MFCC-DPGMM Features for Enhancing Low-Resource ASR*

◎ Bin Wu[1], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]
([1]NAIST, [2]RIKEN AIP)

1 Introduction

For years, Zerospeech challenge 2015, 2017, and 2019 have constantly confirmed the power of Dirichlet Process Gaussian Mixture Model (DPGMM) features to discriminate phonemes across different speakers, different languages under such harsh conditions as interviews with randomly interrupted disfluent dialogues, and read speech in wild or noisy recording environments.

DPGMM clustering can discriminate phonemes well because it dynamically changes the number of Gaussians until each one fits one segmental pattern of the whole speech corpus with the highest probability such that the linguistic units of different segmental patterns are clearly discriminated.

However, to the best of our knowledge, DPGMM features have not been applied to large vocabulary continuous speech recognition (LVCSR) before. Inspired by DPGMM's relatively strong discriminability, we applied it to an LVCSR system by concatenating acoustic features with DPGMM posteriorgrams such that the concatenated features combine the power of both to enhance the ASR system.

2 Method

2.1 DPGMM Clustering

We can treat DPGMM as an infinite GMM with density function $p(x_i) = \sum_{k=1}^{\infty} \pi_k p(x_i|\mu_k, \Sigma_k)$ (alternatively, $p(x_i) = \sum_{k=1}^{\infty} p(Z_i = k)p(x_i|Z_i = k)$).

This generative model samples mixture weights $\{\pi_k\}_{k=1}^{\infty}$ from the stick-breaking process (with concentration parameter α) and the means and variances $\{\mu_k, \Sigma_k\}_{k=1}^{\infty}$ from the normal-inverse-Wishart (NIW) distribution (with a belief of mean μ_0, the belief of variance Σ_0, the belief-strength of mean λ, and the belief-strength of variance ν). The generative model also samples Gaussian cluster indicator hidden variable Z_i by mixture weights and each data point X_i by the Gaussian cluster indicated by Z_i. The joint distribution of model can be described as

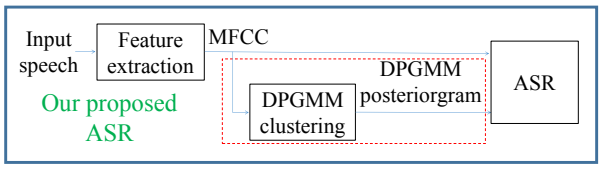

Fig. 1 Proposed features for ASR: concatenation of MFCC features and their DPGMM posteriorgrams.

$\mathrm{DPGMM}(\alpha, \mathrm{NIW}(\mu_0, \lambda, \Sigma_0, \nu))$.

Given the model definition and data $\{x_i\}_{i=1}^{n}$, we can infer from the Gibbs sampling to get posterior $p(z_i|x_i)$ and the cluster k^* of any data point x_i by $k^* = \mathrm{argmax}_k\, p(z_i = k|x_i)$.

2.2 Concatenating DPGMM Posteriorgrams with MFCC Features

Compared with a traditional ASR system which directly extracts such acoustic features as MFCC for recognition tasks, our proposal applies the DPGMM clustering algorithm on the acoustic features, gets the unsupervised DPGMM posteriorgrams and concatenates the DPGMM posteriorgrams with the MFCC features as enhanced features for the ASR system (Fig. 1).

The DPGMM posteriorgrams are of relatively high dimension, the probabilities are usually concentrated on one or two dimensions for each frame, and most of the other dimensions are zeros. MFCC is full of acoustic details in all the dimensions, but the DPGMM posteriorgram is discriminative with few dimensions; they complement each other in feature combinations. We will show that concatenating an MFCC feature and its DPGMM posteriorgram improves the ASR performance.

3 Experiment Settings

We trained an attentional encoder-decoder ASR system. We set batch size to 32 and used the Adam optimizer with an initial learning rate of 0.001. We evaluated our ASR system with a beam search where the beam size was 10.

---

*低リソース ASR の性能を向上するための MFCC-DPGMM 特徴量、
◎ビンウー [1]、 サクリア二サクチイ [1,2]、中村 哲 [1,2]
([1]NAIST, [2]RIKEN AIP)

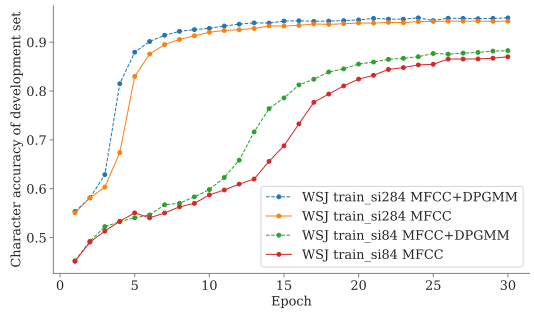

Fig. 2 Comparison between ASR systems with acoustic MFCC features (solid lines) and concatenated MFCC features and their DPGMM posteriorgrams (dashed lines) by character accuracy on the development set "dev93" of WSJ corpus.

We used python to implement the DPGMM model, whose training process used the same parameter setting as previous works [1]. We set the concentration parameter to 1 and the mean and variance of the prior to the global mean and the global variance of the MFCC features with belief-strengths 1 and $D + 2$, where D is the number of dimensions of the MFCC features. We obtained cluster labels after 1500 sampling iterations.

4 Result

We verified the effectiveness and stability of our proposed method with the spontaneous speech recognition task on the WSJ corpus. Table 1 shows that on both tasks with identical ASR system settings, we observed a more constant decrease of CER with the feature with extension (MFCC + DPGMM) than in the original feature (MFCC).

We analyzed the performance of the ASR systems during the entire training process. Fig. 2 shows that the ASR systems with feature extension by the DPGMM posteriorgram converged faster and retained improvement compared to that without the feature extension on the character accuracy of the development set ("dev93"). Our proposed feature improves more obviously on the system trained on the small dataset ("train_si84") than on the large dataset ("train_si284").

5 Conclusion

Since DPGMM features are strong at discriminating phonemes, we propose to concatenate MFCC and DPGMM to improve the ASR system. Results show that the concatenated feature works well on LVCSR, especially with fewer resources.

Table 1 We compared the attentional encoder-decoder ASR systems with or without feature extension of the DPGMM posteriorgrams, along with two baselines [3, 2], by the character error rates (CERs) on the WSJ speech corpus. No systems used pronunciation dictionaries or language models in the decoding process. We divided the WSJ corpus into the following datasets based on the Kaldi recipe: training datasets of "train_si84" (about 15 hours) or "train_si284" (about 80 hours); an identical development dataset of "dev93" and an identical evaluation dataset of "eval92" for all systems.

ASR on WSJ train_si84 (15 hrs)	**CER%**
Att Enc-Dec (Baseline ASR1) [2]	17.01
Att Enc-Dec (Baseline ASR2) [3]	17.35
Att Enc-Dec (Ours MFCC)	16.61
Att Enc-Dec (Ours MFCC+DPGMM)	14.86
ASR on WSJ train_si284 (80 hrs)	**CER%**
Att Enc-Dec (Baseline ASR1) [2]	8.17
Att Enc-Dec (Baseline ASR2) [3]	7.12
Att Enc-Dec (Ours MFCC)	6.57
Att Enc-Dec (Ours MFCC+DPGMM)	5.67

6 Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number JP17H06101.

References

[1] M. Heck, S. Sakti, and S. Nakamura. Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario. *Procedia Computer Science*, 81:73–79, 2016.

[2] S. Kim, T. Hori, and S. Watanabe. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839. IEEE, 2017.

[3] A. Tjandra, S. Sakti, and S. Nakamura. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (IEEE TASLP)*, 28:976–989, 2020.