Using Functional Load for Optimizing DPGMM based Zero Resource Sub-word Unit Discovery

Bin Wu[1,a)] Sakriani Sakti[1,2,b)] Jinsong Zhang[3,c)] Satoshi Nakamura[1,2,d)]

Abstract: Unsupervised sub-word discovery of the zero resource language gains attention recently. One of the methods to tackle the problem is using an unsupervised clustering algorithm to recover the discrete phone-like units from the speech, such as the Dirichlet Process Gaussian Mixture Model (DPGMM), which currently achieves top results in the Zero Resource Speech Challenge. However, the DPGMM model is too sensitive to the acoustic variation and often produces too many types of sub-word units. This paper proposes to apply functional load to reduce the size of sub-word units from DPGMM. The functional load is the measurement of how much information in communication is conveyed by contrasts of these units. Then, the aim is to ignore the contrasts of the sub-word units that contribute little in conveying the information of the speech leading to decrease of the number of sub-word classes. We experiment on the official Zerospeech 2015 measuring with ABX error rate.

1. Introduction

Given the audio corpus merely, can we find the phoneme-like sub-word units? DPGMM finds these clusters and achieves top results [1], [2] in the Zero Resource Speech Challenge [5]. However, it also creates hundreds of sub-word units because of its sensitivity to acoustic details. In our daily speech communication, people are often lazy to listen to every acoustic detail clearly, but sometimes infer some units from their context. The idea of these paper is to ignore the constrasts of the DPGMM sub-word units that can be disambiguated by their context easily — merging the acoustic units with low importance in communication (a.k.a. low functional load [3])

2. Functional Load

2.1 Theory of Functional Load

We will use the measurement of functional load based on entropy loss [3]. 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 [4]. 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 [6]:

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, compute its functional load 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.

$$(x^*, y^*) = \arg\min_{(x,y)} FL(x,y) \tag{3}$$

(3) **Update:** renew label sequence by merging the optimal label pair (x^*, y^*) and output current label sequence.

end while

2.3 Use of Functional Load

We use DPGMM sampling to get the label sequence and merge label pairs with greedy Algorithm 1 iteratively. For evaluation, ABX test is used on ASR posteriorgram of the label sequence at each iteration. (the ASR posteriorgram is more robust to non-linguistic factors such as speakers and channels)

3. Experiment Setup

3.1 Zero-resource Speech Data

We use the Xitsonga corpus (South African read speech). The

[1] Nara Institute of Science and Technology, Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan
[3] Beijing Language and Culture University, China
[a)] wu.bin.vq9@is.naist.jp
[b)] ssakti@is.naist.jp
[c)] jinsong.zhang@blcu.edu.cn
[d)] s-nakamura@is.naist.jp

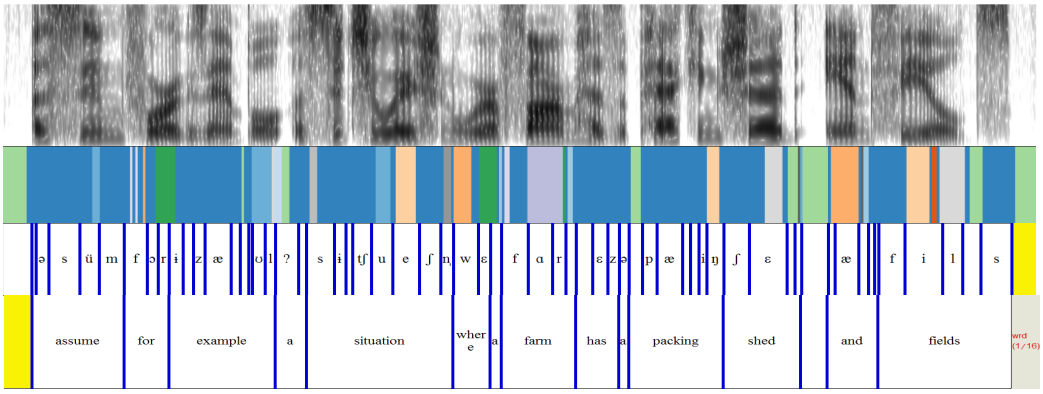

Fig. 1 Example of DPGMM clustering of sub-word units. The top layer is spectrum followed by the DPGMM label layer, phoneme layer and word layer. In the second layer, each color denotes one specific type of sub-word units.

evaluation is conducted on official audio segmentation (about 2 h 29 min) of Interspeech Zero Resource Speech Challenge 2015 .

3.2 Setup of Experiment

We follow [1], [2] to set parameters for DPGMM sampling and ASR posteriorgram extraction. We use the 39-dimensional MFCC+Δ+ΔΔ features with MVN and VTLN.The DPGMM sampling is stopped after 1500 iterations. For functional load, we set the length K of sub-word unit string in equation (1) as 3.

4. Experiment Result and Discussion

4.1 Analysis of DPGMM Clustering

We analyze the sub-word units generated by DPGMM clustering on the training set (3.14 hours) of the TIMIT corpus.

Figure 1 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.

4.2 Evaluation by ABX Discrimination Test

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

Existing systems	**Num. of Labels**	**Within Speakers**	**Across Speakers**
DPGMM (c) [1]	321	9.6	17.2
DPGMM (h) [2]	192	8.9	14.2
DPGMM + PCA (h) [2]	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 follow the same parameter setting of [1], [2] for DPGMM sampling and evaluate on the same official data with ABX error rate.

Table 1 shows that if we merge about half of the labels (188 → 118), we can get a similar ABX error rate to Heck's [2]; if we merge about two thirds of the size of labels (188 → 68), we can get a similar ABX error rate to Chen's [1]. 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 [2] of applying PCA on MFCC features stacking context is also listed.

5. Conclusion

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.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] Chen, H., Leung, C.-C., Xie, L., Ma, B. and Li, H.: Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study, *Sixteenth Annual Conference of the International Speech Communication Association* (2015).
[2] Heck, M., Sakti, S. and Nakamura, S.: Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario, *Procedia Computer Science*, Vol. 81, pp. 73–79 (2016).
[3] Hockett, C. F.: *A manual of phonology*, No. 11, Waverly Press (1955).
[4] Shannon, C. E.: A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol. 5, No. 1, pp. 3–55 (2001).
[5] Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A. and Dupoux, E.: The zero resource speech challenge 2015, *Sixteenth Annual Conference of the International Speech Communication Association* (2015).
[6] Zhang, J.-S., Hu, X.-H. and Nakamura, S.: 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).