Continuous Digits Recognition Leveraging Invariant Structure

Masayuki SUZUKI[1], *Gakuto KURATA*[2], *Masafumi NISHIMURA*[2], *Nobuaki MINEMATSU*[1]

[1]The University of Tokyo, Tokyo, Japan
[2]IBM Research - Tokyo, Kanagawa, Japan
{suzuki, mine}@gavo.t.u-tokyo.ac.jp, {gakuto, nisimura}@jp.ibm.com

Abstract

Recently, an invariant structure of speech was proposed, where the inevitable acoustic variations caused by non-linguistic factors are effectively removed from speech. The invariant structure was applied to isolated word recognition and the experimental results showed good performance. However, the previous method can't apply to continuous speech recognition directly because there was no efficient decoding algorithm. In this paper, we propose a method to leverage the invariant structure in continuous digits recognition. We use a traditional HMM-based Automatic Speech Recognition (ASR) system to get N-best lists with phone alignments. Then we construct invariant structures using these phone alignments and re-rank the N-best lists by investigating which hypothesis is structurally more valid. Experimental results show a relative WER improvement of 17.4% over the baseline HMM-based ASR system.

Index Terms: Invariant Structure, Continuous Digits Recognition, N-best re-ranking

1. Introduction

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

Recently, an *invariant structure* of speech was proposed where the inevitable acoustic variations caused by nonlinguistic factors are effectively removed from speech [1]. In contrast to classical speech processing, invariant structures make use of f-divergence to model only the contrastive aspects of speech and discard the absolute features completely. This approach has been applied to isolated word recognition [2, 3], pronunciation assessment [4], and so on. The experimental results showed robustness and good performance on these tasks.

However, an invariant structure has not been used for continuous speech recognition because there was no efficient decoding algorithm. A decoding algorithm aligns a feature sequence with a hypothesis. But, to get an invariant structure, we need a phone alignment. Thus it is difficult to use invariant structures directly for decoding because of the inherent dilemma. Although Hidden Structure Model (HSM) which includes decoding algorithm was already proposed, HSM has been used only for an artificial task because it is computationally too intensive [5].

In this paper, we propose using an N-best re-ranking method [6] to leverage the invariant structure in continuous speech recognition. This method is the first one to apply the invariant structure to a real continuous speech recognition task. First we use our traditional Hidden Markov Model (HMM)-based ASR system [7] to get N-best lists with phone alignments. Next, we construct invariant structures from each N-best hypothesis. An invariant structure expresses the relationship between phones and can be obtained from a phone alignment. Then we estimate a structure-based score by investigating how valid each hypothesis is in terms of its structure. Finally, we re-rank the N-best lists by combining the ASR score and the structure-based score. This method uses a well-studied HMM-based decoding algorithm for generating N-best lists and hence can be applied to continuous speech recognition.

To confirm whether our proposed method works or not, we conducted experiments with Japanese continuous digits recognition. Experimental results show that the proposed method improved WER by 17.4% relative over our baseline HMM-based ASR system.

The rest of the paper is organized as follows. Section 2 describes the previous structure-based ASR method. Section 3 presents our proposed N-best re-ranking method to leverage the invariant structure in continuous speech recognition. Section 4 presents the experimental results in the continuous digits recognition task. Finally, Section 5 concludes the paper and describes future directions.

2. Invariant Structure

2.1. Theory of invariant structure

Voices of two speakers show different timbre because they have different vocal tract lengths and shapes. In studies of voice conversion, speaker difference is often modeled mathematically as an invertible transformation in the cepstrum domain. Especially, vocal tract length difference can be modeled well by monotonic frequency warping in the spectral domain, which can be converted into a linear transformation in the cepstrum domain. These facts indicate that some transform-invariant features can be robust features.

Consider a feature space $\boldsymbol{X}$ and a pattern $\boldsymbol{S}$ in $\boldsymbol{X}$. Suppose $\boldsymbol{S}$ can be decomposed into M events $\{s_i\}_{i=1}^{M}$. Each event is described as a distribution $s_i(\boldsymbol{x})$ in the feature space. Assume there is an invertible transformation $f : \boldsymbol{X} \rightarrow \boldsymbol{X}'$ which transforms $\boldsymbol{X}$ into a new feature space $\boldsymbol{X}'$. In this way, a pattern $\boldsymbol{S}$ in $\boldsymbol{X}$ is mapped to a pattern $\boldsymbol{S}'$ in $\boldsymbol{X}'$ and event s_i is transformed to event s_i'. Here, what we want is invariant metrics in both spaces $\boldsymbol{X}$ and $\boldsymbol{X}'$. Fig. 1 depicts an intuitive image.

f-divergence between two distributions is invariant with any kind of invertible and differentiable transform [8]. Many well known distances and divergences in statistics and information theory can be seen as special examples of f-divergence. For example, well-known Bhattacharyya distance (BD) and Kullback-Leibler divergence are kinds of f-divergence.

Fig. 1 shows two invariant structures composed only of f-divergences. With multiple events, we can obtain a structure by calculating f-divergences between any pair of them. For example, a structure composed of f-divergences between states of

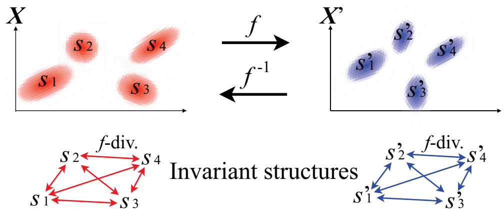

Figure 1: Invariant structures. The structure doesn't change by the invertible transformation f and f^{-1}. (f-divergence is abbreviated as f-div.)

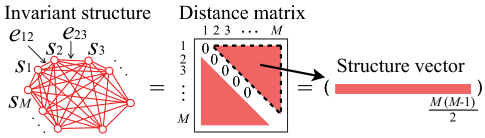

Figure 2: An invariant structure can be represented as a distance matrix and a structure vector.

speaker-dependent phone HMMs can be *speaker-independent*. Since f-divergence is invariant to any invertible transformation, the obtained structure is robust to speaker difference and any other distortions which can be expressed by an invertible transformation of the feature space (e.g. microphone difference).

2.2. Structure-based isolated word recognition

The invariant structure has been applied to isolated word recognition and the experimental results showed good performance [2, 3]. In this section, we explain how we used the invariant structure for isolated word recognition.

To begin with, we define the elements of the invariant structure (Fig. 2). An invariant structure in the left consists of M nodes. We denote individual nodes as $\{s_i\}_{i=1}^{M}$. Each node corresponds to a speech event (e.g. a phone or a state of an HMM) and is expressed as a distribution in a feature space. We denote individual edges as $\{e_{ij}\}$ where $1 \leq i \leq M$ and $i < j \leq M$. Each edge is the f-divergence between two distributions of nodes. An invariant structure can be represented as a distance matrix in the middle. If we use a symmetric distance measure such as BD, the upper-triangle elements of this distance matrix is sufficient to represent the distance matrix. We extract the upper-triangle elements as a feature vector and call it a *structure vector* in the right. The dimensions of a structure vector of M nodes is calculated as $M(M-1)/2$.

The framework of structure-based isolated word recognition is shown in Fig. 3. First of all, we need to define nodes of a structure. We use distributions of states of HMM as the nodes. The left side of the figure shows the procedure to extract a structure vector from an input utterance. First, a cepstrum vector sequence is obtained from an input isolated utterance by acoustic analysis. Then, to convert the vector sequence to a distribution sequence, a left-to-right HMM is trained with this single cepstrum vector sequence. Here, its transition probabilities are omitted. Since all the distributions have to be estimated from a single utterance, the maximum a posteriori (MAP)-based estimation is adopted. Next, we divide a distribution sequence into several sub-streams according to the dimension of cepstrum features [2]. After that, we calculate a distance matrix for each sub-stream. This method is called *multi-stream structuralization*. Geometrically speaking, this method is equivalent to decomposing the feature space into several sub-spaces and constructing an invariant structure in each sub-space. Multi-stream structuralization is effective for reducing excessive invariance in an invariant structure and for finding a rich representation which provides sufficiently discriminative information for classification [2]. Finally, all the upper-triangle elements of the multiple distance matrices are used as a structure vector. Here, a dimension of the structure vector increases by a factor of the number of sub-streams.

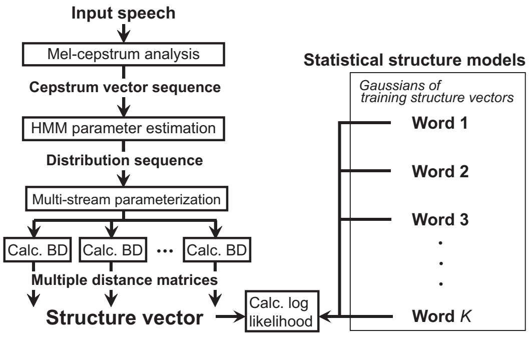

Figure 3: Framework of the structure-based isolated word recognition [2].

Right side of the Fig. 3 shows an acoustic model using the invariant structure. We call it a *statistical structure model*. We denote the number of words as K. We prepare the structure model for each of the K word independently. First, training samples of each word were converted into structure vectors. Here, we use the same topology of a structure, meaning that the nodes of a structure and the dimension of a structure are the same for all words. Then we estimate Gaussians from structure vectors of each word. These Gaussians of the structure vectors are used as statistical structure models. Similarity between an input structure vector and a statistical structure model is calculated as a log likelihood. The statistical structure model showing the maximum log likelihood is the result of recognition. There are several alternatives for statistical structure models. For example, [3] proposed a eigen-structure method based on discriminant analysis.

2.3. Limitation of the previous method

The previous method fixed the number of phones appearing in the utterance to M. In other words, only isolated words consisting of M phones were considered. In continuous speech recognition, multiple words appear in the same utterance and the previous method can't be applied. To overcome this limitation, we need some decoding algorithm. However, it is difficult to use invariant structures for decoding directly because we need phone alignments to get the invariant structure, but phone alignment will be obtained by the decoding.

A possible solution is using HSM [5]. In the framework of HSM-based ASR, a cepstrum vector sequence is first converted to a distribution sequence. HSM provides algorithms for state inference, probability calculation, and parameter estimation for distribution sequences. However, HSM has been used only for an artificial and small task because these algorithms are computationally too intensive.

There is yet another limitation. The previous method needs training samples for each word to build a statistical structure model for each word. However, preparing training samples for every word is difficult when the vocabulary size becomes huge. More generic units that are smaller than words are preferable.

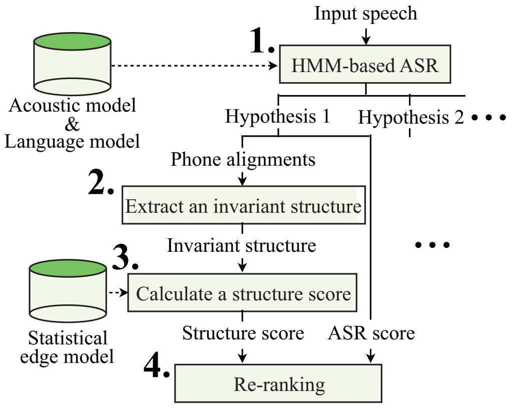

Figure 4: N-best re-ranking leveraging invariant structure

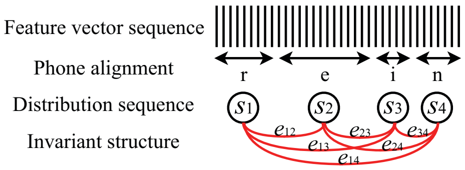

Figure 5: A procedure of extracting an invariant structure from a phone alignment. A hypothesized word is [r e i n].

3. Proposed Method

To overcome the limitation of the previous methods, we propose to leverage the invariant structure when re-ranking N-best lists generated by an HMM-based ASR system. This method is the first one to apply the invariant structure to a real continuous speech recognition task. The key idea is that we can obtain the appropriate invariant structure from the training data in advance. For a testing utterance the invariant structure obtained from the correct hypothesis should be close to the invariant structure from the training data. On the other hand, the structure from the wrong hypothesis should be distorted and different from the structure from the training data.

The framework of our proposed method is shown in Fig. 4. Note that the numbers in Fig. 4 correspond to those of the following subsections.

3.1. HMM-based ASR

We use a traditional HMM-based ASR system to get N-best lists. We can also get the ASR score and the phone alignment for each hypothesis.

3.2. Extract an invariant structure

We extract the invariant structure for each N-best hypothesis using the phone alignment. Fig. 5 shows a procedure of extracting an invariant structure from a feature vector sequence and a phone alignment. First we estimate a distribution for each phone from the feature vector sequences aligned with this phone. Then we extract an invariant structure by calculating f-divergences between each pair of distributions.

3.3. Calculate a structure score

By investigating whether an invariant structure for each hypothesis is appropriate or not, we can select a correct hypothesis from among N-best lists. Concretely, we can use the log likelihood of a statistical structure model for selecting the correct hypothesis.

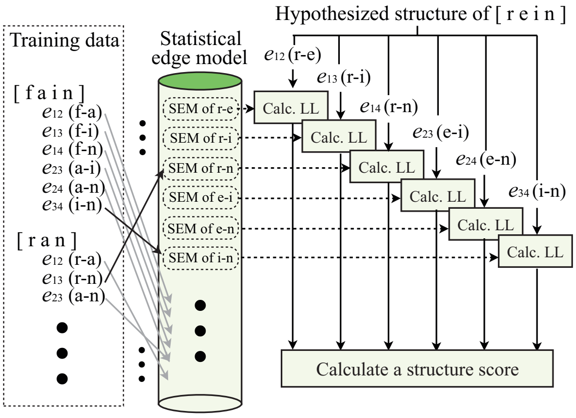

Figure 6: Process for building statistical edge models (SEMs) and process for calculating log likelihoods using SEMs. Log likelihood is abbreviated as LL.

In the previous method, a Gaussian of structure vectors for each word was used as a statistical structure model. However, we can't prepare the statistical structure models for each word if the vocabulary size becomes huge.

To solve this problem, we use edge-level models [9]. We make a statistical model for each pair of phones. We call it a *statistical edge model* (SEM). The left side of Fig. 6 shows a process of building SEMs from the training data. The right side depicts how we calculate the log likelihoods for the hypothesis of the test data using the SEMs. Using edge length data of each phone pair in the training data, we train an SEM of the phone pair as a Gaussian or a Gaussian mixture model (GMM). Suppose there are P phones, we make $P(P-1)/2$ SEMs. Once SEMs are built, we can calculate log likelihoods for each edge e_{ij} of any hypothesized structure independently.

Note that all edges in one word are modeled simultaneously in the previous isolated word recognition. In the SEM, edges are extracted from one utterance. But, each kind of edges are modeled and used independently.

The following formula shows a structure score for the n-th hypothesis h^n:

$$\text{score}_{\text{structure}}(h^n) \;=\; \frac{\sum_{i=1}^{O}\sum_{j=i+1}^{O} L_{ij}(e_{ij}^n)}{O} \qquad (1)$$

where $L_{ij}(e_{ij}^n)$ is a log likelihood of the edge obtained from the n-th hypothesis to the corresponding SEM and O is the number of phones appearing in the n-th hypothesis.

3.4. Re-ranking

We re-rank the the N-best lists combining the ASR score and the structure score with an appropriate weight. The final score of the n-th hypothesis h^n is calculated using the following formula:

$$\text{score}_{\text{proposed}}(h^n) \;=\; \text{score}_{\text{ASR}}(h^n) + w\,\text{score}_{\text{structure}}(h^n) \qquad (2)$$

where w is a weight for the structure score and determined experimentally.

This method uses a well-studied HMM-based decoding algorithm for generating N-best lists with phone alignments. Hence it can be applied to continuous speech recognition. Furthermore, our proposed method has the potential to improve the performance because an invariant structure expresses the relationship between phones in a given utterance that the HMM-based system doesn't take into consideration well.

4. Experiments

4.1. Experimental setup

We conducted experiments in continuous digits recognition in Japanese. We used our conventional HMM-based ASR system to generate N-best lists with phone alignments [7]. We trained an acoustic model (AM) of phones for continuous digits utterances. The training data consists of 27.5 hours of utterances from 667 speakers. Each utterance has 1 to 11 continuous digits. The AM contains 18 phones and these phones are represented as context-dependent, 3-state, left-to-right HMMs. The HMM states are clustered by using a phonetic decision tree. The number of states and Gaussians are 500 and 15,000. For decoding, we used a unigram language model that outputs 10 digits (0 to 9) and the end of sentence symbol with equal probabilities.

Next, we build SEMs. The training data for SEMs consists of 2.5 hours of utterances from 67 speakers, which is a subset of the data for HMM training. A unit of nodes is context-independent 18 monophones. Consequently, the number of SEMs is 18 choose 2 combinational: 136. First we conduct forced alignment on the training data and get the phone alignments. To estimate distributions of the phones, we used 13-dimensional PLP feature sequences which were aligned to the middle state of the corresponding HMM. We assume that a distribution is a Gaussian and the mean of the Gaussian is estimated in a maximum likelihood (ML) manner. For variance, we use the common variance for each phone. As we estimate distributions for the phones independently for each utterance, PLP feature vectors aligned to each phone are limited and the estimation of the variance is unstable. The common variance is estimated in an ML manner using all of the training data for each phone. Then we calculate f-divergence and extract a multi-stream invariant structure for each utterance [2]. Here, we use $\sqrt{\mathrm{BD}}$ as a f-divergence as used in [3]. For the multi-stream structuralization, we divide the 13-dimentional feature vectors into 12 multiple sub-streams as used in [3]. Consequently, the dimension of SEMs is 12. We assume that an SEM can be represented as a GMM and estimate parameters of the GMM using training data. The number of mixtures of Gaussians was set to 1, 2, 4, 8, 16.

Finally, we calculate the structure score and re-rank the N-best lists. We used test data that had at least 2 hypotheses. The total amount was 1.0 hour consisting of 95 speakers. Using an alignment for each hypothesis, we extract a structure for each hypothesis in a similar way as training data. Then we calculated a structure score using equation (1) for each hypothesis and combined the structure score with the HMM-based ASR score using equation (2). We determined the weight w using 1-person-leave-out 95-fold cross validation.

4.2. Results

Fig. 7 shows the word error rate (WER) for the test data. It also shows the baseline performance obtained by our HMM-based ASR and N-best oracle. The proposed method outperformed the baseline with all number of mixtures of Gaussians. When we use 4 mixtures, we achieved the WER of 1.17% that is relative 17.4% improvement from the baseline WER of 1.38%. Considering that the N-best oracle WER was 0.87%, our proposed method achieved 41.2% error reduction.

Theoretically speaking, the edges are invariant with regard to speaker change and hence a complexity of SEMs should be very small. However, the highest performance was achieved by 4 mixtures, not 1 mixture. Since we only use a context-independent monophone instead of the context-dependent models like a triphone, we consider that SEMs needed to model the context information.

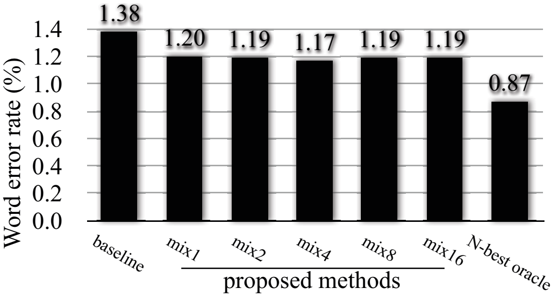

Figure 7: Word error rate

5. Conclusion

In this paper, we propose edge-level structure modeling and an N-best re-ranking method to leverage an invariant structure. The proposed method is the first one to apply an invariant structure to a real continous digits recognition task. Experimental results show that a relative WER improvement of 17.4% over a baseline ASR system was achieved.

For future work, we're going to change the HMMs to sophisticated models such as discriminatively trained and adapted models. Using the sophisticated HMM, we can get more precise phone alignments and hence more accurate invariant structures can be obtained. Since an invariant structure uses contrastive aspects of speech which aren't used in HMM-based ASR system effectively, our proposed method has potential to improve the performance of the more sophisticated HMM-based system. Additionally, we're going to apply the proposed method to Large Vocabulary Continuous Speech Recognition (LVCSR). Using a generic edge-level structure model instead of a word-level model can be of help in LVCSR.

6. 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,” *Jornal 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.*, “Integration of multilayer regression with structure-based pronunciation assessment,” *Proc. INTERSPEECH,* pp.586–589, 2010.

[5] Y. Qiao, *et.al.*, “A study of Hidden Structure Model and its application of labeling sequences,” *Proc. ASRU,* pp.118–123, 2009.

[6] Brian Roark, *et.al.*, “Corrective language modeling for large vocabulary ASR with the perceptron algorithm,” *Proc. ICASSP,* pp.749–752, 2004.

[7] S. Chen, *et.al.*, “Advances in Speech Transcription at IBM under the DARPA EARS Program,” *IEEE Transactions on Speech and Audio Processing,* 2006.

[8] Y. Qiao and 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.

[9] D. Saito, *et.al.*, “Experimental study of acoustic modeling using speaker-invariant speech contrast as modeling unit,” *Technical report of IEICE*, No.77, pp.7–12, 2019. (In Japanese)