Using Phonetic Context for Continuous Speech Recognition with Invariant Structure

☆Congying Zhang (Tokyo Univ., IBM), Masayuki Suzuki, Gakuto Kurata,
Masafumi Nishimura (IBM), Nobuaki Minematsu (Tokyo Univ.)

1 Introduction

The speech signal inevitably varies according to non-linguistic acoustic factors, such as age, gender, microphone, background noise, and so on. These variations often degrade the performance of Automatic Speech Recognition (ASR).

Recently, an invariant structure of speech was proposed, through which the speech is represented without effect of variations by these non-linguistic factors [1]. The invariant structure models spectral contrast between acoustic events (e.g. phonemes). This approach has been applied both in isolated word recognition [2][3], and N-best candidates reranking for continuous speech recognition [4][5]. It showed robustness and good performance on these tasks.

It has been shown that phonetic context-dependent models result in better performance in ASR than phonetic context-independent models. However, in existing study, the invariant structures are merely extracted from phonetic context-independent condition. Therefore, the effect of using phonetic context in structural approach should be further studied.

In this paper, the approach of extracting invariant structure in phonetic context-dependent conditions for continuous speech recognition is proposed. The task of continuous digits speech recognition task and large vocabulary continuous speech recognition (LVCSR) task are applied for testing the performance.

2 Related works

2.1 Invariant structure

Voices of different speakers show different timbre because they have different vocal tract lengths and shapes. By using a mathematical model of voice mapping or transformation characterizing variations of vocal tract length and shape, voice from one speaker can be converted into another speaker's. This fact indicates that if we can find any transform-invariant features, they will be robust features.

A necessary and sufficient condition for a feature to be invariant is that the feature is represented by *f*-divergence. *f*-divergence between two distributions is a family of divergences which are invariant to invertible differentiable transform. For example, the well-known Bhattaryya distance is a kind of *f*-divergences. Consider a feature space $\boldsymbol{X}$ and a pattern $\boldsymbol{S}$. Suppose $\boldsymbol{S}$ has $\boldsymbol{M}$ events $\{s_i\}_{i=1}^{M}$. Each is described as a distribution $s_i(x)$ in the feature space $\boldsymbol{X}$. Assume there is an invertible transformation $f : \boldsymbol{X} \to \boldsymbol{X}'$, which transforms feature space $\boldsymbol{X}$ into a new feature space $\boldsymbol{X}'$. In this way, $\boldsymbol{M}$ events $\{s_i\}_{i=1}^{M}$ in $\boldsymbol{X}$ is mapped into $\{s_i'\}_{i=1}^{M}$ in $\boldsymbol{X}'$. Here, the *f*-divergence between events s_j and s_k $(1 \le j < k \le M)$ is invariant to any arbitrary invertible transform *f*. Therefore, it is equal to the *f*-divergence between s_j' and s_k'.

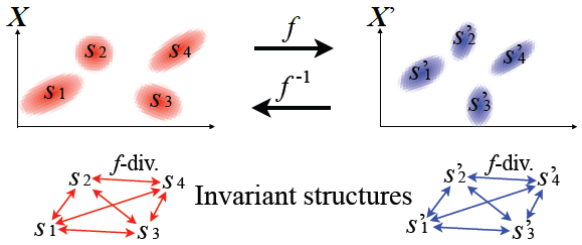


Fig. 1 Invariant structure

Fig. 1 shows two invariant structures. By calculating the *f*-divergences between each two events (phonemes) pair in a pattern, we can obtain a structure. Each pattern is consist of 4 events, so there are 6 edges in each structure. Since *f*-divergence (edge length) is invariant to any invertible transformation, the obtained structure is robust to acoustic condition variation which can be expressed by an invertible transformation *f* to the feature space, such as speaker difference and microphone difference.

2.2 Discriminative statistical edge model (SEM)

In continuous speech recognition, by calculating the invariant structure, the phoneme alignment of an acoustic input can be mapped into a feature vector $\Phi(x, y)$. It is a function of acoustic input x and its candidate y. $\Phi(x, y)$ is formed as the same way of $\Phi_{dr}(x, y)$ in the

---

構造的表象による連続音声認識における音素コンテキストの利用 by C. Zhang (Tokyo Univ., IBM), M. Suzuki, G. Kurata, M. Nishimura (IBM), N. Minematsu (Tokyo Univ.)

Input: Training samples $\left(x_i, \overline{y_i}, \underline{y_i}\right)$ for $i = 1 \ldots I$

Initialization: $\alpha_0^I = 0$

1: **for** $t = 1 \ldots T$ **do**

2: $\quad \alpha_t^0 = \alpha_{t-1}^I$

3: $\quad$ **for** $i = 1 \ldots I$ **do**

4: $\quad\quad$ **if** $\alpha_t^{i-1} \cdot \boldsymbol{\Phi}(x_i, \overline{y_i}) + \phi_0(x_i, \overline{y_i}) > \alpha_t^{i-1} \cdot \boldsymbol{\Phi}\left(x_i, \underline{y_i}\right) + \phi_0\left(x_i, \underline{y_i}\right)$ **then**

5: $\quad\quad\quad \alpha_t^i = \alpha_t^{i-1} + \lambda\left(\boldsymbol{\Phi}\left(x_i, \underline{y_i}\right) - \boldsymbol{\Phi}(x_i, \overline{y_i})\right)$

Output: $\alpha = \sum_{i,t} \alpha_i^t / IT$

Fig. 2 A variant of averaged perceptron algorithm

discriminative reranking for LVCSR. Number of words found in candidate y can be a feature. For example, as in equation (1), the number of word "foo" or "bar" in candidate y can be applied in feature of $\Phi_{dr}(x, y)$. One possible way of using the invariant structure for $\Phi(x, y)$ is that using the classes of edges instead of the words in (1).

$$\Phi_{dr}(x, y) = \begin{bmatrix} \text{the number of "foo" in } y \\ \text{the number of "bar" in } y \\ \vdots \end{bmatrix} \quad (1)$$

Averaged perceptron algorithm is applied for discriminative modeling of the feature vector $\Phi(x, y)$. This model is named as discriminative SEM. Parameter α of the model is learned according to $\Phi(x, y)$. It is interpreted as degree of importance to all classes in the feature. For a candidate y, $\alpha \cdot \Phi(x, y)$ is the structure score, and $\phi_0(x, y)$ is a scalar parameter, which is the log likelihood by a traditional ASR system. $\alpha \cdot \Phi(x, y) + \phi_0(x, y)$ is the new score for candidate y. Fig. 2 shows the learning process of α by averaged perceptron algorithm. $\overline{y_i}$ and $\underline{y_i}$ shows the highest-WER candidates and the lowest-WER candidates in N-best candidates of x_i, respectively. i is the number of training data. t is the number of iterative training. λ is a parameter of learning rate.

For the feature applied in large vocabulary continuous speech recognition (LVCSR) task, language model score is also added to the feature.

2.3 Reranking framework

Discriminative reranking has been applied in our previous study of leveraging the invariant structure in continuous speech recognition. The framework is shown in Fig.3. There are four steps in this process. First, acoustic input x is recognized as N-best candidates set y by a baseline ASR system. Second, invariant structure features

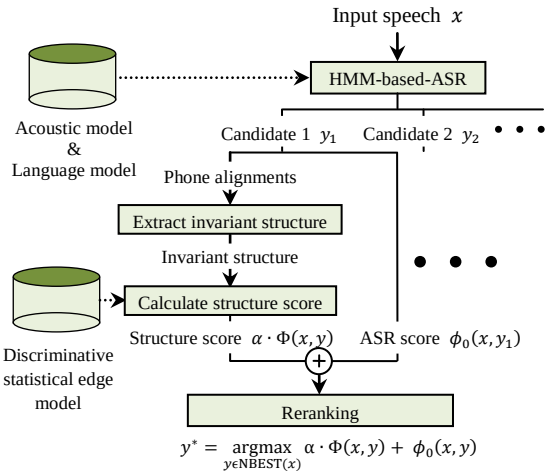


Fig. 3 Reranking frame work

are calculated according to phone alignment of each candidate. Third, structure score is calculated according to the discriminative SEM. In the final step, these candidates are reranked according to the new scores each of which is obtained by combining the structure score and the ASR score. The best candidate can be acquired as in equation (2). Through this approach, better performance than baseline ASR system has been acquired.

$$y^* = \underset{y \in \mathrm{NBEST}(x)}{\mathrm{argmax}}\ \alpha \cdot \Phi(x, y) + \phi_0(x, y) \quad (2)$$

3 Proposed approach

Although pervious approaches resulted in performance improvement to the baseline ASR system, for calculating the invariant structure, edges between phonemes are labeled in a phoneme context-independent way. In the early study of traditional ASR approaches, it has already been proven that phonetic context dependent (triphone) condition resulted in better recognition performance [6]. Therefore, phonetic context condition should be introduced for the extraction of invariant structure.

In this approach, triphone condition is applied for defining phoneme-to-phoneme edges for continuous speech recognition. The discriminative SEM in chapter 2.2 and the framework in chapter 2.3 remain to be applied. However, the acoustic events will be set as triphone alignment instead of monophone alignment.

The problem in triphone condition is that it is not practical to apply all triphone classes because its number is too huge. The number will double in order while it turns to edge classes corresponding to phoneme pairs. This will result in

sparseness for training discriminative SEM. Therefore, the classes of triphone need to be clustered to a relatively optimum number to ensure both enough information of phonetic context and also an acceptable number of triphone classes.

In this paper, supervised clustering of phonetic contexts is examined for both test tasks of continuous digits speech recognition and LVCSR.

For the continuous digits task, since the cross-word context of digits is less likely to contain important information, the cross-word triphone is not necessary. Also, since only 11 words are used in this task, the variety of phonetic context information is not large. Therefore, the triphone classes are clustered as intra-word triphone.

For the LVCSR task, a rule-based tentative strategy of triphone classes clustering is proposed. First, cross-word triphone classes should be applied since the cross-word context is important information in LVCSR task. Second, for there is a trend that vowels are more likely to be affected by the phonetic context, the consonants and silence are dealt as monophones, while leave only the vowels to be cross-word triphones. In order to further suppress the number of triphone classes, classes of the left or right phone in a triphone is clustered into three classes, silence, consonant and vowel. In this way, the classes of triphones are suppressed into an acceptable level.

4 Experiment

4.1 Experiment setup

Experiments of Japanese continuous digits recognition and Japanese LVCSR are conducted with this approach. The experiment conditions are shown in Table 1 and Table 2. Traditional ASR recognition system [8] is used, which generates 10-best candidates for averaged perceptron learning and reranking.

13 dimension PLP sequence is applied to form distributions of phonemes, and the second state of each phoneme is applied for calculation of edge lengths in invariant structure. The common variance is shared for all the phonemes.

4.2 Results

Word Error Rate (WER) is used to evaluate the proposed method in continuous digits task, and Fig. 4(a) and Fig. 4(b) show the results based on monophone and triphone condition respectively. Character Error Rate (CER) is used for evaluation in LVCSR, and Fig. 5(a) and Fig. 5(b)

Table 1 Experiment condition for Japanese continuous digits

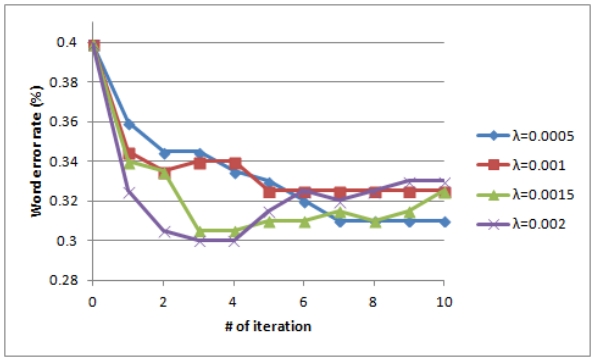
Experiment	monophone	triphone
Utterances	1 to 11 continuous Japanese digits	
Training data for HMM	27.5 hrs/667 spks/ 17316 utters	
Training data of discriminative SEM	27.5 hrs/667 spks/ 17316 utters	
Test data	1.5 hrs/100 spks/ 7382 utters	
# of HMM states	500	
# of HMM Gaussians	15000	
Language model	Unigram that outputs 10 digits (0 to 9) and the end of sentence symbol with equal probabilities	
# of phone classes (P)	18	37
# of phone pairs	171	703

Table 2 Experiment condition for Japanese LVCSR

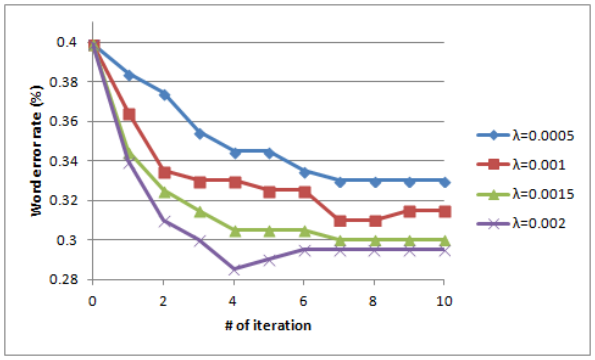
Experiment	monophone	triphone
Utterances	Japanese Dictation task	
Training data for HMM	352 hrs / 1325 spks / 196475 utters	
Training data of discriminative SEM	352 hrs / 1325 spks / 196475 utters	
Test data	1.5 hrs / 20 spks / 600 utters	
# of HMM states	5000	
# of HMM Gaussians	150000	
Language Model	Word 3-gram estimated with modified Kneser-Ney smoothing [7]	
# of phone classes (P)	57	92
# of phone pairs	1653	4278

show the results obtained on the two conditions. In this experiment, it is find out that, in the averaged perceptron algorithm, the sequence of training data also affects the performance of the model to some extent. Therefore, in the training process, training data are fixed and shared among the baseline approach and proposed approach.

As in the figures, triphone conditions provide performance improvement. λ is the

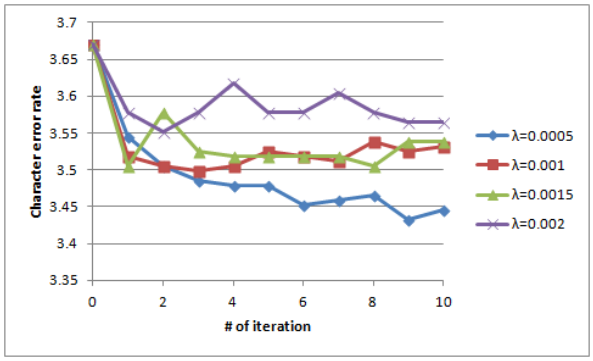


(a) monophone condition

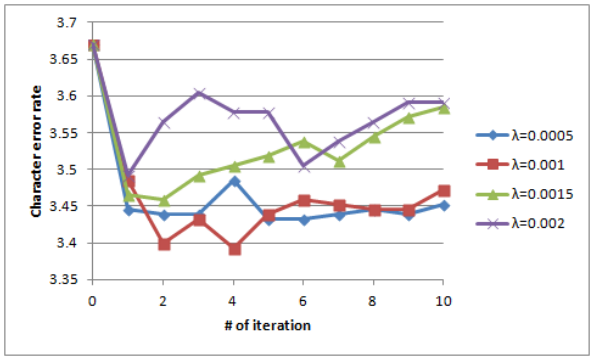


(b) intra-word triphone condition

Fig. 4 Performance of WER on Japanese continuous digits speech task



(a) monophone condition



(b) rule-based triphone condition

Fig. 5 Performance of CER on Japanese LVCSR task

learning rate in averaged perceptron. Since value of λ affects performance of learned model, performances to several values are tested. In Fig. 4(b), the best performance appears at 4th iteration when $\lambda = 0.002$. There is 4.7% WER reduction than baseline approach. In Fig. 5(b), the best performance appears at 4th iteration when $\lambda = 0.001$. There is 1.2% CER reduction than baseline approach.

5 Conclusion

In our previous study, the invariant structure was defined as f-div.-based acoustic contrast between monophones. Here in this study, it was defined as contrasts between triphones. Performance improvement is acquired both in continuous digits speech recognition task and LVCSR task. This proves that phonetic context information is able to improve effectiveness of invariant structure feature.

In LVCSR task, in order to suppress the kind of triphones used to define the invariant structure, triphone classes are tentatively clustered according to a rule based strategy. However, it is not likely to be optimal. It is promising that better performance could be acquired if more sophisticated clustering strategies are discussed.

Reference

[1] N. Minematsu, Proc. ICASSP, pp.585–588, 2004.
[2] N. Minematsu, et.al., Journal of New Generation Computing, Vol. 28, No. 3, pp.299–319, 2010.
[3] Y. Qiao, *et.al.*, Proc. INTERSPEECH, pp.3055–3058, 2009.
[4] M. Suzuki, et.al., Proc. INTERSPEECH, pp.993–996, 2011.
[5] M. Suzuki, et.al., Proc. INTERSPEECH 2012
[6] R. Schwartz, et.al., Proc. ICASSP, pp.1205-1208, 1985
[7] S.Chen, et,al., Computer Speech & Language, vol. 13, no.4, pp.359-393, 1999.
[8] S.Chen, et.al., IEEE Trans. on Speech and Audio Processing, pp.1596-1608, 2006.