LEVERAGING PHONETIC CONTEXT DEPENDENT INVARIANT STRUCTURE FOR CONTINUOUS SPEECH RECOGNITION

Congying Zhang[1,2], *Masayuki Suzuki*[2], *Gakuto Kurata*[2], *Masafumi Nishimura*[2], *Nobuaki Minematsu*[1]

[1]The University of Tokyo, Tokyo, Japan,
[2]IBM Research-Tokyo, Tokyo, Japan

ABSTRACT

Speech acoustics intrinsically vary due to linguistic and non-linguistic factors. The invariant structure extracted from a given utterance is one of the long-span acoustic representations, where acoustic variation caused by non-linguistic factors can be removed reasonably. It expresses spectral contrasts between acoustic events in an utterance. In previous studies, the invariant structure was leveraged in continuous speech recognition for reranking the N-best candidates hypothesized by a traditional automatic speech recognition (ASR) system. Use of the invariant structure features for reranking showed good effects, however, the features were defined or labeled in a phonetic-context-independent way. In this paper, use of phonetic context to define invariant structure features is examined. The proposed method is tested in two tasks of continuous digits speech recognition and large vocabulary continuous speech recognition (LVCSR). The performances are improved relatively by 4.7% and 1.2%, respectively.

Index Terms— Phonetic context, Invariant structure, Continuous digits speech recognition, LVCSR, N-best candidates reranking

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 ASR.

Recently, a method of extracting the invariant structure from an utterance was proposed, where speech acoustics are 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 was applied both to 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.

Generally speaking, it has been shown in many studies that phonetic-context-dependent models resulted in better performance in ASR than independent models. However, in our previous studies, the invariant structures were merely extracted in a phonetic-context-independent way. Therefore in this study, the effect of using phonetic context in defining the invariant structure is studied.

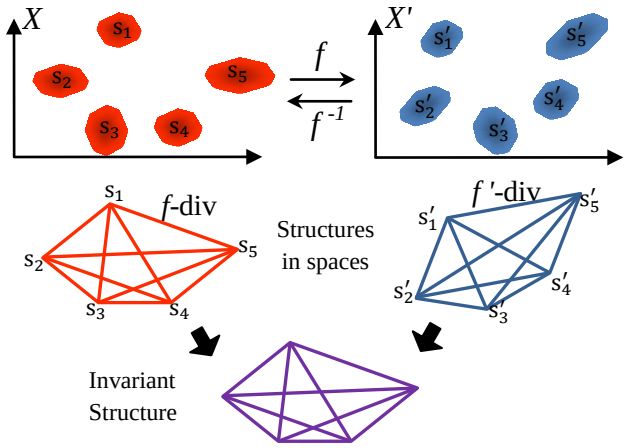


Fig. 1. Invariant structure

In this paper, the method of extracting phonetic-context-dependent invariant structure is proposed and tested in reranking the candidates hypothesized by a baseline ASR engine. In chapter 2, related works about invariant structure are explained. In chapter 3, our proposal of phonetic-context-dependent speech structure is introduced and, in chapter 4, our proposed method is examined in the tasks of continuous digits speech recognition and LVCSR and the results are discussed. Finally in chapter 5, this paper is concluded with future directions.

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 to characterize variations of the vocal tract length and shape, voices from one speaker can be converted into another speaker's. This fact indicates that if we can find any transform-invariant features, they can be used as robust features.

A necessary and sufficient condition for a feature to be invariant for any continuous and convertible transform is

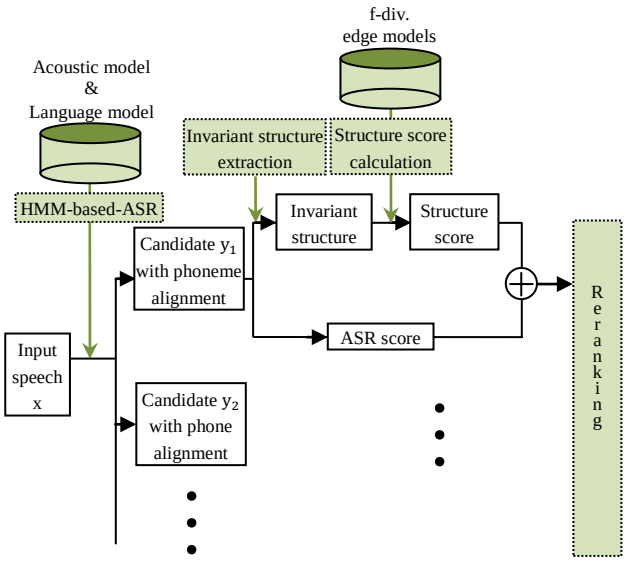


Fig. 2. Reranking framework

that the feature is represented by *f*-divergence [6]. *f*-divergence between two distributions is a family of divergences such as Bhattaryya distance and KL divergence.

Consider a feature space $\boldsymbol{X}$ and a pattern $\boldsymbol{S}$. Suppose $\boldsymbol{S}$ has 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 $f: \boldsymbol{X} \rightarrow \boldsymbol{X}'$, which transforms feature space $\boldsymbol{X}$ into a new feature space $\boldsymbol{X}'$. In this way, M events $\{s_i\}_{i=1}^{M}$ in $\boldsymbol{X}$ are mapped into $\{s_i'\}_{i=1}^{M}$ in $\boldsymbol{X}'$. Here, the *f*-divergence between events s_j and s_k $(1 \leq j < k \leq M)$ is invariant to any arbitrary invertible transform *f*. Therefore, it is equal to the *f*-divergence between s_j' and s_k'.

Fig.1 shows two invariant structures which are extracted from two utterances of the same linguistic content that were spoken by two different speakers. By calculating all the *f*-divergences between any pair of events (phonemes) in a pattern, we can obtain a structure. Each pattern consists of 5 events, so there are a total of 10 edges in each structure.

2.2. Reranking framework

In our previous study, invariant structure was leveraged in reranking the candidates hypothesized by a baseline ASR system [4][5]. The framework is shown in Fig.2. There are four steps in this process. First, a baseline ASR system generates N candidates $\{y_i\}$ from acoustic input x. Second, invariant structure features are calculated according to phoneme alignment of each candidate. Third, the structure score is calculated according to the models trained by using the invariant structure features of training data. In the final step, these candidates are reranked according to the new scores obtained by combining the structure score and the ASR score. The candidate of the highest score will be chosen as the best candidate. Through this approach, a better performance than baseline ASR system was acquired.

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.3. A variant of the averaged perceptron algorithm

2.3. Discriminative reranking with f-div. edge models

For discriminative reranking of multiple candidates generated from a speech recognizer, a *d*-dimensional feature vector $\Phi(x, y)$ is often used, where x is acoustic input and y is a specific candidate. $\Phi(x, y)$ characterizes one aspect of that candidate. For example, in LVCSR, the *i*-th dimension of $\Phi(x, y)$ is the number of word w_i $(1 \leq i \leq d$, where d is the size of the vocabulary) in y.

In our previous study [7], phoneme-to-phoneme (p-to-p) edges in candidates were applied instead of words. Here, by using training data, for each kind of p-to-p edges, its instances (f-div. values) were clustered into *N* classes. Using f-div.-dependent classes for each kind of p-to-p, the following feature $\Phi(x, y)$ was derived for acoustic input x and candidate y, where class identification was done easily by using thresholds. In equation (1), $e_{j,k}$ $(1 \leq j \leq k \leq P)$ is edge kind, which has N classes. As a result, the $\Phi(x, y)$ was formed as follows.

$$\Phi(x, y) = \begin{bmatrix} \vdots \\ \text{the number of edge } e_{j,k} \text{ class 1 in } y \\ \text{the number of edge } e_{j,k} \text{ class 2 in } y \\ \vdots \\ \text{the number of edge } e_{j,k} \text{ class } N \text{ in } y \\ \vdots \end{bmatrix} \quad (1)$$

The averaged perceptron algorithm was applied for reranking candidates discriminatively using $\Phi(x, y)$. In the algorithm (See Fig.3), weighting vector α for $\Phi(x, y)$, the *i*-th element of which can be interpreted as dgree of importance of the *i*-th element of $\Phi(x, y)$, is trained so that the lowest-WER candidate $\underline{y}$ will show higher integrated new score of $\alpha \cdot \Phi(x, y) + \phi_0(x, y)$ and vice versa. $\alpha \cdot \Phi(x, y)$ is the structure score, and $\phi_0(x, y)$ is the log likelihood caculated by a traditional ASR system. In Fig.3, x_i is the *i*-th speech in training data. And $\overline{y_i}$ and $\underline{y_i}$ are its highest-WER

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 for averaged perceptron and f-div. thresholds	the same above	
Testing data	1.5 hrs/100 spks/ 7382 utters	
# of HMM states	500	
# of HMM Gaussians	15000	
Language model	Digit-based unigram	
# 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 for averaged perceptron and f-div. thresholds	the same above	
Testing data	1.5 hrs / 20 spks / 600 utters	
# of HMM states	5000	
# of HMM Gaussians	150000	
Language Model	Word 3-gram [9]	
# of phone classes (*P*)	57	92
# of phone pairs	1653	4278

candidate and the lowest-WER candidate among the N-best candidates, respectively. t is the sequence number of iterative training. λ is a parameter of learning rate.

When the feature is applied to the LVCSR task, language model scores are also added to the feature.

3. PROPOSED APPROACH

Although our method proposed in [7] resulted in performance improvement from that of the baseline ASR system, for calculating the invariant structure, an edge between two phonemes was labeled in a phonetic-context-independent way. In the early studies of traditional ASR models, it was well known that phonetic-context-dependent models, e.g. triphone models, resulted in better recognition performance [8]. Therefore, context-dependent definition of edge labels is expected to improve the performance.

In our previous study [7], an edge was labeled by using the names of phonemes existing at the two ends. In that case, monophone labels were applied, e.g. /ah/–/m/. However, in the current study, the names of triphones are applied for defining p-to-p edges instead, e.g. /f-aa+dh/–/aa-dh+ax/. We will examine experimentally whether use of context-dependent labels will increase the effectiveness of the framework described in chapters 2.2 and 2.3.

The problem in the proposed method is that it is not practical to apply all triphone classes because their number is too huge. This will lead to data sparseness and reliable optimization of α becomes difficult. Therefore, merging phonetic context of the triphones is needed to ensure both enough information of phonetic context and also an acceptable number of triphone classes. In this paper, supervised merging of phonetic contexts is examined only in the task of LVCSR.

For the continuous digits task, as labels of triphones, we used intra-word triphones, not inter-word triphones. In this task, the vocabulary size is small and the number of kinds of triphones found in utterances is small. Merging phonetic context in defining triphones is not needed.

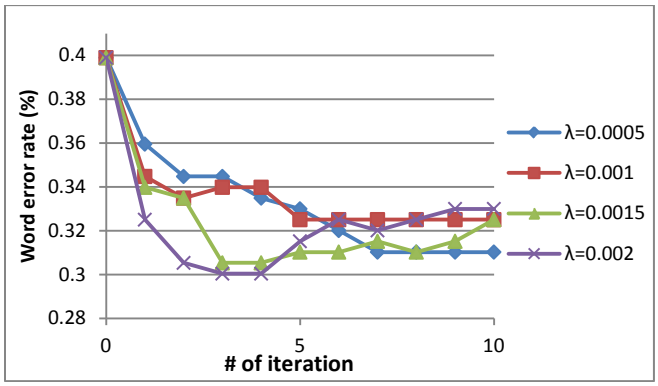
For the LVCSR task, however, since the number of kinds of triphones found in utterances is very large, context merging is required. A rule-based tentative strategy of phonetic context merging is examined. Because it is supposed that vowels are more likely to be context dependent than consonants, triphone labels are introduced only to vowels and monohpone labels are used for consonants. This means that, for consonant-to-consonant edges, their names (labels) are the same as those used in our previous study [7]. Phonetic context merging in triphones that have a vowel as central phoneme is done in the following way. The context is classified into three cases of consonants (con), silence (sil), and vowels (vow); [sil|vow|con] – vow + [sil|vow|con]. So, a monohpone of a specific type of vowel comes to have 9 triphone classes. Further, it should be noted that, different from the triphone definition used in the task of continuous digits recognition, not intra-word triphones but inter-word triphones are used because the inter-word context provides important information in the task of LVCSR.

4. EXPERIMENTS AND RESULTS

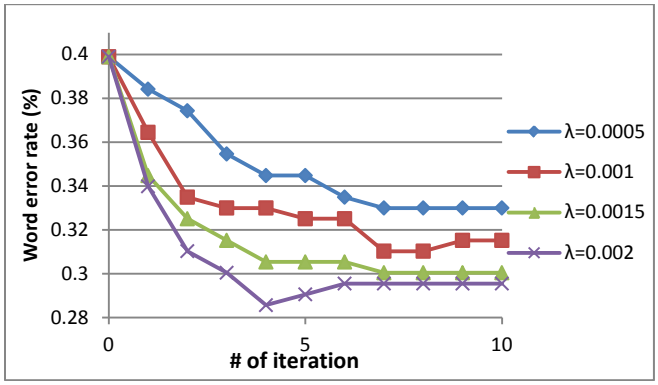
4.1. Experiment setup

Experiments of Japanese continuous digits recognition and Japanese LVCSR are conducted with the proposed method. The experimental conditions are shown in Table 1 and Table 2. In this experiment, an ASR engine, which was proposed in [10] is used as baseline recognizer, which generates 10-best candidates for averaged perceptron learning and reranking.

For extracting the structure, 13-dimensional PLP sequences were converted into distribution sequence by referring to phoneme alignment. From the frames corresponding to the central state of each phoneme HMM, their mean vector is calculated but the only and global diagonal variance matrix is shared and used for f-div. calculation.

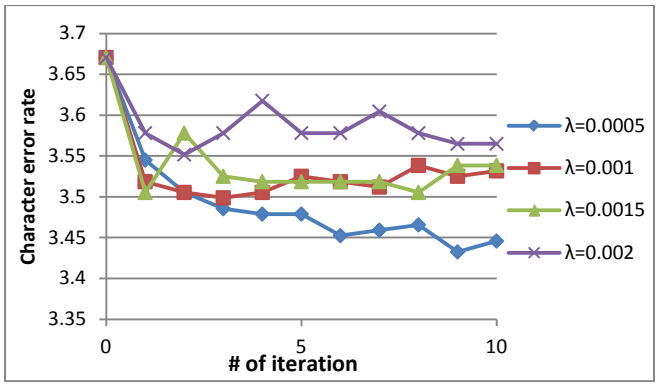


(a) monophone condition

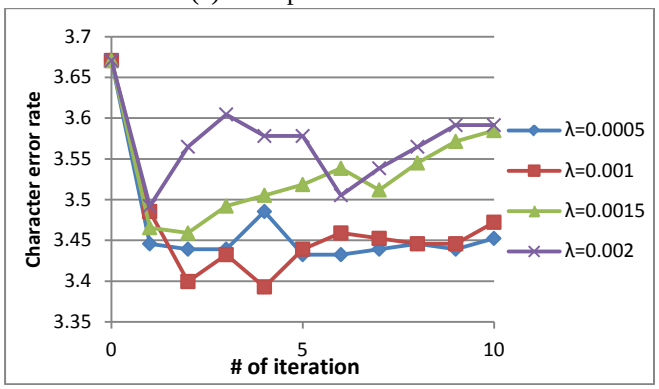


(b) intra-word triphone condition

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

4.2. Results and discussion

Word Error Rate (WER) is used to evaluate the proposed method in continuous digits task. Fig.4(a) and Fig.4(b) show experimental results when applying monophone-based and triphone-based invariant structures, respectively.

On the other hand, Character Error Rate (CER) is used for evaluation in LVCSR. Fig.5(a) and Fig.5(b) show the results obtained in the two experiments. In each of the figures, different values of learning rate λ are tested.

As in the figures, the triphone method provides performance improvement. In Fig.4(b), the best performance appears at 4th iteration when $\lambda = 0.002$. In this case, 4.7% WER reduction from the baseline performance was obtained. In Fig.5(b), the best performance appears at 4th iteration when $\lambda = 0.001$. Here, 1.2% CER reduction was found from the baseline approach.

In experimental results, recognition performance degradation occurred while λ or t reached a certain high value. This is considered to be because learning of parameter α causes overfitting. However, the triphone-based invariant structure methods performed well with relatively larger λ than the monophone based methods. It is because the larger number of edge kinds results in smaller number of samples for each kind. And this requires larger learning amount by increasing the learning rate λ. Therefore, finding appropriate λ and t is important for the optimum performance.

5. CONCLUSION

The invariant structure is extracted by using f-div.-based speech contrasts. In our previous work, each speech edge was labeled by using the monophone labels at the two ends. In this study, however, triphone labels are introduced to name the invariant speech edges. Performance improvement is acquired both in continuous digits speech recognition task and LVCSR task. This proves that phonetic context information is able to improve the effectiveness of invariant structure features.

In LVCSR task, in order to suppress the number of triphone classes used to define the invariant structure, triphone context are tentatively merged according to a rule based strategy. However, it is not likely to be optimal. We expect that more sophisticated merging will show better performance.

6. ACKNOWLEDGEMENT

The work was done when first author was a research intern at IBM Research - Tokyo.

7. REFERENCES

[1] N. Minematsu, “Yet another acoustic representation of speech sounds, ” *Proc. ICASSP*, pp.585–588, 2004.

[2] N. Minematsu, et.al., “Speech structure and its application to robust speech processing,” *Journal of New Generation Computing*, Vol. 28, No. 3, pp.299–319, 2010.

[3] Y. Qiao, et.al., “On invariant structural representation for speech recognition : theoretical validation and experimental improvement,” *Proc. INTERSPEECH*, pp.3055–3058, 2009.

[4] M. Suzuki, et.al., “Continuous digits recognition leveraging invariant structure,” *Proc. INTERSPEECH*, pp.993–996, 2011.

[5] M. Suzuki, et.al., “Discriminative reranking for LVCSR leveraging invariant structure,” *Proc. INTERSPEECH*, 2012

[6] Y. Qiao, N. Minematsu, “A study on invariance of f-divergence and its application to speech recognition,” IEEE Trans. on Signal Processing, vol.58, no.7, pp.3884-3890, 2010

[7] M. Suzuki, G. Kurata, M. Nishimura, “Discriminative edge model for invariant structure of speech, ” *IBM Research Report*, 2014

[8] R. Schwartz, et.al., “A Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech,” *Proc. ICASSP*, pp.1205-1208, 1985.

[9] S.Chen, et,al., “An empirical study of smoothing techniques for language modeling,” *Computer Speech & Language*, vol. 13, no.4, pp.359-393, 1999.

[10] S.Chen, et.al., “Efficient discriminative training of error corrective models using high-WER competitors, ” *IEEE Trans. on Speech and Audio Processing*, pp.1596-1608, 2006.