2007 RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'07)
Shanghai Jiao Tong University, Shanghai, China, Mar. 3-6, 2007

DISTINCTIVE PHONETIC FEATURE (DPF) BASED PHONE SEGMENTATION USING 2-STAGE MULTILAYER NEURAL NETWORKS

Huda Mohammad Nurul, Muhammad Ghulam, Kouichi Katsurada, Yurie Iribe and Tsuneo Nitta
Toyohashi University of Technology, Aichi, Japan

ABSTRACT

Segmentation of speech into its corresponding phones has become very important issue in many speech processing areas such as speech recognition, speech analysis, speech synthesis, and speech database. In this paper, for accurate segmentation in speech recognition applications, we introduce Distinctive Phonetic Feature (DPF) based feature extraction using a two-stage MLN (Multi-Layer Neural Network) system consists of an MLN$_{\text{LF-DPF}}$ in the first stage and an MLN$_{\text{Dyn}}$ in the second stage. The MLN$_{\text{LF-DPF}}$ maps continuous acoustic features, Local Feature (LF), onto discrete DPF patterns, while the MLN$_{\text{Dyn}}$ constraints DPF context or dynamics in an utterance. The experiments are carried out using Japanese triphthong data. The proposed DPF based feature extractor provides good segmentation and high recognition rate with a reduced mixture-set of HMMs (Hidden Markov Models) by resolving co-articulation effect.

KEY WORDS

Distinctive Phonetic Feature, Recurrent Neural Network, Multi-Layer Neural Network, Hidden Markov Model, Local Features.

1. Introduction

A variety of methods have been proposed to accomplish phone segmentation [1]-[4]. Although some of them showed acceptable performances, most of the HMM (Hidden Markov Model)-based method uses context sensitive tri-phone model to resolve co-articulation effect. However, these models need a large number of speech parameters and a large volume of speech corpus. In order to overcome these drawbacks, the proposed method incorporates context effect to neural networks. Because it is not a parametric model, it eliminates the degradation of performance in noisy acoustic environment. Besides, current HMM-based ASR (Automatic Speech Recognition) systems always reject a new vocabulary or so called out-of-vocabulary (OOV) word. Accurate phone segmentation is expected to get an accurate phonetic typewriter functionality that will assist the next generation ASR systems in solving the OOV-word problem of HMM through short interaction (talk back) by enrolling the word into the database automatically rather than manually.

In this paper, from the segmentation performance and computational points of view, we investigate and evaluate four types of DPF-based feature extraction methods together with the conventional method of MFCC (Mel Frequency Cepstral Coefficient). Four types of feature extractors are (i) MLN (Multilayer Neural Networks)[5], (ii) RNN (Recurrent Neural Network), (iii) RNN-MLN and (iv) two-stage MLN. RNN can represent dynamics in a sequence of acoustic features, that is, uncertain evidence is accumulated over many time frames in order to build up an accurate representation of the long term context variables. But the main problem of RNN is the slowness of network of large size during training. Without proper training for large network better context effect is impossible [6]. The proposed two-stage MLN approach generates more precise phonetic segment as well as higher recognition accuracy with less computation by resolving co-articulation effect without using RNN.

The paper is organized as follows: Section 2 discusses about the phonological features. Section 3 explicates the system configuration of the existing methods. Section 4 illustrates the system configuration of RNN-MLN approach. The proposed two-stage MLN based method is explained in Section 5. Experimental database and setup are provided in Section 6, while experimental results are analyzed in Section 7. Finally, Section 8 draws some conclusion.

2. The Articulatory Features

A phone can easily be identified by using its unique articulatory features or Distinctive Phonetic Feature (DPF) set [7], [8]. Besides, a DPF-set can classify a phone at lower mixture of HMM [5]. Because the traditional DPF set (high, low, anterior, back, coronal, plosive, continuant, fricative, nasal, voiced, semi-vowel) is not designed for ASR system; the feature vector space composed of the traditional DPF is not necessarily

Table 1: DPFs for vowels.

	High	Low	Nil	Back	Nil
a	-	+	-	+	-
i	+	-	-	-	+
u	+	-	-	+	-
e	-	-	+	-	+
o	-	-	+	+	-

suitable for classifying speech signal [5]. A novel DPF set with 15 elements is introduced by modifying traditional set [9]. These DPF values are mora, <high, low, nil>, <anterior, back, nil>, coronal, plosive, affricate, continuant, voiced, unvoiced, nasal and semi-vowel. Table 1 shows the distinguishable features of vowels. Here, present and absent elements of the DPF are indicated by "+" and "-" signs, respectively.

3. DPF-based Phonetic Segmentation using MLN

Figure 1 shows a process of DPF-based phonetic segmentation using MLN. At the acoustic feature extraction stage, input speech is converted into Local Features(LF) (LF/Δ_t, LF/Δ_f), which represent a variance of spectrum along time and frequency axes by using three-point linear regression (LR) calculation. LFs are input to an MLN with three layers, including 2 hidden layers, after combining preceding t-3 th and succeeding t+3 th frames with the current t th frame. Each frame of LF consists of 25 values (12Δt+12Δf+ΔP). The MLN has 45 output units (15x3) corresponding to context-dependent DPF vector that consists of three DPF vectors of 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. The MLN is trained by using a standard back-propagation algorithm. However, single MLN suffers from its inability to model dynamic information.

4. DPF-based Phonetic Segmentation using RNN

We proposed a DPF-based system with RNN (Recurrent Neural Networks) to overcome the problem of one-stage MLN [11]. RNN for its capability of handling a longer context window is the basics of this network. The aim of the network is to map a sequence of frames of parameterized speech onto a sequence ofphone labels associated with those frames by resolving context. A single frame of parameterized speech is replaced by an acoustic vector containing several adjacent frames along with the original central frame to increase the dimensionality of the acoustic vector. Figure 2 shows a block diagram of a DPF-Based phonetic segmentation method using an RNN. Higher dimensional input acoustic vector is formed by taking preceding (t-3) th and succeeding (t+3) th frames together with the current t th frame. Each input frame is formed by 25 LF values same as DPF-based phonetic segmentation using MLN. The RNN generates 45 DPF values of which 15 are for the preceding frame, 15 for the current frame and the rest for the succeeding frame. An MLN is used to integrate these 45 DPF values to obtain 15 DPF output for the current frame.

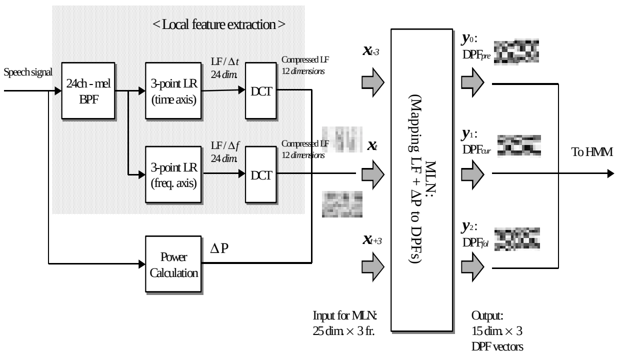

Figure 1: System for MLN-based DPF extractor.

A fully recurrent neural network (FRNN) of two layers is used for this approach. Hidden layer consists of 180 hidden units. Each time total input vector is formed by taking output layer (OL) feedback values and hidden layer (HL) feedback values together with external input (25x3) LF values of that time. Feedback values of hidden and output layer at time t_0 are assumed to be 0.1. Epoch wise back-propagation through time algorithm is used for training. MLN for this approach has 3 layers (2 hidden layers, output layer). Hidden layer one and two are of 90 and 30 hidden units, respectively. Standard back-propagation is used as learning algorithm.

The main problem of RNN is the slowness of network of large size during training. Without proper training for large network better context effect is impossible.

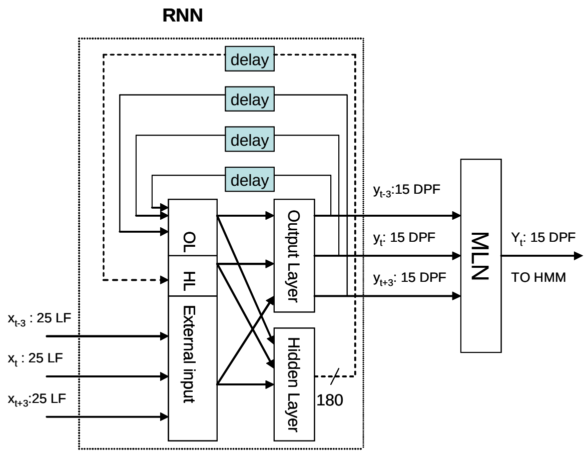

Figure 2: System for RNN-based DPF extractor

5. Proposed Method with dynamics in DPF space

Figure 3 shows the system configuration of DPF-based phonetic segmentation using two-stage MLN. Single MLN can not handle context effect rigorously. Again, DPF-based segmentation using RNN can not provide good segment and contextual information without proper training of large and complex network. The highest level performance of ASR faces challenge to capture the context information. The proposed system replaces the first stage RNN of DPF-based system with RNN described in Section 4, by a MLN to reduce the time complexity to a certain limit. Again better context effect is maintained by forming higher dimensionality of input acoustic vector of preceding (t-3) th and succeeding (t+3) th frames together with the current t th frame. Each input frame is formed by 25 LF values same as DPF-based phonetic segmentation using MLN.

The first stage MLN, named $\mathrm{MLN}_{\mathrm{LF\text{-}DPF}}$, generates 45 DPF values of which 15 are for the preceding frame, 15 for the current frame and the rest for the succeeding frame from LF. The second stage MLN, named $\mathrm{MLN}_{\mathrm{Dyn}}$, is used to correlate these 45 DPF values to obtain 15 DPF output for the current frame, and thereby, realizes dynamic property of utterance one degree higher than that using only one stage MLN. The $\mathrm{MLN}_{\mathrm{Dyn}}$ also reduces the DPF fluctuation caused at starting of an utterance and the effect of previous phone on the current phone.

The $\mathrm{MLN}_{\mathrm{LF\text{-}DPF}}$ uses three layers (2 hidden layers, output layer). First hidden and second hidden layer from the input side and output layer are of 256, 96, and 45, units respectively. Each time total input vector is formed by taking the external input (25x3) LF values of preceding, current and succeeding time. Standard back-propagation algorithm is used for training the network of reasonable size. The $\mathrm{MLN}_{\mathrm{Dyn}}$ has 3 layers (2 hidden layers, output layer). Hidden layer one and two are of 90 and 30 hidden units, respectively. Standard back-propagation algorithm is also used for training $\mathrm{MLN}_{\mathrm{Dyn}}$.

6. Experiments

6.1. Speech Database

A subset of “ASJ (Acoustic Society of Japan) continuous speech Database”, consisting of 1316sentences uttered by male speakers is used as training. 60 tripthongs (three consecutive vowels only, such as /aei/, /uoi/, etc.) are used as test data. Sampling rate is 16KHz.

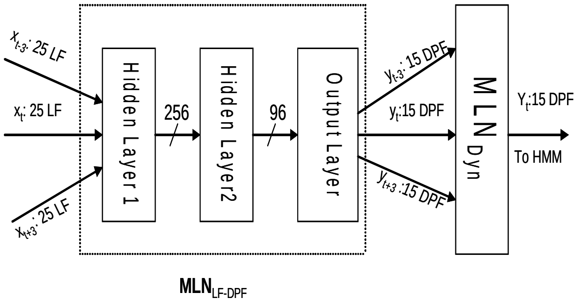

Figure 3: Proposed DPF extractor with dynamics.

6.2. Experimental Set up

Frame length and frame rate are set to be 25 ms and 10 ms, respectively. MFCC consists a vector of 38 dimensions (12 MFCC, 12Δ, 12ΔΔ, ΔP and ΔΔP, where P is log energy of raw signal). Conventional approach of ASR systems uses MFCC as feature vector to be fed into an HMM-based classifier. Minimum Maharanobis distance is calculated for each frame of test utterance and hence, a minimum distance phone is obtained for phone segmentation.

For recognition, input features for HMM-based classifier are:

i) MFCC
ii) DPF from LF using MLN [3]
iii) DPF from LF using RNN
iv) DPF from LF using RNN+MLN
v) The proposed method

Delta parameters (15Δ DPF, 15ΔΔ DPF) are appended with extracted DPF from methods iii), iv) and v) to make input dimension 45 for HMM classifier. The classifier in this experiment adopts a standard monophone-based HMM with 5-state 3-loop left-to-right models.

Segmentation is evaluated as frame-by-frame basis and frame correct rate is calculated by the following equation:

$$\text{frame correct rate} = (1 - \frac{No.of\ erroneus\ frames}{Total\ no.of\ \ frames}) \times 100\%$$

7. Experimental Results and Discussion

Table 2 shows frame correct rate for all the investigated methods. MFCC shows frame correct rate 54.72%, which is the worst over all the methods. The proposed method with dynamics in DPF space shows the highest number of frame correct rate (88.16%). DPF using MLN generates frame correct rate (70.62%), which is 14.15% lower than the proposed method.

Table 3 shows that the highest recognition accuracy (81.16%) using Gaussian distribution at mixture 4 is obtained by using the proposed method. The proposed method provides 78.26% recognition rate at mixture 2 whereas other methods take more mixture(s) to generate same accuracy. Additional mixture(s) requires more computation cost. So, less computation demand by the proposed method is realized.

Figure 4 shows the phonetic segmentation for utterance [iai] where considering DPF value is "back". DPF using MLN and RNN+MLN show deviation of phone boundary from its ideal position. The proposed method shows no shifting of phone boundary.

8. Conclusion

This paper has presented a relatively simple phone segmentation system by resolving co-articulation effect. DPF-based method using two-stage MLN achieved accurate segmentation on triphthong task. The proposed method has also provided high recognition accuracy with less computation. The effect of the proposed method on all types of phone will be investigated in future.

Acknowledgement

This work was supported in The 21st Century COE Program "Intelligent Human Sensing", from the ministry of Education, Culture, Sports, Science and Technology, Japan.

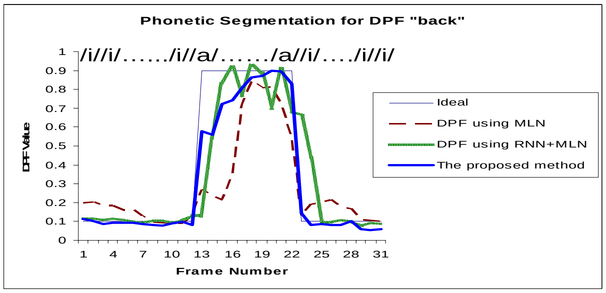

Figure 4: Investigated Segmentation for /iai/.

Table 2: Frame correct rate (%) with different methods.

Investigated methods	Frame correct rate(%)
MFCC	54.72
DPF using MLN	70.17
DPF using RNN	70.62
DPF using RNN+MLN	76.01
The proposed method	88.16

Table 3: Recognition rate (%) with different methods.

Investigated Methods	Recognition Rate(%)		
	Mix.1	Mix.2	Mix.4
MFCC	69.57	76.81	71.01
DPF using MLN	68.12	72.46	75.36
DPF using RNN	69.57	73.91	72.46
DPF using RNN+MLN	72.46	76.81	78.26
The proposed method	75.36	78.26	81.16

REFERENCES

[1] Victor W. Zue, "The Use of Speech Knowledge in Automatic Speech Recognition," Proceedings of the IEEE, Vol. 73, pp. 1602-1615, Nov. 1985.

[2] Weinstein, C. J., McCandless, S. S., Mondshein, L. F., and Zue, V.W. "A System for Acoustic-Phonetic Analysis of Continuous Speech," IEEE Trans. ASSP, Vol. 23, pp. 54-67, Feb. 1975.

[3] Grayden, D. B and Scordilis, M. S. "Phonemic Segmentation of Fluent Speech," Proc. ICASSP-94, pp. 73-76, 1994.

[4] Buniet, L. and Fohr, D. "Continuous Speech Segmentation with the Gamma Memory Model," Proc. of EUROSPEECH'95, pp. 1685-1688, 1995.

[5] T. Fukuda, et al, "Orthogonalized Distinctive Phonetic Feature Extraction for Noise-robust Automatic Speech Recognition," IEICE Trans. Info. & Sys., vol.E87-D, no.5, pp.1110-1118, 2004.

[6] T. Robinson, "An application of Recurrent Nets to Phone Probability Estimation," IEEE Trans. Neural Networks, Volume 5, Number 3, 1994.

[7] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech using Neural Networks," Computer Speech and Language 14(4), pp. 333-345, 2000.

[8] E. Eide, "Distinctive Features for Use in an Automatic Speech Recognition System," Proc. Eurospeech 2001, vol.III, pp.1613-1616, 2001.

[9]T. Fukuda and T. Nitta, "A Study on Japanese Distinctive Phonetic Feature Set for Robust Speech Recognition," The 2003 Autumn Meeting of The Acoustical Society of Japan, Vol.I, 1-6-5, pp.9-10, September 2003, *in Japanese*.

[10] Nitta, et al, "Representing local features for speech recognition in cepstrum domain," Proc. Spring Meeting of ASJ, pp.131-132, 2001.

[11] Huda, et. al, "DPF Based Phonetic Segmentation using Recurrent Neural Networks," Proc. Autumn Meeting of ASJ, pp. 3-4, 2006.