

# Noise-robust ASR by Using Distinctive Phonetic Features Approximated with Logarithmic Normal Distribution of HMM

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

Various approaches focused on noise-robustness have been investigated with the aim of using an automatic speech recognition (ASR) system in practical environments. 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 replace normal distributions (NDs) of DPFs with logarithmic normal distributions (LNDs) in HMMs because DPFs show skew symmetry, or positive and negative skewness. The HMM with the LNDs was firstly evaluated in comparison with a standard HMM with NDs in an experiment using an isolated spoken-word recognition task with clean speech. Then noise robustness was tested with four types of additive noise. In the case of DPFs as an input feature vector set, the proposed HMM with the LNDs can outperform the standard HMM with the NDs in the 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 DPFs with static MFCCs and  $\Delta P$ .

## 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 using a standard HMM with normal distributions (NDs). In the DPF extraction stage, after converting speech signals to acoustic features composed of local features (LFs) and  $\Delta P$ , the 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 the DPFs. The output DPFs could reduce the word error rate on an isolated spoken-word recognition task with additive noise.

On the other hand, in our previous method, the MLN is trained to output the value of 1 for the corresponding DPF element. Because it outputs DPF values close to 1 in the DPF element corresponding to the input phoneme, DPF distributions appear to show asymmetries. Thus, we conjecture that the HMM with the NDs causes recognition error by deviating from real distributions. A HMM recognizer requires distributions that explicitly represent the DPF with skew symmetry.

In this paper, we attempt to represent the DPF distributions by using logarithmic normal distributions (LNDs) instead of the standard NDs. The DPF distribution in a real acoustic model is firstly observed and compared with the MFCC distribution, then a method for replacing NDs with LNDs in the HMM is provided. Experiments are conducted to evaluate recognition accuracy using an isolated spoken-word recognition task with clean speech and with speech contaminated by four types of additive noise. A combined feature vector set of DPFs and static MFCCs is also evaluated.

This paper is organized as follows. Section 2 outlines the implementation of a DPF extractor, and Section 3 shows the comparison between MFCC and DPF distributions. Section 4 describes the experimental setup and results, and provides a discussion, and Section 5 finishes with some conclusions.

## 2. Overview of a 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 $\times$ 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

\*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 previously investigated in speech recognition[6, 7].

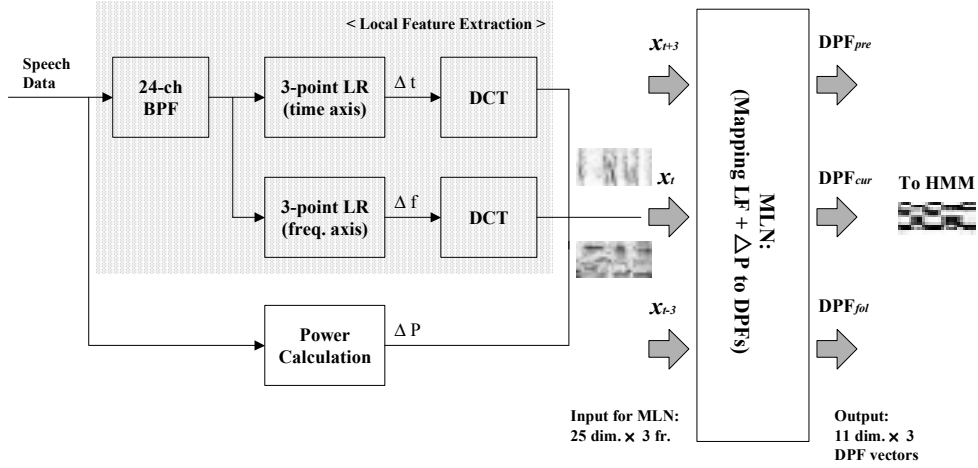


Fig. 1 DPF feature extraction

maximum of 30 and the data is selected using nearest neighborhood clustering. Finally, the outputs of the MLN are used as inputs to an HMM classifier as a sequence of DPF vectors.

### 3. Logarithmic normal distribution (LND)

#### 3.1. MFCC and DPF vector distribution

Figure 2 shows a distribution of MFCC and DPF vector observed from the D1 data set in section 4.1 (phoneme /a/, third state (or, middle state), 6th dimension (voiced)). DPF values are multiplied by 10 for avoiding negative big values by calculating the log of DPF amplitude. As illustrated in Figure 2, the DPF distribution shows negative skewness, whereas MFCC can be efficiently represented by the ND.

Other phonemes and dimensions also showed almost the same tendency except a certain part (DPF elements, not corresponding to input phoneme, showed a distribution with positive skewness). On the other hand, LND is known in the field of statistics. In this distribution, a mode value shifts to the left side from the center position in comparison with ND, namely, it has positive skewness. Figure 3 illustrates an example of ND and LND with negative skewness after converting stochastic variable  $X$  by  $Z = (1.0 - X) \times 10.0$  (dotted line is an example of a standard LND).

#### 3.2. Introduction of LND to HMM

Applying the HMM with ND, output probabilities in an  $m$ -th Gaussian mixture caused by transition from state  $i$  to state  $j$  are calculated as follows.

$$b_{ijm}(\mathbf{o}) = \prod_k \frac{1}{\sqrt{2\pi\sigma_{ijmk}^2}} \exp\left\{-\frac{1}{2\sigma_{ijmk}^2}(o_k - \mu_{ijmk})^2\right\} \quad (1)$$

where  $o_k$ ,  $\mu_{ijmk}$ , and  $\sigma_{ijmk}^2$  are the amplitude of input feature, mean and variance in  $k$ -th dimension, respectively. In contrast with ND, output probabilities of the HMM with LND are computed as follows.

$$b_{ijm}(\mathbf{o}) = \prod_k \frac{1}{o_k \sqrt{2\pi\sigma_{ijmk}^2}} \exp\left[-\frac{1}{2\sigma_{ijmk}^2} \{\log(o_k) - \mu_{ijmk}\}^2\right] \quad (2)$$

The mean  $\mu$  and the variance  $\sigma$  are estimated by calculating a sample average and variance after converting the DPF vector in linear space into that in log space (in case of negative skewness,  $((1.0 - o_k) \times 10.0 \rightarrow o_k)$ ). Therefore, we can apply conventional training methods, such as an EM algorithm, to parameter estimation for HMM with LND. Logarithmic output probabilities based on ND and LND are calculated as follows, respectively.

#### [Normal distribution]

$$b_{ijm}(\mathbf{o}) = -\frac{1}{2} \sum_k \frac{1}{\sigma_{ijmk}^2} (o_k - \mu_{ijmk})^2 - \frac{1}{2} \sum_k \log(2\pi) - \frac{1}{2} \sum_k \log \sigma_{ijmk}^2 \quad (3)$$

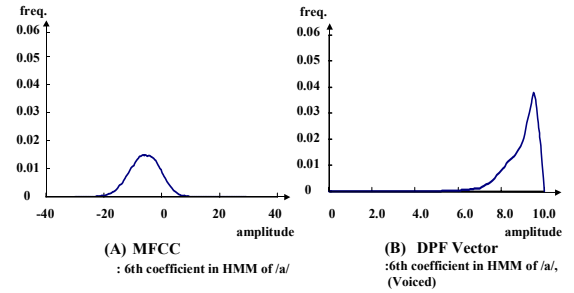


Fig. 2 Examples of MFCC and DPF vector distributions

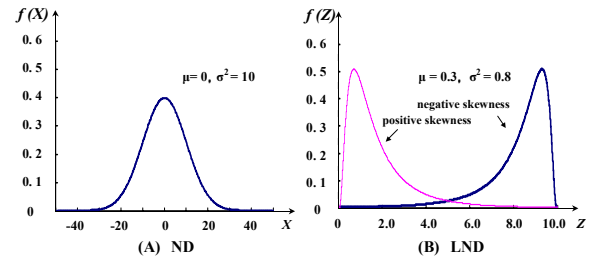


Fig. 3 ND and LND

### [Logarithmic normal distribution]

$$b_{ijm}(\mathbf{o}) = -\frac{1}{2} \sum_k \frac{1}{\sigma_{ijmk}^2} (\log(o_k) - \mu_{ijmk})^2 - \frac{1}{2} \sum_k \log(2\pi) - \frac{1}{2} \sum_k \log \sigma_{ijmk}^2 - \sum_k \log(o_k) \quad (4)$$

## 4. Experiments

### 4.1. 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.

### 4.2. 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 ND and LND with only negative skewness, and diagonal matrices are used. Speaker-independent isolated spoken-word recognition tests were carried out with the D2 data set.

### 4.3. Experimental results and discussion

#### 4.3.1. Comparison between ND and LND

##### (A) Case of DPFs as a feature vector set

Figure 4 shows the experimental results. In the baseline system, the input of HMM with ND is the conventional acoustic feature set with 38 dimensions which consists of MFCC with CMN, dynamic features ( $\Delta_t$ ,  $\Delta_t \Delta_t$ ),  $\Delta P$  and  $\Delta \Delta P$ . The HMM with the LND significantly improved recognition accuracy at the comparatively lower mixtures of 1, 2, and 4 in comparison with the HMM with the ND when DPF parameters were input. At the mixture of 16, HMM with the LND showed better performance than that with the standard ND using conventional MFCC parameters.

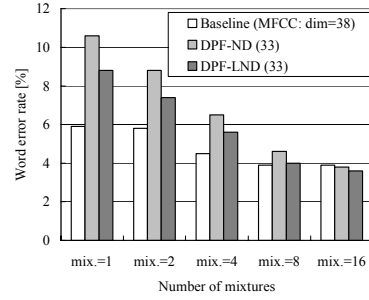


Fig. 4 ND vs. LND

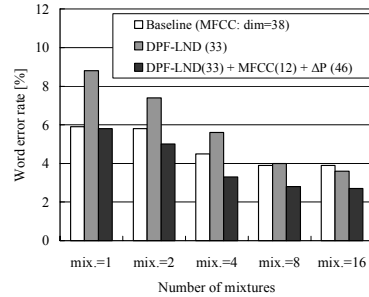


Fig. 5 Performance of DPF together with MFCC

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

Figure 5 illustrates the recognition result by using the DPFs together with the static MFCC with 12 dimensions and  $\Delta P$ . The total number of dimensions is 46. The output probability of DPF vectors was calculated with the LND, and that of MFCC and  $\Delta P$  were computed with the ND. We obtained an improvement of error rate in comparison with the use of only DPFs. The combined usage of DPF and MFCC compensates for each other’s errors because DPF and MFCC represent different characteristics, or complementary information. This result is consistent with other works[2, 8].

#### 4.3.2. Evaluation of noise robustness

##### (A) Case of DPFs as a feature vector set

Figures 6 and 7 illustrate the recognition result after adding the D3 noise data set and white noise to the D2 data set with SNR=10 dB and SNR=5 dB, respectively. The number of mixtures is 16 in all the models. The proposed HMM showed word error reduction for all types of noise in comparison with the HMM with ND for the input DPF vectors. As compared with the baseline system, the proposed HMM degraded the performance with respect to “Particles” (see [1]). It is considered that the recognition accuracy decreased because the high-level noise distributed locally on a time-spectrum space appears in local features (LFs) as large variations.

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

The performance of a combined feature vector set of DPF, MFCC and  $\Delta P$  is shown in Figures 8 and 9. Combining the DPFs with the static MFCCs, the proposed ASR system achieved significant improvements of performance for speech

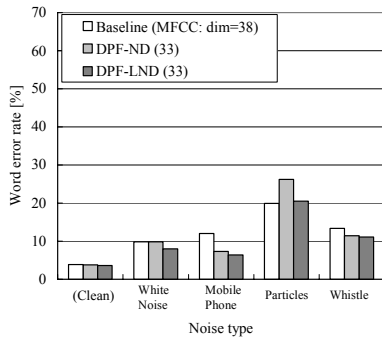


Fig. 6 Noise robustness: DPF, SNR = 10dB

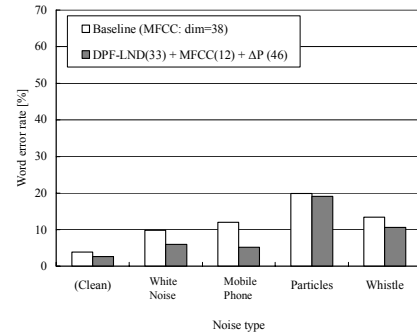


Fig. 8 Noise robustness: DPF+MFCC, SNR = 10dB

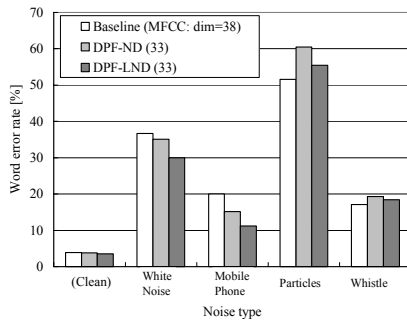


Fig. 7 Noise robustness: DPF, SNR = 5dB

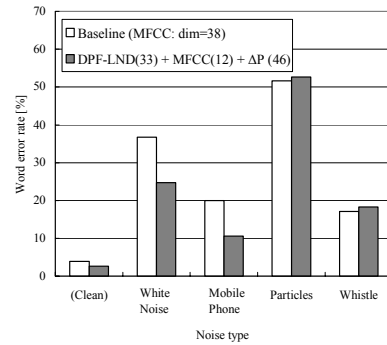


Fig. 9 Noise robustness: DPF+MFCC, SNR = 5dB

contaminated by additive noise, and showed the same accuracy with regard to “Particles”. Particularly, the proposed HMM using DPF together with MFCC reduced the word error rate from 9.8% to 6.0% in SNR=10 dB and from 36.7% to 24.7% in SNR=5 dB for “White Noise”, and from 12.0% to 5.2% in SNR=10 dB and from 20.0% to 10.6% in SNR=5 dB for “Mobile Phone”.

## 5. Conclusion

Application of the LND approximation of DPF, instead of ND, to HMMs was investigated. The LND explicitly represents the distribution of the DPF vector in comparison with standard ND. The proposed HMM with LND showed better performance than the HMM with the ND on the speaker-independent isolated spoken-word recognition tests both with clean speech and with speech contaminated by high-level additive noise. Furthermore, the combined usage of DPF and MFCC could significantly improve the word error rate over the baseline HMM system based on MFCC parameters.

In future work, we will discuss how to improve the DPF extractor, including applying distributions with positive and negative skewness, and investigate the use of DPF in practical environments.

## Acknowledgements

This work was supported in The 21<sup>st</sup> 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.