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

Bin Wu[1] , Sakriani Sakti[1,2] , Jinsong Zhang[3] and Satoshi Nakamura[1,2]
{wu.bin.vq9,ssakti,s-nakamura}@is.naist.jp, jinsong.zhang@blcu.edu.cn

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

Background

Unsupervised subword unit discovery

- Modern ASR system depends on rich resources:
  - annotated training corpora
  - carefully designed dictionary
  - high-order language model
- However, for 7000 living languages, most of them are low resources:
  - lack of expert knowledge of phonemic system
  - some not have written forms
- Unsupervised subword unit discovery of Zerospeech was proposed (Park, 2008; Versteegh, 2015)

Previous methods

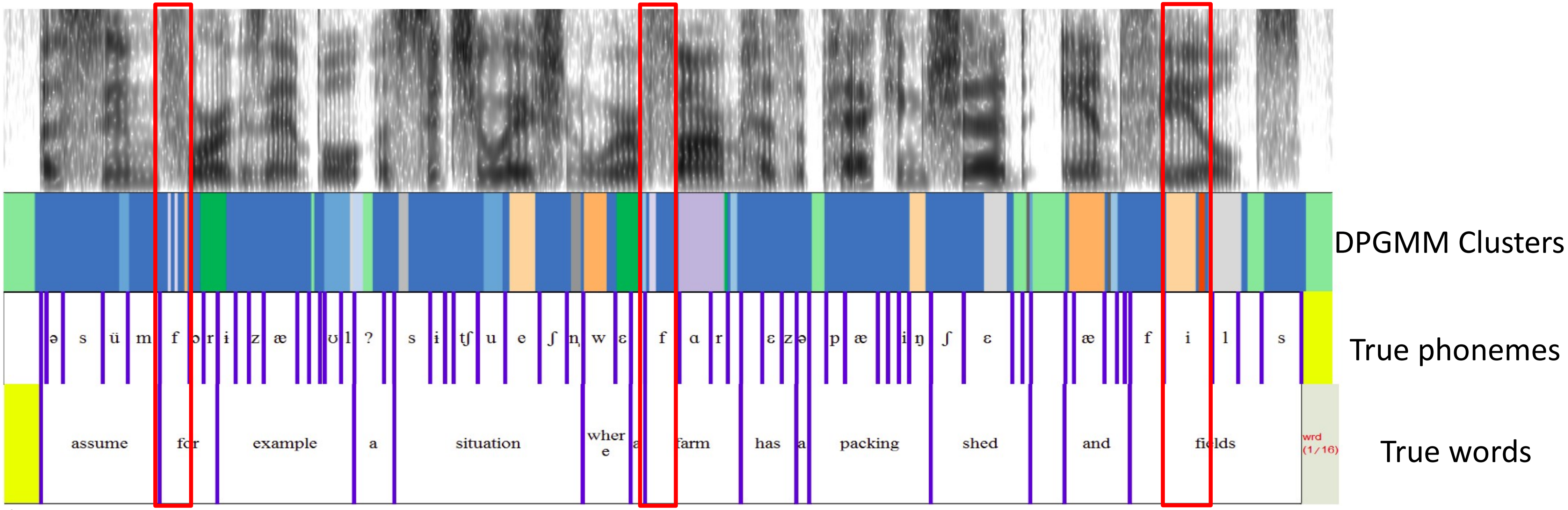
- Unsupervised subword unit discovery of Zerospeech
  - Early works: DTW (Park, 2008); HMM-split (Varadarajan, 2008)
  - DNN based method:
    - Spoken term detection + autoencoder(Badino 2014, Kamper, 2015; Pitt, 2015)
    - Spoken term detection + ABnet (Synnaeve 2014, Thiolliere, 2015)
  - Nonparameteric Bayesian based methods
    - Variational autoencoders (Ondel, 2016; Ebber, 2017)
    - Dirichlet Process Gaussian Mixture Model (**DPGMM Clustering**) (Lee, 2012; Chen, 2015; Heck, 2016)

DPGMM Clustering

- DPGMM clustering method gets relative good performance:
  - top results of the Zerospeech Challenge 2015 (Chen, 2015)
    → improved by feature transformations + iterative ASR optimization (Heck, 2016)
  - top results of the Zerospeech Challenge 2017 (Heck, 2017)

- Problem of DPGMM clustering:
  - # of sub-word clusters > # of phonemes of usual languages
    - e.g. # of sub-word clusters - 321 (Chen, 2015); 192 (Heck, 2016)
  - posteriorgram with high dimension (high computational cost; overfitting)

Problem



DPGMM Clusters
True phonemes
True words
assume
for
example
a
situation
where
farm
has
a
packing
shed
and
fields

Proposal

Basic idea

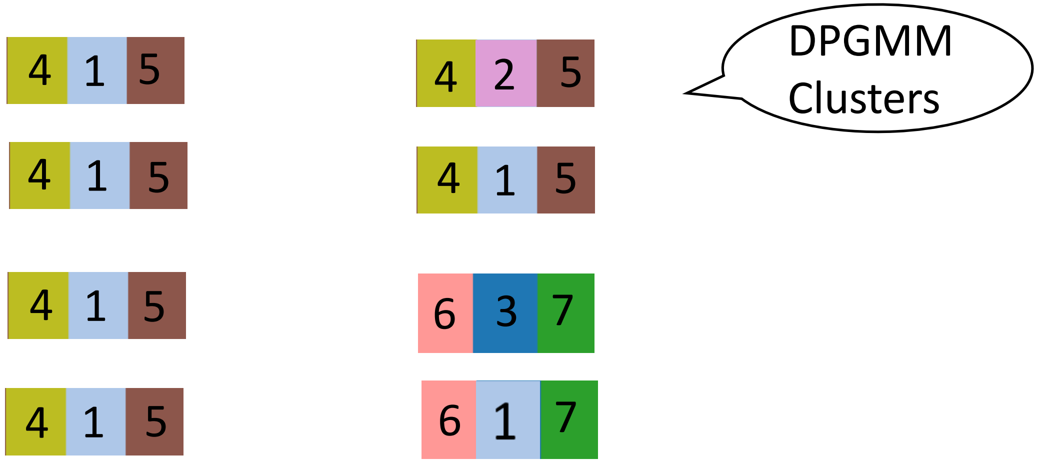
- Merge the pairs of sub-word units **with different context**

Minimal pairs

Merge Minimal pairs

Complementary distribution

Merge Complementary distribution



Functional load

- Any pair of sub-word units that can be disambiguated by the context (their surrounding units) easily has low functional load. (Wang, 1967)
- Functional load of a contrast
  - Definition: the importance of the contrast in speech communication
  - Computation: on loss of entropy (Hockett, 1955)
    - e.g. functional load of a contrast of a sub-word unit pair x and y

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

- Proposal: merge the sub-word units with low functional load to reduce the redundancy of the DPGMM sub-word units

Training framework

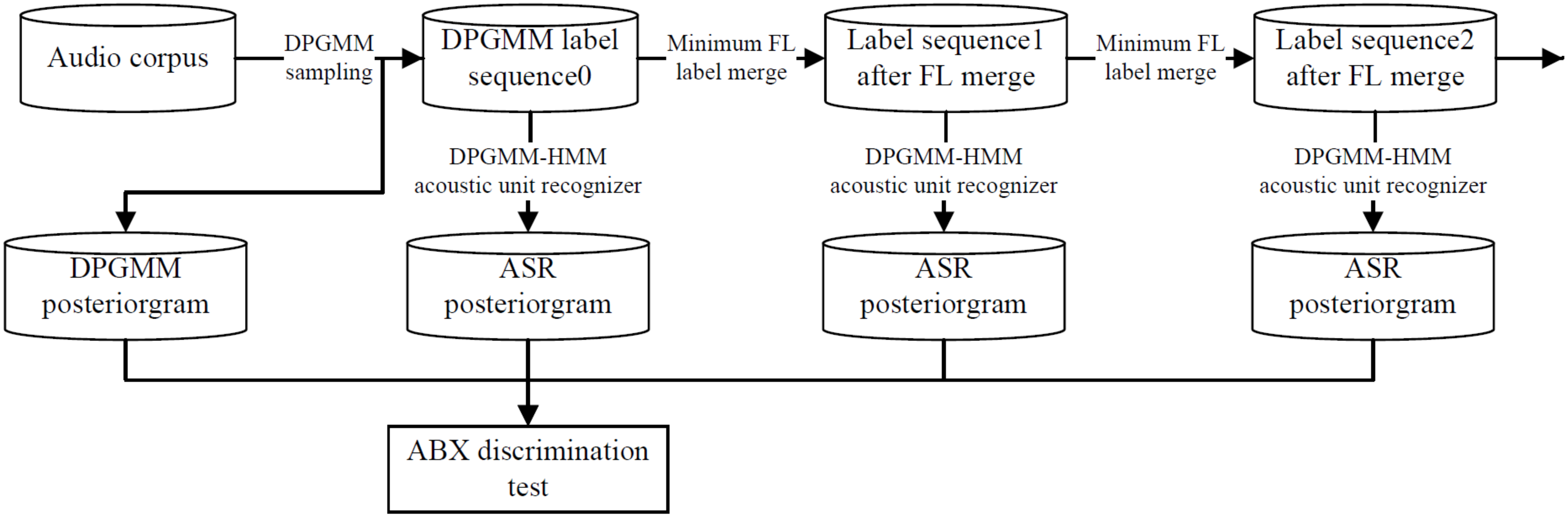


Figure 1: *System to optimize DPGMM based on functional load.*

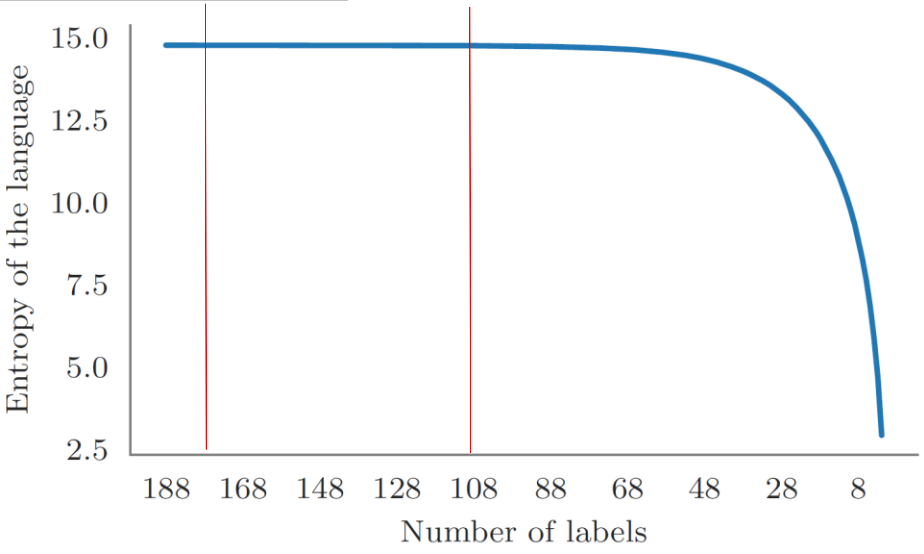
Experiment & Result

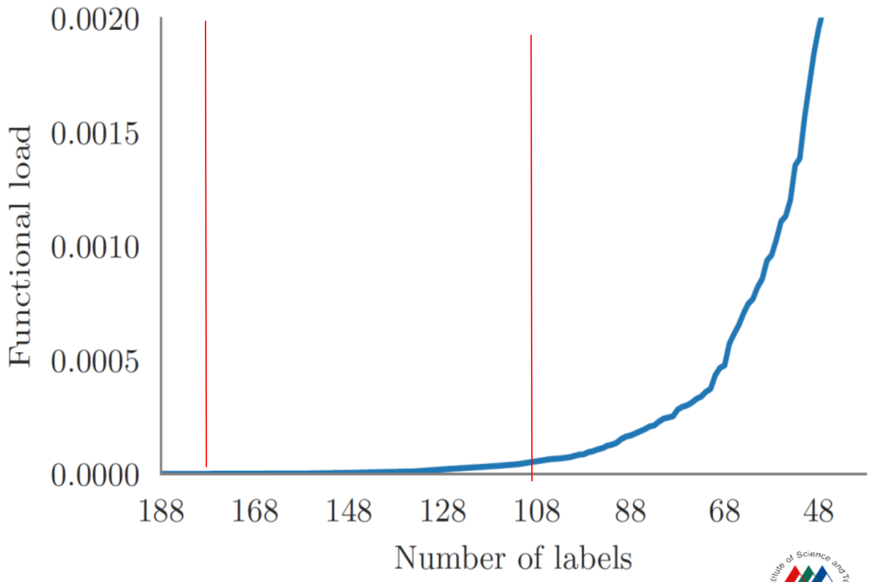
Corpus

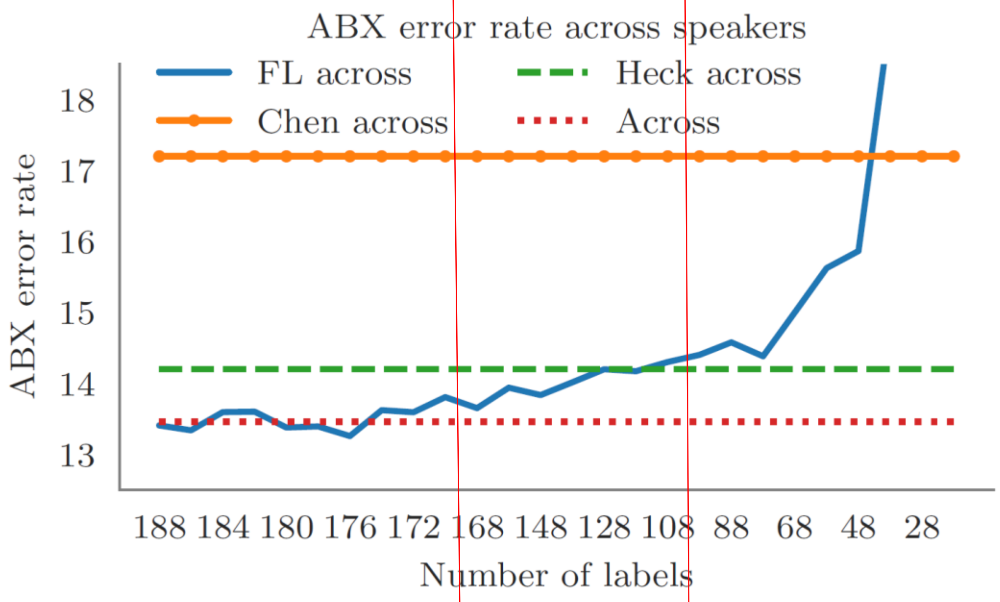
- Xitsonga corpus
  - an excerpt the NCHLT corpus of South African read speech
  - with the official segmentation of Interspeech Zero Resource Speech Challenge 2015
  - length: 2 h 29 min

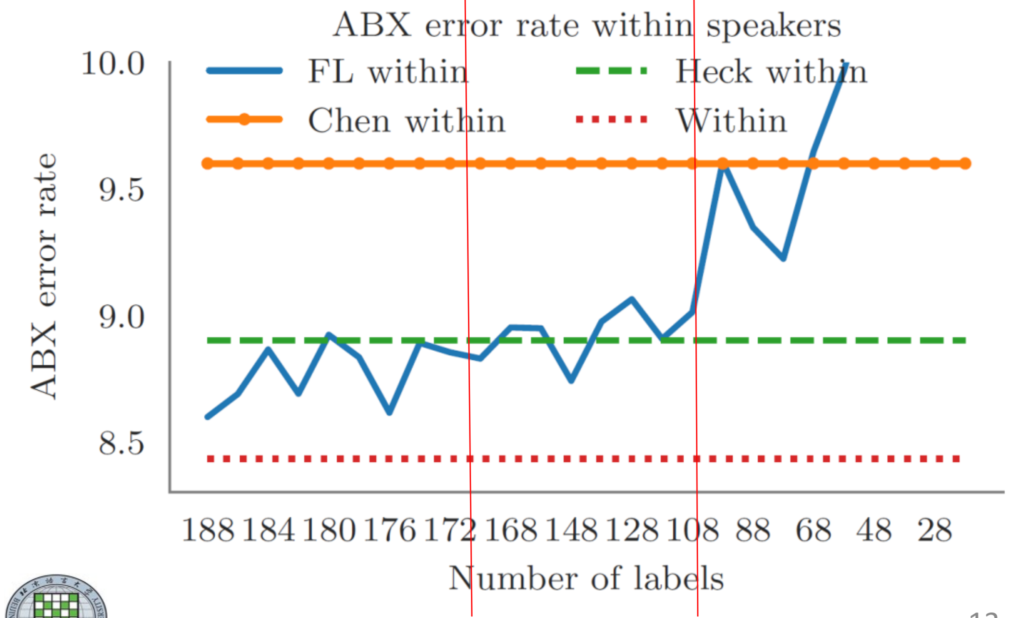
Between 188 and 171:
FL(the first 17 pairs) = 0
No information loss

After 108:
FL(contrast of label pairs)
starts to increase quickly









Results

Table 1: ABX error rate from Chen, Heck and this paper
(FLm: result after m iterations of functional load merge of DPGMM label pairs)

Existing systems	**Number of labels**	**Within speaker**	**Across speaker**
DPGMM (Chen, 2015)	321	9.6	17.2
DPGMM (Heck, 2016)	192	8.9	14.2
DPGMM + PCA (Heck, 2016)	239	9.8	16.4

Conclusion

- We merge DPGMM sub-word units greedily with low functional load.
- We reduce the number of sub-word units by more than two thirds and still get relatively good ABX error rate.
- The number of remaining units is close to that of phonemes in human language.

Thank you for listening!