Optimizing DPGMM Clustering in Zero-Resource Setting based on Functional Load

Bin Wu[1], *Sakriani Sakti*[1,2], *Jinsong Zhang*[3], *Satoshi Nakamura*[1,2]

[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan
[3]Beijing Language and Culture University, China

{wu.bin.vq9,ssakti,s-nakamura}@is.naist.jp, jinsong.zhang@blcu.edu.cn

Abstract

Inspired by infant language acquisition, unsupervised subword discovery of zero-resource languages has gained attention recently. The Dirichlet Process Gaussian Mixture Model (DPGMM) achieves top results evaluated by the ABX discrimination test. However, the DPGMM model is too sensitive to acoustic variation and often produces too many types of subword units and a relatively high-dimensional posteriorgram, which implies high computational cost to perform learning and inference, as well as more tendency to be overfitting.

This paper proposes applying functional load to reduce the number of sub-word units from DPGMM. We greedily merge pairs of units with the lowest functional load, causing the least information loss of the language. Results on the Xitsonga corpus with the official setting of Zerospeech 2015 show that we can reduce the number of sub-word units by more than two thirds without hurting the ABX error rate. The number of units is close to that of phonemes in human language.

Index Terms: Zero-resource speech recognition, functional load, Dirichlet process Gaussian mixture model

1. Introduction

Modern speech recognition systems rely on large amounts of human-generated annotations and language resources based on human knowledge to achieve relatively good performance. However, it is known that with almost no previous knowledge, a first-year infant can recognize sub-words and words from the human language [1]. Inspired by the study of infant language acquisition, the speech processing community set up a challenge of discovering the word and sub-word units of a language completely from scratch [1]. Building a quantitative model for this task also serves as the basis for developing a universal speech recognition system for a completely unknown language.

One of the methods to tackle this problem is to use an unsupervised clustering algorithm to recover the discrete phone-like units from speech, such as the DPGMM model, which currently achieves the top results evaluated by the ABX discrimination test: Chen et al. achieved the top results of the Zerospeech Challenge 2015 using DPGMM [2]; Heck et al. further improved the results by using feature transformations before DPGMM [3, 4] and iteratively training DPGMM-HMM acoustic unit recognizers [5].

However, the DPGMM model is too sensitive to acoustic variation; it often produces hundreds of types of sub-word units, which is too many compared to the usual number of phone classes of the phonological system. This implies there exists redundancy in DPGMM sub-word units, which may contributes little in modeling sub-word units.

Too many sub-word units also produces a posteriorgram with high dimension; the high dimensionality usually causes high computational cost to perform learning and inference, as well as more tendency to be overfitting. Decreasing the number of types of sub-word units also reduces the dimension of the posteriorgram. In [3, 4], they use Linear Discriminant Analysis (LDA) and Principle Component Analysis (PCA) to reduce the dimension of feature vector before DPGMM sampling. However, they do not necessarily reduce the number of DPGMM sub-word units and the dimension of the posteriorgram — the final representation for each frame. In addition to posteriorgram representation, we can also reduce the dimension of frame-based embedding representation using auto-encoder [6, 7].

This paper proposes to merge the sub-word units with low functional load to reduce the redundancy of the sub-word units generated by the DPGMM clustering.The key idea is that any pair of sub-word units, acoustically similar or different, that can be disambiguated by the context (their surrounding units) easily has low functional load [8]. We ignore such contrasts of units that can be easily recovered by context; the load of their communicative function actually is quite low; even if we ignore them, it will cause little information loss in speech communication. For example, in our daily speech communication, people are often lazy to listen to every phonetic unit clearly, but sometimes infer some units from their context.

Functional load is a measure of the work that two phonological units — such as two sub-word units — do in keeping the utterances apart [9, 10]. For example, in English, hundreds of word pairs differ only in /p/ and /b/ (e.g. pat vs. bat), but very few word pairs differ only in /ʃ/ and /ʒ/ (e.g. asher vs. azure). We just presume that the contrast of /p/ and /b/ does more work than that of /ʃ/ and /ʒ/ in telling complete utterances apart, implying high functional load.

We can quantify the functional load of the contrast of two phonological units by how much information they convey in speech communication using information theory [10, 8, 11]; functional load has already been applied in the study of phonological system [12, 13, 14, 15], sound change [10, 8], speech recognition [16], etc. Diachronic linguistics study shows that for human language, some contrasts of phones with high functional load remain in language evolution while contrasts with low functional load tend to disappear [9, 8]. It is also shown that merging the sub-word units with low functional load helps to improve the performance of speech recognition systems [16]. Inspired by these findings, this paper aims to ignore the contrasts of the sub-word units with low functional load, which contribute little in conveying the information of speech communication. Then we can decrease the number of types of subword units to optimize the DPGMM clustering.

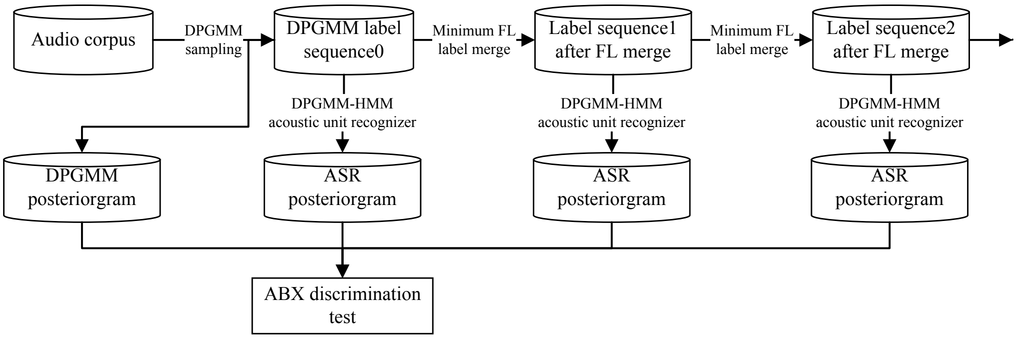


Figure 1: *System to optimize DPGMM based on functional load.*

2. Functional Load

2.1. Theory of functional load

We will use the measurement of functional load based on entropy loss [10]. Assume that our language is a sequence of labels — in this paper, as a sequence of sub-word units identified by the DPGMM — generated from a stationary and ergodic stochastic process [17]. Then we can approximate entropy H of the language L as

$$H(L) = -\frac{1}{K}\sum_{i=1}^{n} p(s_i)\log p(s_i)\,, \tag{1}$$

where s_i is any label string with length K, and n is number of different types of label strings occurring in the language.

The functional load of the contrast of label x and label y is computed by the decrease of the entropy if we ignore their difference: replacing each label x with label y in given language L.

$$FL(x,y) = \frac{H(L) - H(L_{xy})}{H(L)}\,, \tag{2}$$

where L_{xy} is the new language with label x and label y merged.

2.2. Minimum functional load based label merge

We design Algorithm 1 to compact the redundancy of the label set of a language by greedily merging the pairs of labels by the least functional load criteria similar to [16]:

```
Algorithm 1 Minimum functional load-based label merge
while number of label types is greater than threshold do
    1. Functional Load Calculation: for each merge of label pair in the language, compute its functional load with the order $K$ based on Eq. (1) and Eq. (2).
    2. Merge Decision: merge the pair of labels that leads to the least information loss with the minimum functional load.
           $(x*, y*) = \arg\min_{(x,y)} FL(x,y)$        (3)
    3. Update: renew the language label sequence by merging the optimal label pair ($x*$, $y*$) and output current label sequence of the language.
end while
```

3. Optimizing DPGMM based on Functional Load

3.1. DPGMM sampling

DPGMM can be regarded as an infinite Gaussian mixture model. Given the observations $x = x_1, \ldots, x_N$, a DPGMM can be constructed as follows:

1. Mixture weights $\pi = \{\pi_k\}_{k=1}^{\infty}$ are generated from the stick-breaking process[18].
2. The Gaussian mixture parameters $\theta = \{\theta_k\}_{k=1}^{\infty}$ are generated from a prior called Normal-Inverse-Wishart distribution [19] with the parameter $\theta_0 = (m_0, S_0, \kappa_0, \nu_0)$, where θ_k includes the mean and variance of kth mixture of Gaussian, and κ_0, ν_0 are the belief-strengths of the prior mean m_0 and the prior variance S_0 respectively.
3. Assign a label z_i to every observation x_i according to the mixture weights π.
4. Generate x_i according to z_i-th Gaussian component.

After constructing the DPGMM, we can generate the posteriorgram for each frame [2].

3.2. DPGMM-HMM acoustic unit recognizer

In this paper, we use the ASR system to get a posteriorgram that is more robust to non-linguistic factors such as speakers and channels. We follow the same procedure as [5]: build an ASR system from the DPGMM labels; use the posteriorgram of the ASR system for the ABX discrimination test. We use the typical ASR [20] that includes monophone training, triphone training followed by LDA, maximum likelihood linear transforms and speaker adaptive training.

3.3. Use of functional load

We build the system as depicted in Figure 1. First, from the raw audio corpus, we apply the DPGMM sampling to get the DPGMM cluster posteriorgram and the DPGMM label for each frame. Then, we send the DPGMM posteriorgram to the ABX discrimination test; at the same time, the DPGMM label classes are iteratively merged according to the minimum functional load criteria such that the contrasts of the labels that do little in conveying information will be ignored.

Along the way, we extract the ASR posteriorgam by building the ASR system on the label sequence; the dimension of the ASR posteriorgram will be lower and lower after doing more and more mergers of label pairs greedily based on the minimum

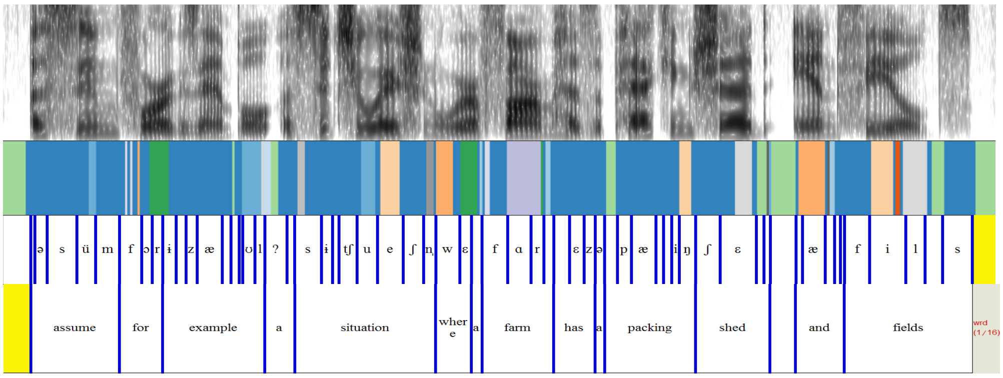


Figure 2: *Example of DPGMM clustering of sub-word units. The top layer is spectrum followed by the DPGMM label layer, phoneme layer and word layer. In the second layer, each color denotes one specific type of sub-word units.*

functional load criteria. Finally, we evaluate the ASR posteriorgram by the ABX discrimination test.

4. Experiment setup

4.1. Zero-resource speech data

We conduct all our experiments on the official data set of Interspeech Zero Resource Speech Challenge 2015. We use the Xitsonga corpus, which is an excerpt the NCHLT corpus of South African read speech. The evaluation is conducted on the official segmentation of the audio corpus with a length of 2 h 29 min.

4.2. DPGMM label extraction

The DPGMM sampling experiment is done by toolkit [21]. The setting of parameters we use is the same as [2, 3]. We use the 39-dimensional MFCC+Δ+$\Delta\Delta$ features with a window size of 25ms and a window shift of 10ms, which are followed by the mean and variance normalization (MVN) and vocal tract length normalization (VTLN).

The DPGMM sampling is stopped after 1500 iterations with the parameter setting the same as [2, 3]: the concentration parameter $\alpha = 1$; the prior of the mean m_0 and the variance S_0 as the global mean and the global variance with belief-strengths $\kappa_0 = 0$ and $\mu_0 = D + 3$ respectively, where D is dimension of MFCC features (D = 39) .

4.3. DPGMM-HMM acoustic unit recognizer

We use the Kaldi toolkit [20] and follow the standard recipe of the TIMIT corpus [22] except for three modifications:

1. We use the 1-state HMM instead of the 3-state HMM because lots of DPGMM labels exist with only a few frames; 3-state HMM fails to capture the temporal variation of DPGMM labels.
2. We don't make a model for silence. In our experiment on the TIMIT corpus, we found that DPGMM clustering does well in recognizing the silence part and has already assigned some labels for the silence (Figure 2).
3. We use language model with higher order. As the DPGMM algorithm is sensitive to the acoustics, it generates many of its sub-word units with shorter duration than usual phones of human language. The default bigram of the TIMIT recipe is a bit short for modeling the context of DPGMM labels. In our experiment, we use the 4-gram for the language model as [5], which will improve the performance of the ABX test as we have better context.

4.4. Functional load computation

For the computation of functional load, we set the length K of the sub-word unit string in equation (1) as 3.

4.5. ABX discrimination test

Suppose that we want to get the ABX error rate [23] of two categories with distribution $\mathbb{P}$ and distribution $\mathbb{Q}$ given some distance measure d. First, we randomly draw two observations a, x from distribution $\mathbb{P}$ and an observation b from distribution $\mathbb{Q}$; then we say we make an error of distinguishing these two categories if the distance of two observations (d_{ax}) from the same category is larger than the distance of two observations (d_{bx}) from different categories. Thus, the ABX error rate is defined as

$$e(\mathbb{P},\mathbb{Q}) = \mathop{\mathbb{E}}_{a\sim\mathbb{P},b\sim\mathbb{Q},x\sim\mathbb{P}} (I(d_{ax} \geq d_{bx}) - \frac{1}{2} I(d_{ax} = d_{bx})), \tag{4}$$

where $\mathbb{E}$ is the expectation, and I is the indicator function.

We use the ABX toolkit [24] for experiments. The ABX error rate is computed on the categories of the minimum pairs of triphones with distance measure as KL-divergence, as suggested by [1] when evaluating the posteriorgram representation.

5. Experiment Result and Discussion

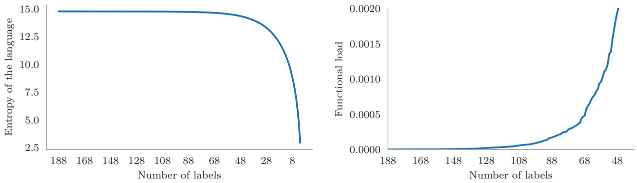
5.1. Analysis of DPGMM clustering

To analyze the sub-word units generated by DPGMM clustering, we want to compare it with the true phones in the language. As Xitsonga has no annotation of time information of each phone, we do some preliminary experiments by running the DPGMM algorithm on the training set (3.14 hours with clean read speech) of the TIMIT corpus [22].

Figure 2 shows that DPGMM does well in discovering segments of silence; the fricative *s* and the fricative *f* are quite fragmentary — one phone corresponding to several different short DPGMM sub-word units — because the fricatives have high frequency; the vowel *i* in the word *field* is fragmentary because of the sharp change of the formants.

We conclude that DPGMM clustering is sensitive to acoustic variation such as high frequency and change of formats, while if there is no change of acoustics (e.g. the silence segments), DPGMM does well in recognition of sub-word units.

5.2. Analysis of functional load merger



(a) *Entropy of language* (b) *Functional load of label pair*

Figure 3: *Entropy of language and functional load of label pair merged after each iteration*

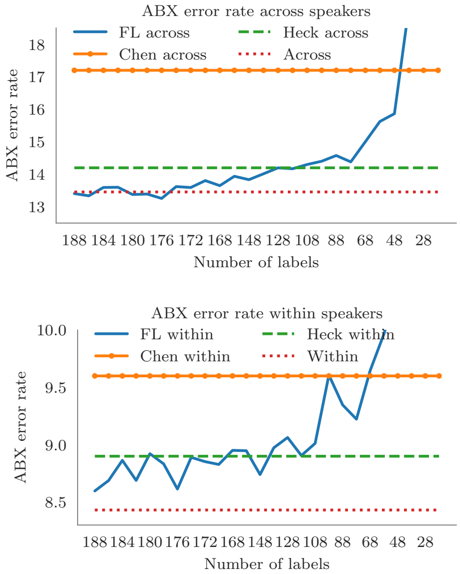


Figure 4: *ABX error rate across and within speakers. Chen within, Heck within and Within are results from [2], Heck et al. [3] and this paper using the same setup of parameters. The blue line is ABX error rate after each iteration of functional load merge.*

Figure 3a shows that as we merge more and more label pairs, the entropy of the language decreases monotonically. This is because it will cause more damage to the information transmission of the language as we can't distinguish more and more label pairs. Figure 3b shows that the functional load at the beginning of mergers is very small. In our experiments, we find that the functional load of the first 17 pairs of the labels is zero.

Merging pairs with zero functional load does not cause any information loss of the language. In [8], it can be proved that the functional load of a label pair is zero if and only if the label pair is in complementary distribution. That means each of the first 17 label pairs has completely a different context. Even if the pair of labels is merged, we can still distinguish them from their context —surrounding labels.

5.3. Evaluation by ABX discrimination test

Table 1: *ABX error rate from [2], [3] and this paper. Paper [2] achieved the top results in Zerospeech 2015; paper [3] improved the performance of [2]. FLm: result after m iterations of functional load merge of DPGMM label pairs*

Existing systems	**Num. of Labels**	**Within Speakers**	**Across Speakers**
DPGMM (c) [2]	321	9.6	17.2
DPGMM (h) [3]	192	8.9	14.2
DPGMM + PCA (h) [3]	239	9.8	16.4
Proposed system			
DPGMM + FL0	188	8.4	13.4
DPGMM + FL12	176	8.6	13.2
DPGMM + FL70	118	8.9	14.2
DPGMM + FL120	68	9.6	15.0

For comparison, we use the same parameter setting as [2, 3] for DPGMM sampling; we use officially provided voice activity detection segmentations (2.5 hours) like [3] while [2] used all Xitsonga corpus (6.52 hours) for training DPGMM; we use the same official data to do the ABX error rate evaluation as [2, 3].

Figure 4 shows that the ABX error rate of the ASR posteriorgram of the first 20 pairs of label mergers is relatively stable. Actually, our experiments show that merging the first 17 pairs of labels doesn't change the entropy of the language.

Table 1 and Figure 4 shows that if we merge about half of the labels $(188 \rightarrow 118)$, we can get a similar ABX error rate to Heck's [3]; if we merge about two thirds of the size of labels $(188 \rightarrow 68)$, we can get a similar ABX error rate to Chen's [2]. This implies that by merging the labels with low functional load, we can reduce the size of the DPGMM labels without hurting much of the performance in the ABX test. In Table 1, the result [3] of applying PCA on MFCC features stacking context is also listed.

6. Conclusions

In this paper, we reduced the number of DPGMM sub-word acoustic units by merging units with the least information loss of the language: the minimum functional load. Even if we lose the contrasts of these units, they can be recovered from the surrounding context easily, as indicated by their low functional load. Results show that we can reduce the number of sub-word units by more than two thirds without hurting the ABX error rate. The number of units is close to that of phonemes in human language.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237. We thank our colleague Michael Heck for introducing topics related to DPGMM and ABX test.

8. References

[1] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, “Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[4] ——, “Supervised Learning of acoustic models in a Zero Resource Setting to Improve DPGMM clustering,” in *INTERSPEECH*, 2016, pp. 1310–1314.

[5] ——, “Iterative training of a DPGMM-HMM acoustic unit recognizer in a zero resource scenario,” in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 57–63.

[6] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, “Unsupervised neural network based feature extraction using weak top-down constraints,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5818–5822.

[7] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, “A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] W.-Y. Wang, “The measurement of functional load,” *Phonetica*, vol. 16, no. 1, pp. 36–54, 1967.

[9] M. André, “Economie des changements phonétiques,” *Berne: Francke*, 1955.

[10] C. F. Hockett, *A manual of phonology*. Waverly Press, 1955, no. 11.

[11] J. Zhang, W. Li, Y. Hou, W. Cao, and Z. Xiong, “A study on functional loads of phonetic contrasts under context based on mutual information of Chinese text and phonemes,” in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010, pp. 194–198.

[12] Y. M. Oh, F. Pellegrino, C. Coupé, and E. Marsico, “Cross-language comparison of functional load for vowels, consonants, and tones.” in *Interspeech*, 2013, pp. 3032–3036.

[13] D. Surendran and P. Niyogi, “Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals,” *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, vol. 279, p. 43, 2006.

[14] B. Wu, J. Zhang, and Y. Xie, “A clustering analysis of Chinese consonants based on functional load,” in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–4.

[15] Y. Chen, Y. Xie, and J. Zhang, “A comparison study of information contributions of phonemic contrasts in Mandarin,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 1579–1582.

[16] J.-S. Zhang, X.-H. Hu, and S. Nakamura, “Using mutual information criterion to design an efficient phoneme set for Chinese speech recognition,” *IEICE TRANSACTIONS on Information and Systems*, vol. 91, no. 3, pp. 508–513, 2008.

[17] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[18] J. Sethuraman, “A constructive definition of Dirichlet priors,” *Statistica sinica*, pp. 639–650, 1994.

[19] M. West and J. Harrison, *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[21] J. Chang and J. W. Fisher III, “Parallel sampling of DP mixture models using sub-cluster splits,” in *Advances in Neural Information Processing Systems*, 2013, pp. 620–628.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, 1993.

[23] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, “Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline,” in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1–5.

[24] T. Schatz, R. Thiolliere, E. Dupoux, G. Synnaeve, and E. Dunbar, “Abxpy v0. 1,” 2015.