DISTINCTIVE PHONETIC FEATURE EXTRACTION FOR ROBUST SPEECH RECOGNITION

Takashi FUKUDA, Wataru YAMAMOTO, and Tsuneo NITTA

Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi JAPAN

E-mail: fukuda@vox.tutkie.tut.ac.jp, yamamoto@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

ABSTRACT

This paper describes an attempt to extract distinctive phonetic features (DPFs) that represent articulatory gestures in linguistic theory by using a multi-layer neural network (MLN) and to apply the DPFs to noise-robust speech recognition. In the DPF extraction stage, after converting a speech signal to acoustic features composed of local features (LFs), 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 proposed DPF parameters without MFCC were firstly evaluated in comparison with a standard parameter set of MFCC and dynamic features on a word recognition task using clean speech and the result showed the same performance as that of the standard set. Noise robustness of these parameters was then tested with four types of additive noise and the proposed DPF parameters outperformed the standard set except one additive noise type.

1. INTRODUCTION

A set of MFCC parameters, which is based on the short-term power spectrum and combined with dynamic features, has long been used in automatic speech recognition (ASR) systems. MFCC parameters can represent a log-spectrum envelope of a speech signal efficiently, however, because they are often deformed by the difference of transmission characteristics and/or contaminated by noise, the recognition accuracy of MFCC-based ASR is decreased.

On the other hand, linguists have proposed distinctive phonetic features (DPFs) that represent the manner of articulation (vocalic, consonantal, continuant, …), tongue position (high, front, end, …), etc. and can separate each phoneme. The use of DPFs had been investigated previously in speech recognition, and has been actively discussed again in recent years [1,2,3,4,5,6]. In [1], a set of multi-layer neural networks (MLNs) was used to map acoustic features into DPFs. Each MLN was trained to extract a corresponding DPF, then in the recognition stage, DPFs output from MLNs are combined and used in an HMM classifier. In [2], a Gaussian mixture model (GMM) was used to extract DPFs by comparing the likelihood in articulatory feature presence models with that in articulatory feature absence models. In these previous works, DPFs have been used together with conventional MFCC parameters for an input to an HMM classifier.

We aim to achieve high performance in noise-robust ASR by using DPFs only. The main differences between our proposed method and the previous works are:

(1) Input acoustic features of an MLN are not MFCC but local features (LFs) [7] described in section 2, and
(2) Output DPFs of an MLN, which are extracted using a single MLN, are context-dependent, that is, DPFs with 33 dimensions consist of 11 DPFs, 11 preceding context DPFs, and 11 following context DPFs.

In this paper, firstly the proposed method is compared with a baseline HMM-based ASR system with a standard parameter set of MFCC and dynamic features on an isolated spoken-word recognition task using a clean speech database. Experiments are also carried out to evaluate the differences in input and output parameters. Finally, noise robustness is evaluated using various types of additive noise.

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

2. DPF EXTRACTOR

2.1 Distinctive Phonetic Features (DPFs)

Figure 1 shows a three-dimensional DPF space converted from an original 11-dimensional space of Japanese distinctive features (vocalic, consonantal, high, back, low, anterior, coronal, obstruent, voiced, continuant, nasal) by using the multi-dimensional scaling (MDS) method. As shown in Figure 1, phonologically similar phonemes, such as the group of vocalic and consonantal phonemes, are distributed closely while the others are separated. The merits of using DPFs in speech recognition are described as follows.

A) DPFs can express those phonemes for which the manner of utterance is similar, as close distance vectors.

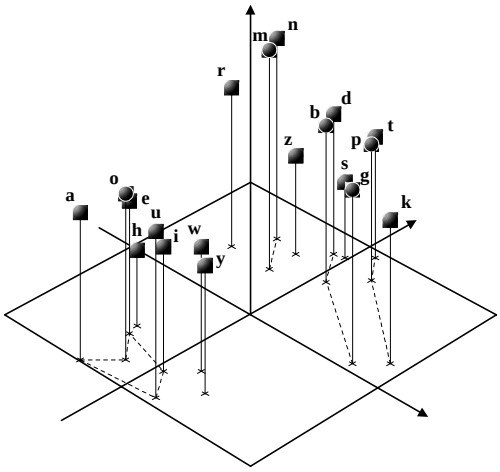


Figure 1. Three-dimensional DPF space by using MDS.

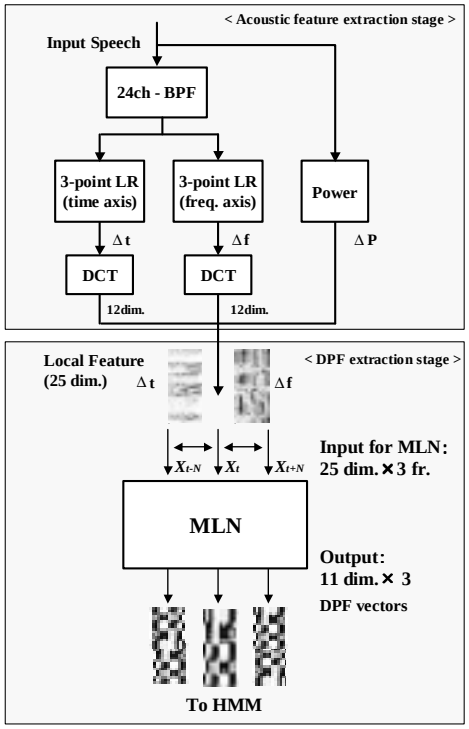


Figure 2. DPF feature extractor.

B) The intermediate expression of DPFs is located between continuous acoustic-feature vectors and discrete representations of words and might be more robust because DPFs explicitly characterize the property in speech production.

We conjecture that the DPFs will achieve accurate recognition in adverse environments.

In the actual implementation of a feature extractor, two articulatory features of "vocalic/non-vocalic" and "consonantal/ non-consonantal" were replaced by "semi-vowel (/j, w, r/)/ non-semi-vowel" and "fricative(/s, z, h/) /non-fricative".

2.2 Design of a DPF Extractor

It is difficult to directly apply DPFs to speech recognition because articulatory gestures are not always the same between individuals, speech acquisition environments and speakers' articulatory organs vary, and so forth. This section describes the mapping procedure from acoustic features to DPFs with a single MLN.

The proposed feature extractor is illustrated in Figure 2. At the acoustic feature extraction stage, firstly, an input speech is converted into LFs. The LFs are then entered into an MLN after combining a current frame X_t with the other two frames that are N-points before and after from 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 MLN is trained to output the value of 1 for the corresponding DPF elements with an input phoneme and its adjacent phonemes. Finally, the outputs of the MLN are used for an input 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 Acoustical 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.

3.2 Spectrogram of Noise

Figure 3 shows the spectrum patterns of the three types of additive noise used in the experiments. As shown in Figure 3, "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.

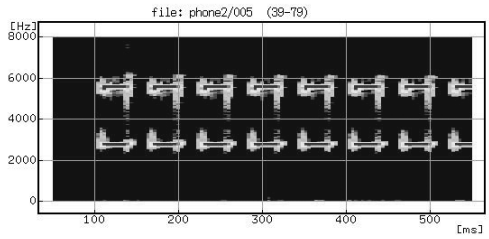
3.3 Experimental Setup

3.3.1 Acoustic feature parameters

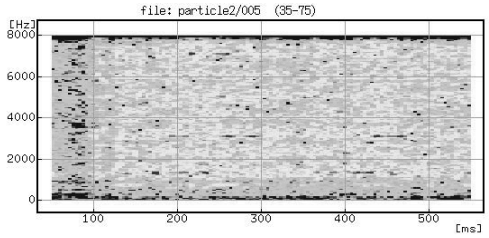
The following two acoustic features were investigated for the input of MLN.

(A) DPF(MFCC)

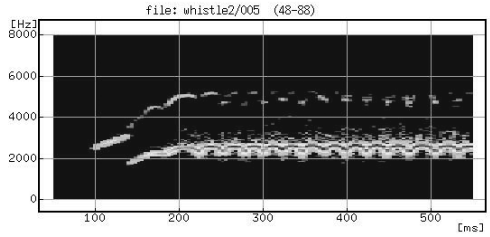
An input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segments is applied every 10 ms. The resultant FFT power spectrum is then integrat-



(A) Mobile phone



(B) Particles(the sound when particles fall onto a metal plate)



(C) Whistle

Figure 3. Spectrogram of noise.

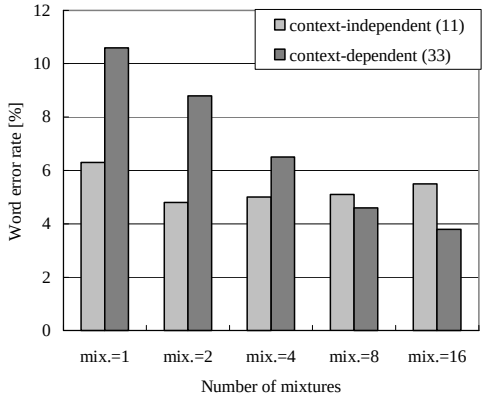


Figure 4. Comparison of the configuration of MLN output unit.

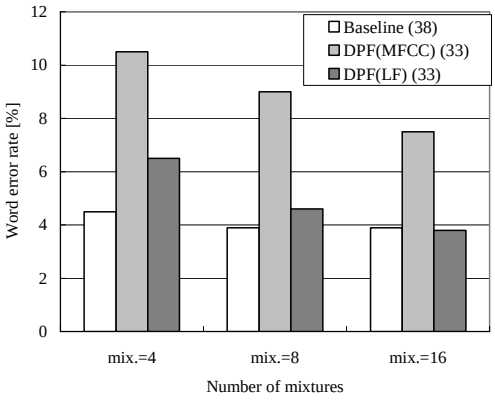


Figure 5. Experimental result: clean speech.

ed into 24-ch BPFs output with mel-scaled center frequencies. Then, 25 feature parameters including 12 static parameters (mel-cepstrum), 12 dynamic features (Δ_t) and ΔP (logarithmic power) are extracted after converting the output of BPFs into cepstrum coefficients (MFCC) by using DCT. MFCC parameters are processed with CMN for every utterance.

(B) DPF(LF: Local Features)

Two LFs are firstly extracted by a three-point linear regression (LR) calculation along the time and frequency axis on a time spectrum pattern [8]. Then, after converting the two LFs into cepstrum with 12 dimensions by using DCT, respectively, 25 feature parameters including ΔP are composed.

3.3.2 MLN structure

The acoustic feature parameters described above were used to train a single MLN with four layers including two hidden layers. Each layer consists of 75, 256, 64, and 33 units from the input layer, respectively. An MLN input of MFCC-based parameters is combined with continuous three frames, while the input of LF-based parameters is combined with the current frame and two adjacent frames that are three points before and after.

Two input acoustic feature parameters of DPF(MFCC) and DPF(LF) use the same frame length. The MLN is trained using a back-propagation algorithm. The number of training data of each tri-phone is limited to a maximum of 30 and the data is selected using nearest neighborhood clustering.

3.3.3 HMM acoustic model

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.

3.4 Experimental Results and Discussion

Speaker-independent isolated spoken-word recognition tests were carried out with the D2 data set.

(A) Comparison of MLN structure

The difference of MLN structures was evaluated. Figure 4 shows the experimental result. The MLN with 33 context-dependent output units yields higher performance than that with 11 context-independent output units at the comparatively higher mixtures of 8 and 16.

The improvement of performance by MLN with context-dependent is considered to be as follows. The MLN with context-independent could separate the phonological information between phonemes by mapping into 11-dimensional DPFs, however, the performance was not increased at the higher mixtures because it often occurred the mapping error in an adjacent phoneme boundary. On the other hand, although the MLN with context-dependent also occurred the mapping error, the context of output could reduce the effect of that.

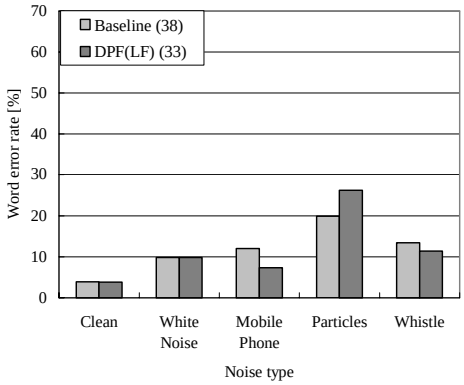


Figure 6. Experimental result: noisy data. (SNR=10 dB, Mixture=16)

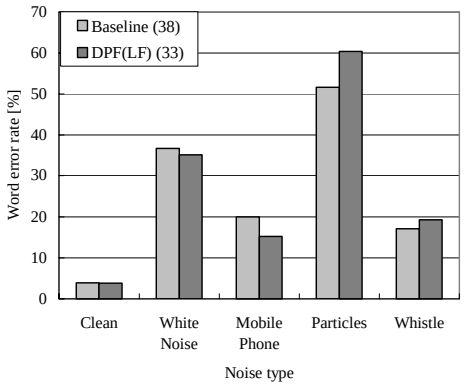


Figure 7. Experimental result: noisy data. (SNR=5 dB, Mixture=16)

(B) Comparison of input acoustic parameters of MLN

Figure 5 shows the experimental result. In the baseline, the input of HMM is the conventional acoustic feature set with 38 dimensions which consists of MFCC with CMN, dynamic features (Δ_t, $\Delta_t\Delta_t$), ΔP and $\Delta\Delta$P. DPF(MFCC) degraded the performance regardless of the number of mixtures while DPF(LF) showed better performance than DPF(MFCC). The proposed DPF(LF) without MFCC parameters achieved the same performance in comparison with the baseline parameter.

(C) Evaluation of noise robustness

Figures 6 and 7 illustrate the recognition result after adding D3 noise data set and white noise to the D2 data set with SNR=10 dB and SNR=5 dB, respectively. The proposed DPF(LF) showed word error reduction for the three types of noise except "Particles". Particularly, in "Mobile Phone", DPF(LF) significantly improved the word error rate from 12.0% to 7.3% in SNR=10 dB and from 20.0% to 15.2% in SNR=5 dB.

The LF parameter represents the variance of the log-power spectrum along the time and frequency axis. Thus, when speech is contaminated by noise distributed over all frequency bands such as "Particles", then LFs obtain little phonologic information, and hence the proposed DPF(LF) parameter reduces the performance. With respect to the "Particles" in another experiments, by using the DPFs together with the conventional MFCC, we obtained the improvement of error rate from 26.2% to 23.7% in SNR=10 dB and from 60.4% to 55.9% in SNR=5 dB in comparison with the use of only DPFs. We need further study on this point.

4. CONCLUSION

A novel feature extractor based on distinctive phonetic features was proposed. Acoustic parameters were mapped to DPFs by using an MLN with context-dependent output units. LFs showed better performance than MFCC as an input of the MLN. The proposed DPF without the conventional MFCC parameter provided almost the same results as the standard MFCC-based feature parameter in HMM-based isolated spoken-word recognition experiments with clean speech, and could significantly reduce the effect of high-level additive noise, particularly the ring tone of a 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 conducted using the non-speech sounds in an anechoic room (dry sources) data of the RWCP Sound Scene Database in Real Acoustic Environment.

REFERENCES

[1] B. Launay, O. Siohan, A. Surendran and C. H. Lee, "Towards Knowledge-based Features for HMM Based Large Vocabulary Automatic Speech Recognition," Proc. of IEEE ICASSP'02, pp.817-820 (2002).

[2] E. Eide, "Distinctive Features For Use in an Automatic Speech Recognition System," Proc. of Eurospeech'01, pp.1613-1616 (2001).

[3] 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).

[4] H. Hermansky and S. Sharma, "TRAPS – Classifiers of Temporal Patterns," Proc. of ICSLP'98, pp.1003-1006 (1998).

[5] P. Jain, H. Hermansky and B. Kingsbury, "Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features," Proc. of ICSLP'02, pp.473-476 (2002).

[6] H. Tolba, S. A. Selouani and D. O'Shaughnessy, "Comparative Experiments to Evaluate the Use of Auditory-based Acoustic Distinctive Features and Formant Cues for Automatic Speech Recognition Using a Multi-stream Paradigm," Proc. of ICSLP'02, pp.2113-2116 (2002).

[7] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. of IEEE ICASSP'99, pp.421-424 (1999).

[8] T. Fukuda, M. Takigawa and T. Nitta, "Peripheral Features for HMM-based Speech Recognition," ICASSP'01, pp.129-132 (2001).