1-P-26b

Use of Invariant And Structural Features in Discriminative Models for Isolated Word Speech Recognition

○Congying Zhang, Yousuke Ozaki, Daisuke Saito, Nobuaki Minematsu, Keikichi Hirose

(The University of Tokyo)

1 Introduction

Automatic Speech Recognition (ASR) plays an important role in human-computer interaction. For its development, MFCC-based features and HMM-based speech models have been the most popular technologies. These days, however, a new trend of ASR is found where new speech models are introduced such as DNN and discriminative models [1]. Even with these new technologies, it is well-known that acoustic mismatch between training and testing conditions easily leads to performance degradation. Differences of speakers, microphones, channels as well as background noises are typical reasons of the mismatch. In order to improve the robustness of ASR in terms of designing robust features, the structural features were proposed [2]. In our previous work, the features were mainly used in generative models of GMM. In this paper, the new features are applied in discriminative models. Experiments show that the proposed approach acquires 35.7% word error reduction.

2 Structural Feature

The structural feature was proposed in [2]. The advantage is its transform-invariant property, which is expected to lead to feature invariance against differences of speakers, microphones and etc. Basically speaking, the structural features are spectral contrasts and Fig. 1 shows the procedure of feature extraction. A sequence of feature vectors is converted into that of feature distributions. This was implemented as HMM training for an input utterance. The Bhattacharyya distance (BD) between any pair of distributions is calculated to form a distance matrix, called speech structure. Due to transform-invariance of BD, the speech structure can become a very robust feature.

In [2], the new features were used in the task of isolated word recognition. Each word in the vocabulary was statistically modeled as GMM only using structural features. Spoken word samples of simulated very tall and small speakers were used in the experiment. Fig. 2 shows the performance of structure-based word recognition. α is the warping factor that can control the vocal tract length of speakers. Negative / positive values corresponds to lengthening / shortening the length. $\alpha = 0$ means no warping.

The curve labeled with HMM is the results of word-based HMMs, while the curves marked with w is the results of the structure-based model. Here, w is a warping parameter. For details, readers should refer to [2]. It can be seen clearly that the performance of word-based HMMs is easily degraded by speakers of mismatched body height. On the other hand, that of the structure-based model is more robust among different speakers though the performance at the matched condition (α=0) is higher in the case of HMMs.

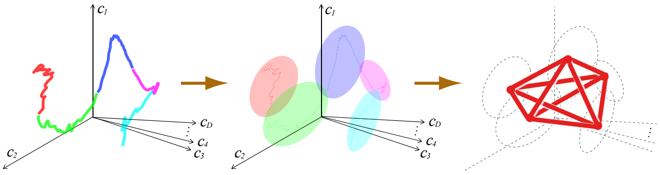

Fig.1 Extraction of structural features from an utterance

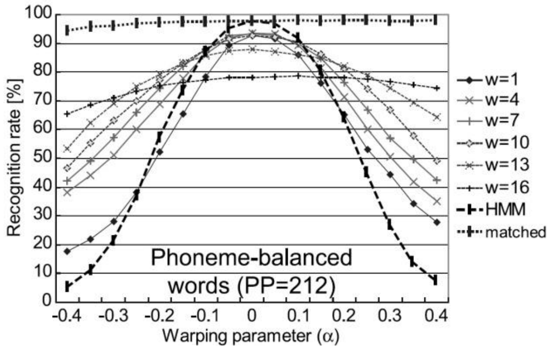

Fig.2 Performance comparison between HMM and the structure model

In Fig.1, a single feature stream gives us a speech structure. In [3], an additional technique was introduced to control the invariant property of BD adequately. A feature stream is divided into multiple sub-feature streams, where division was done by separating a full vector into partial vectors. By using multiple sub-streams, multiple structures, i.e. multiple distance matrices, can be made. This multiple structuralization was proved to be effective to improve the performance [3].

One of the drawbacks of the structure feature is high dimensionality. Since the feature is obtained as spectral contrast and the multiple structuralization can increase the number of matrices, the dimension of parameters in the speech structure tends to become very large. For this, in [2], LDA was introduced in two stages. The first LDA was carried out for each sub-stream. Dimension-reduced structures were concatenated

変換不変な構造的特徴と識別モデルを用いた孤立単語音声認識 Congying Zhang, Yousuke Ozaki, Daisuke Saito, Nobuaki Minematsu, Keikichi Hirose (The University of Tokyo) into a big structure, which again went through LDA. The resulting structural feature was eventually used in isolated word recognition [2].

3 Proposed method

In [2,3], the structural feature obtained after the two-stage LDA, which is discriminative feature transformation, was modeled statistically using a generative model of GMM. As shown in Fig.2, its performance is not satisfactory enough. Recently, use of discriminative models has become very popular and it can show better performance in various recognition tasks including visual object recognition. In this paper, we test a discriminative model of Support Vector Machine (SVM) as word recognizer with structural features as input. Use of SVM can be realized in different ways. Here, we test the following two structural features; A) without LDA and B) with the first LDA. These features are input to SVM. To realize 1-to-N classification, first, 1-to-1 classifications are done between each two classes and there will be a result for each classification. Then the final classification result will be the class which wins the most classifications. Here we used libsvm [4].

4 Experiments and results

The database used in the experiment is Matsushita isolated word speech database. It contains speech samples of 212 different words. The number of speakers is 60. Isolatedly spoken word utterances from 30 speakers are used for training, and the rest are used for testing. In this experiment, we did not apply warping of the vocal tract length, which corresponds to $\alpha=0$.

The condition of acoustic analysis is shown in Table 1. MFCC-based features are used. 12 dimensions MFCC and 1 dimension Energy are applied for structuralization. Feature division is done at the highest resolution, where the dimension of each sub-vector is 1.

The results of the experiment are shown in Table 2. Compared to the baseline method of [2], use of SVM shows better performances in either of feature A and feature B. The best performance is obtained by using feature B. Although SVM is a very powerful discriminative classifier, it seems that the dimension of structure features with no LDA is too large and the first LDA with label information is effective to be used as input feature to SVM. Quantitative comparison between the baseline method and SVM with the first LDA shows that the word error reduction rate is so high as 35.7 %.

After the experiments, we did error analysis and extracted the words that were not correctly recognized in the baseline system but were correctly recognized in SVM with the first LDA. We divided these words into mora units and calculated mora-based statistics. Then, we found the fact that 41.3 % of the morae in those words included vowel /i/. If vowel distribution among these morae is unbiased, the proportion will be 20 %. Although we cannot associate the difference of recognition mechanism between the baseline system and SVM with the first LDA to this fact, we consider that this is an interesting fact.

Table 1 Condition of acoustic analysis

Sampling frequency	16bit / 16kHz
Window length	25 ms length / 10 ms shift
Feature	MFCC (12dim) + Δ + ΔΔ + E + ΔE + ΔΔE (39 dim)
Other operation	CMN (MFCC)
For structuralization	MFCC(12dim)+E(13 dim)

Table 2 Recognition result of three approaches

(LDA+LDA)_Str+Multi-Gaussian	90.2%
Str+SVM	91.4%
LDA_Str+SVM	93.7%

5 Conclusion

For the structural feature, the proposed approach shows better performance than the existing approach. The word error rate is reduced by 35.7%. It is also interesting to find that words with certain phones are better recognized by SVM recognizer. In the future, structural feature will be calculated in phone level but not state level. Then structural feature calculated from the phone with /i/ should be considered to be given a better weight.

In [5,6], the structure feature has improved recognition accuracy by re-ranking the N-best hypotheses recognized by generative model. It is also possible to be applied in re-ranking approach of iterative decoding [7] which is to re-rank the hypotheses represented in confusion network form with less computation cost.

References

[1] Sadaoki furui, et.al., Signal Processing Magazine, IEEE, vol.29, Issue: 6, pp.16-17, 2012

[2] N. Minematsu, et.al., Journal of New Generation Computing, vol.28, no.3, pp.299-319, 2010

[3] S. Asakawa, et. al., ICASSP 2008, pp.4097-4100,2008

[4] R.-E. Fan, et.al., Journal of Machine Learning Research 6, pp.1889-1918, 2005

[5] M. Suzuki, et.al., Proc. INTERSPEECH, pp.993-996, 2011

[6] M. Suzuki, et.al., Proc. INTERSPEECH, 2012

[7] Deoras, A. et. al., ARSU 2009, pp.282-286, 2009