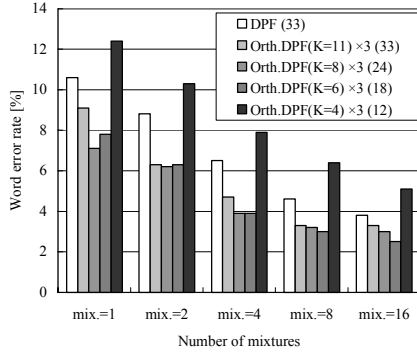


Fig. 2 Orth. DPF vs. non-orth. DPF

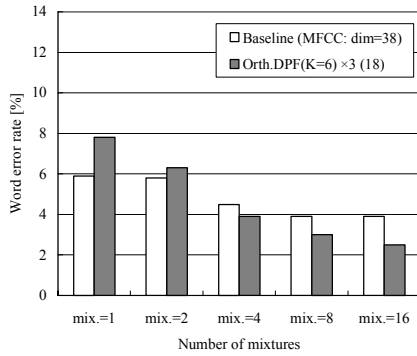


Fig. 3 Orth. DPF vs. MFCC

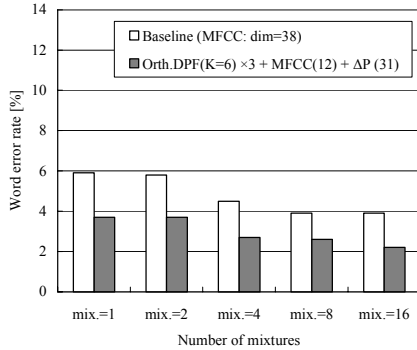


Fig.4 Performance of orth. DPF together with MFCC

(C) Case of a combined feature vector set of DPFs and MFCCs

The recognition result by using the orthogonalized DPF ($K=6$) together with the static MFCC with 12 dimensions and ΔP is shown in Figure 4. The total number of dimensions is 31. The combined usage of DPF and MFCC yielded higher performance than the baseline system at all mixtures because it compensates for each other's errors by representing different characteristics of speech, or complementary information. This result is consistent with other works[2, 8].

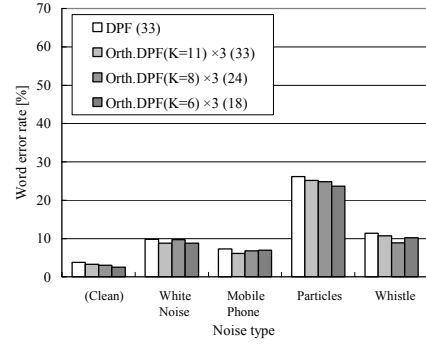


Fig. 5 Comparison of noise robustness: orth. DPF vs. non-orth. DPF, SNR=10 dB

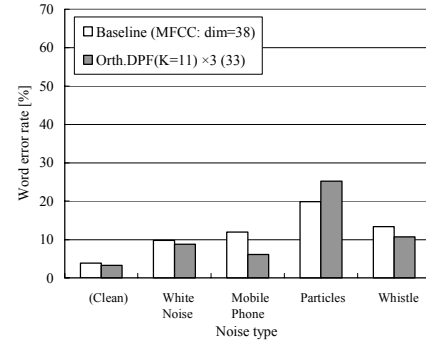


Fig. 6 Comparison of noise robustness: orth. DPF vs. MFCC, SNR=10 dB

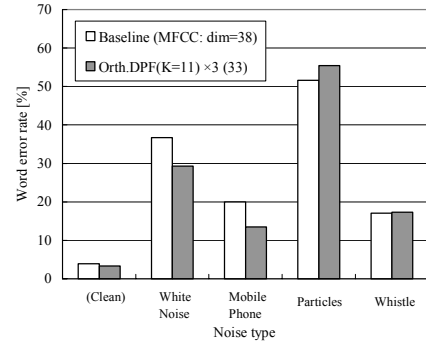


Fig. 7 Comparison of noise robustness: orth. DPF vs. MFCC, SNR=5 dB

3.3.2. Evaluation of noise robustness

(A) Comparison between orth. DPF and non-orth. DPF

Figure 5 shows the recognition result after adding the D3 noise data set and white noise to the D2 data set with SNR=10 dB. As shown in Figure 5, the orthogonalized DPF decreased the word error rate caused by additive noise. The difference of performance by changing the feature dimension is slight, and the orthogonalized DPF showed the best performance at $K=11$ on average.

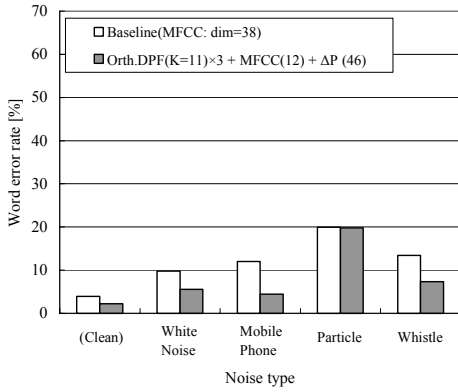


Fig. 8 Comparison of noise robustness: orth. DPF together with MFCC, SNR=10 dB

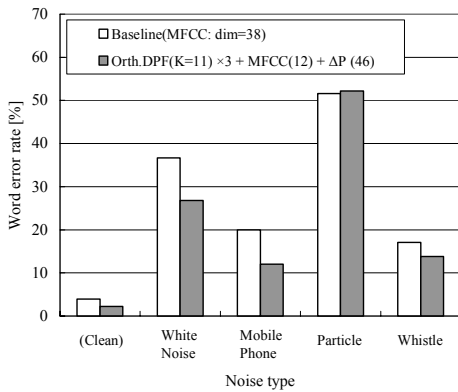


Fig. 9 Comparison of noise robustness: orth. DPF together with MFCC, SNR=5 dB

(B) Comparison with MFCC parameter

Figures 6 and 7 illustrate the experimental result on D2 data contaminated by D3 noise data with SNR=10 dB and SNR=5 dB, respectively. We obtained significant error reduction by using the orthogonalized DPF over the baseline system. Particularly, the orthogonalized DPF showed significant improvement in “White Noise” and “Mobile Phone”. However, the proposed DPF degraded the performance with respect to “Particles” (see [1]). It is considered that the word error rate increased because the high-level noise distributed locally on a time-spectrum space appears in local features (LFs) as large variations.

(C) Case of a combined feature vector set of DPFs and MFCCs

The performance of the orthogonalized DPF ($K=11$) together with MFCC with 12 dimensions and ΔP is shown in Figures 8 and 9 (the total number of dimensions is 46). Combining the orthogonalized DPFs with the static MFCC and ΔP , our ASR system achieved significant improvements for almost all cases, and showed the same performance with regard to even “Particles”. Particularly, the combined feature vector set of DPF and MFCC significantly reduced the word error rate from 9.8% to 5.5% in SNR=10 dB and from 36.7% to 26.8%

in SNR=5 dB for “White Noise”, and from 12.0% to 4.4% in SNR=10 dB and from 20.0% to 12.0% in SNR=5 dB for “Mobile Phone”.

4. Conclusion

The method for orthogonalizing original DPF parameters, output from MLN, by KLT was proposed, and we applied orthogonalized DPF parameters as the input of HMM. In the experiment, the orthogonalized DPF with only 18 dimensions showed higher performance than the baseline system based on conventional MFCC parameters on the isolated spoken-word recognition task with clean speech. Moreover, the DPF vector showed noise robustness, and the combined usage of the orthogonalized DPF and MFCC could significantly improve performance, particularly in “White Noise” and “Mobile Phone”.

In future work, we will discuss how to improve the DPF extractor, and investigate the use of DPF in practical environments.

Acknowledgements

This work was supported in The 21st Century COE Program “Intelligent Human Sensing”, from the ministry of Education, Culture, Sports, Science and Technology, and conducted using the non-speech sounds in an anechoic room (dry sources) data of the RWCP Sound Scene Database in Real Acoustic Environments.

References

- [1] T. Fukuda, W. Yamamoto and T. Nitta, “Distinctive Phonetic Feature Extraction for Robust Speech Recognition,” Proc. ICASSP’03, 2003.
- [2] K. Kirchhoff, G. A. Fink and G. Sagerer, “Combining acoustic and articulatory feature information for robust speech recognition,” Speech Communication, 37, pp.303-319, 2002.
- [3] P. Jain, H. Hermansky and B. Kingsbury, “Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features,” Proc. ICSLP’02, pp.473-476, 2002.
- [4] E. Eide, “Distinctive Features for Use in an Automatic Speech Recognition System,” Proc. Eurospeech’01, pp.1613-1616, 2001.
- [5] N. Chomsky and M. Halle, “The Sound Pattern of English,” New York, Harper and Row, 1968.
- [6] T. B. Martin, “Practical Application of Voice Input to Machine,” Proc. IEEE, 64-4, 1976.
- [7] S. Makino, S. Homma and K. Kido, “Speaker independent word recognition system based on phoneme recognition for a large size (212 words) vocabulary,” J. Acoust. Soc. Jpn., (E) 6, 3, pp.171-180, 1985.
- [8] B. Launay, O. Siohan, A. Surendran and C. H. Lee, “Towards Knowledge-based Features for HMM Based Large Vocabulary Automatic Speech Recognition,” Proc. ICASSP’02, pp.817-820, 2002.