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

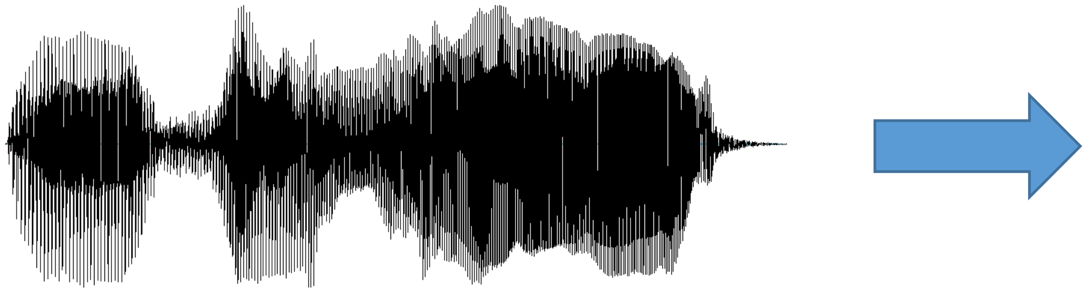
Bin Wu[1] , Sakriani Sakti[1,2] , Jinsong Zhang[3] and Satoshi Nakamura[1,2]

{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

Research Question



o h a y o o sil

- How to find phoneme-like units from zero-resource speech?

line-girl1-ohayou1

Why important

- Problem: zero-resource phoneme-like unit discovery
- Why the problem important?
    - State-of-art DNN needs labels (phonemes,…)
        - manual labelling needs money and effort
        - Knowledge of the labels (phonological system, …)
- Zero-resource technology helps to create these labels (phonemes, …)

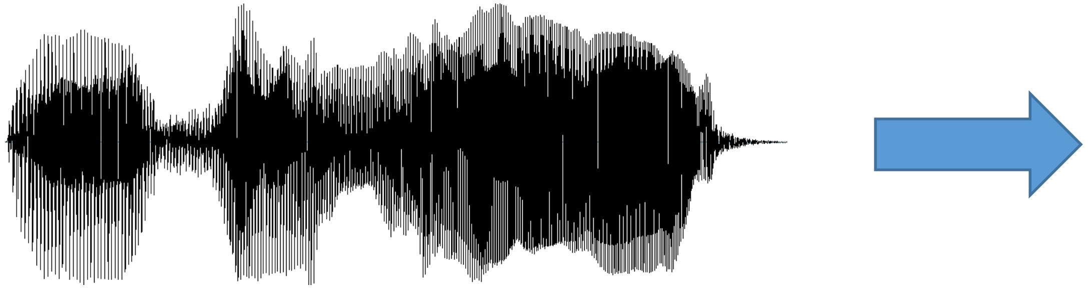
Previous methods

- Unsupervised sub-word unit discovery of Zerospeech
  - **Pre-trained labels + DNN**
    - spoken term detection + autoencoder *[Badino 2014, Kamper, 2015; Pitt, 2015]*
    - spoken term detection + ABNet *[Synnaeve 2014, Thiolliere, 2015]*
  - **Unsupervised clustering**
    - Variational autoencoders [Ondel, 2016; Ebber, 2017]
    - Dirichlet Process Gaussian Mixture Model (**DPGMM** Clustering) *[lee, 2012; Chen, 2015]*
      - DPGMM + ASR feature transformations *[Heck, 2016]*
      - DPGMM + ASR alignment *[Heck, 2017]*
- DPGMM clustering gets top results of the Zerospeech Challenge 2015, 2017

Problem

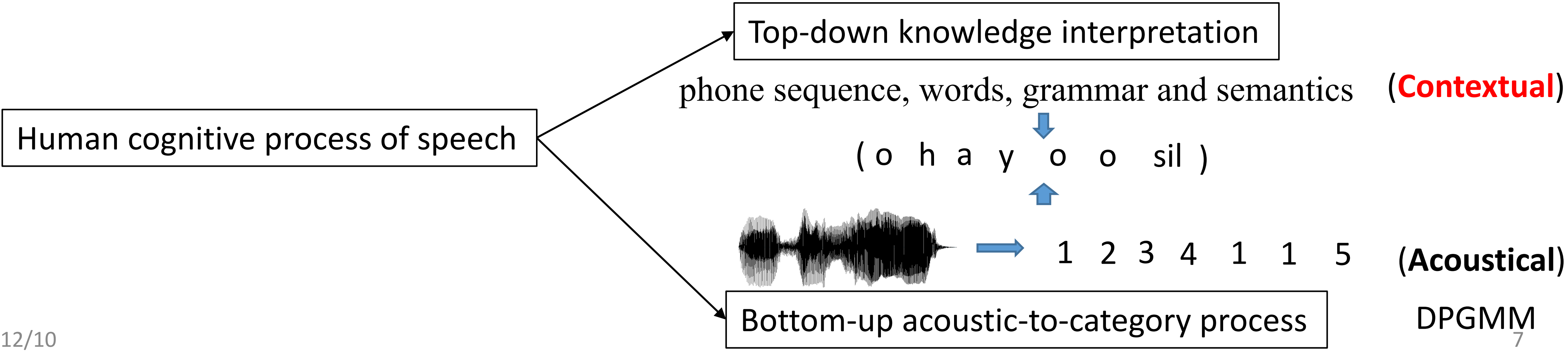
Human cognitive process of phoneme

- Goal: Audio -> Phoneme-like units



o h a y o o sil

- How does the human find the phonemes?



Human cognitive process of speech

Top-down knowledge interpretation

phone sequence, words, grammar and semantics (**Contextual**)

(o h a y o o sil)

1 2 3 4 1 1 5 (**Acoustical**)

Bottom-up acoustic-to-category process

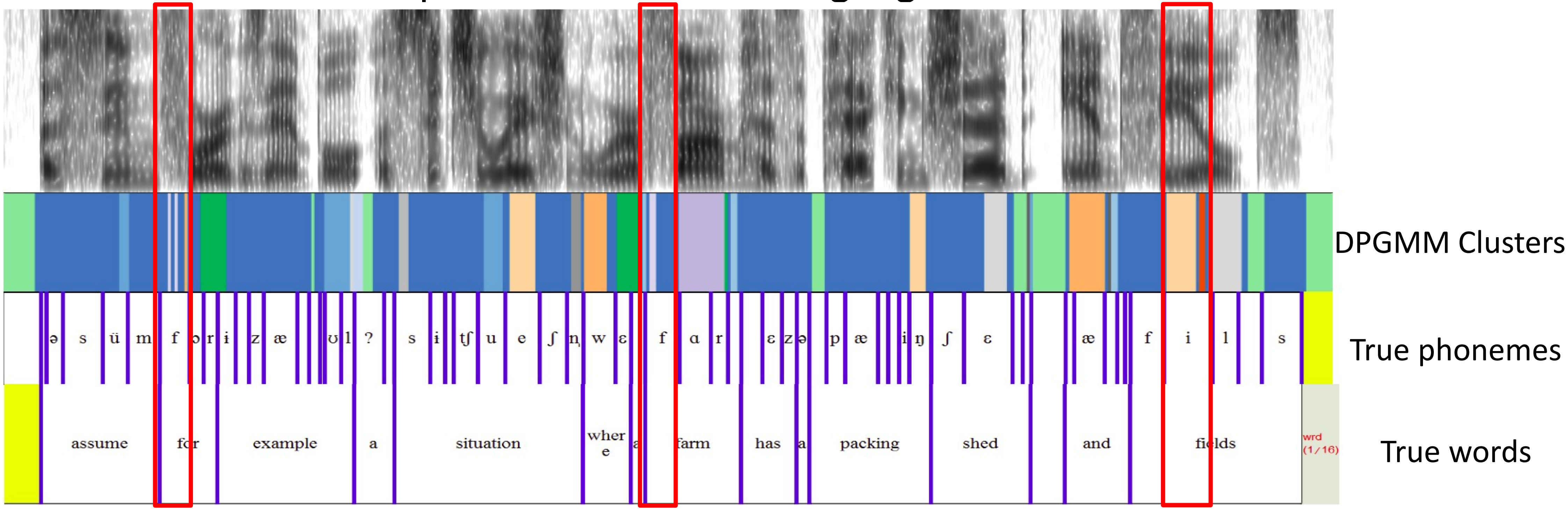
DPGMM

Problem1:DPGMM is too sensitive to acoustics

Problems of DPGMM clustering

- Problem1: DPGMM is too sensitive to acoustics
  - High frequency acoustics make lots of small DPGMM clusters
  - Rapid formant changes make lots of small DPGMM clusters
  - # of clusters > # of phonemes of usual languages

Example:
f: high frequency
i: rapid format change



DPGMM clustering results on timit training corpus

Problem2: DPGMM is weak in contextual modelling

Contextual modelling

- Context is important

School

/s**k1**u:l/

Kite

/**k2**ait/

K1 and K2 is acoustically different
However,
K1 is **always** following s
K2 is **always** following some word boundary

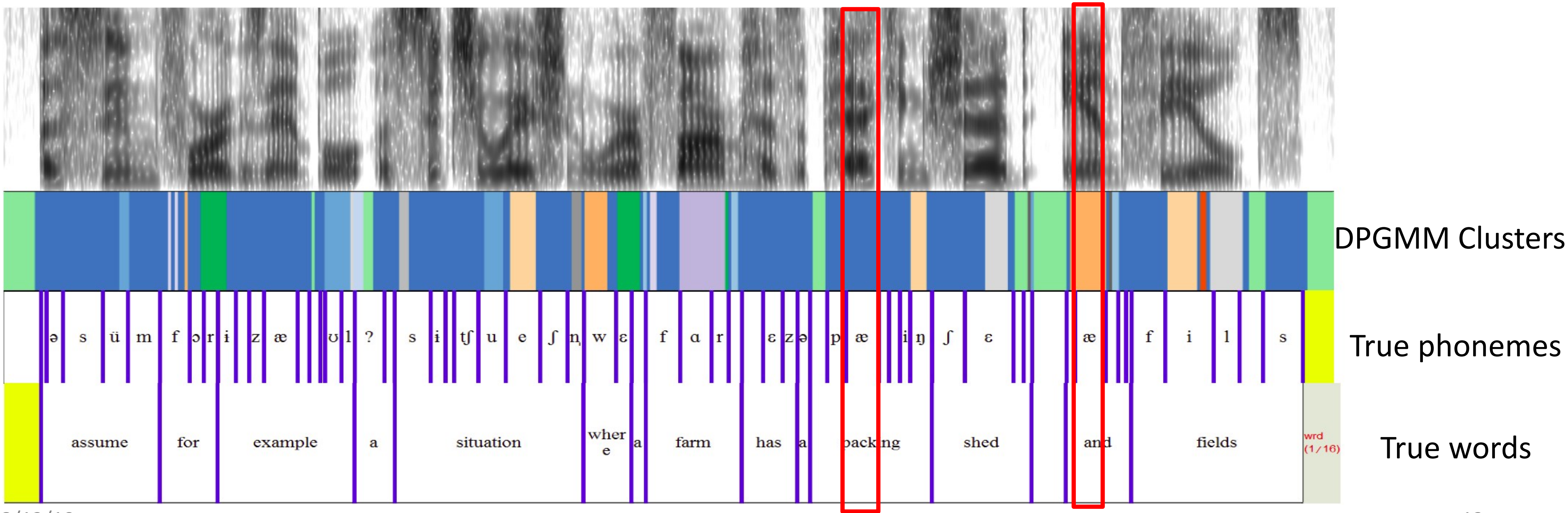
→

K1 and K2 are in completely different context
They belong to same phoneme.

Problems of DPGMM clustering

Example:
- pack: /æ1/ after p
  and: /æ2/ before word boundary
- acoustically different and
  but complementary distribution
- /æ1/ and /æ2/ belong to same
  phoneme /æ/

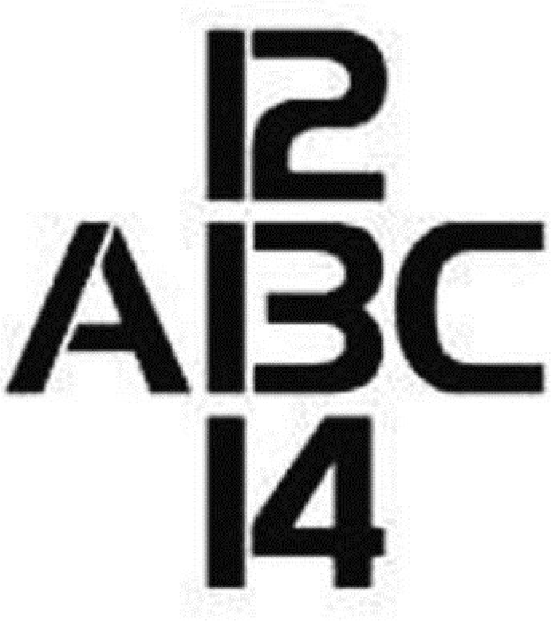
- Problem2: DPGMM is weak in contextual modelling
  - Acoustically different sub-word units are always treated as different labels by DPGMM.
  - Although they are in completely different context and belongs to same phoneme



DPGMM clustering results on timit training corpus

Contextual modelling

- Context is important



Assume B and 13 are two different phonemes,
But they are acoustically similar,
Sometimes B is between A and C
Sometimes 13 is between 12 and 14

→

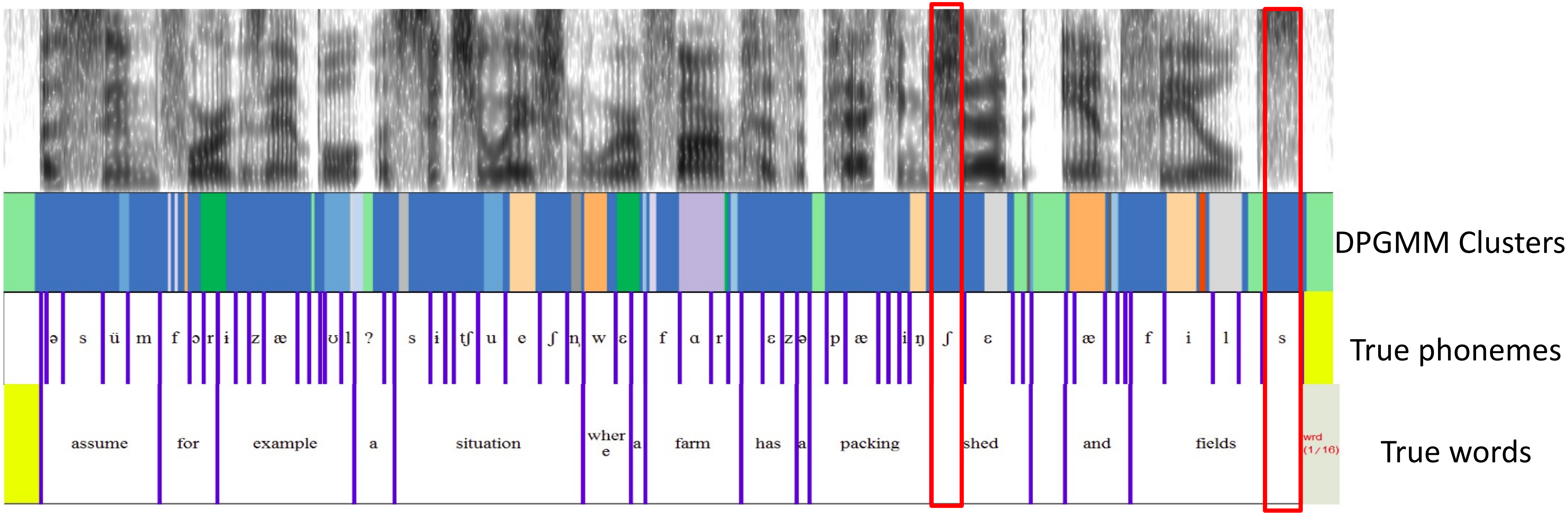
We can distinguish B and 13 by the specific context A, C and 12, 14

Problems of DPGMM clustering

Example:

- Shed: /ʃ/ and fields: /s/
- /ʃ/ and /s/ acoustically similar
- Only /s/ will following /d/
  fields can’t be ended as /d/ + /ʃ/

- Problem3: DPGMM is weak in contextual modelling
  - Context can help distinguish acoustically similar phonemes



DPGMM clustering results on timit training corpus

Problems of DPGMM

- Human use context to distinguish phonemes
  - Acoustic different units with completely different context tends to be the same phoneme
  - Context also helps distinguishing acoustic similar phonemes

- Problems of DPGMM
  - weak in context modeling (top-down)
  - sensitive to acoustics (bottom-up)

Proposal

Proposal

- But How to deal with the contextual effects?
  - Statement:
    - If two units can be easily distinguished by the context.
    - It means the contrast of two units are not important in communication
      - (a.k.a Functional Load (FL) is small)
    - Equivalently, the contrast conveys little information in communication
    - Extremely,

if two units are in
Completely different context,
It means **FL = 0**;
It means conveying **no info**.

Computation of functional load

- The measurement of functional load of the contrasts
  - Information loss ignoring the contrast (Hockett, 1955)
    - functional load of a contrast of a label pair x and y

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

    - eg. In English, K1 and K2 are in completely different context
      - Mathematically, $FL(k1,k2)=0$

School

/s**k1**u:l/

Kite

/**k2**ait/

System configuration

- **Proposal:** greedy mergers based on **least functional load** criteria
  - Iteratively merge the DPGMM label pairs with lowest functional load and enhance our features by ASR

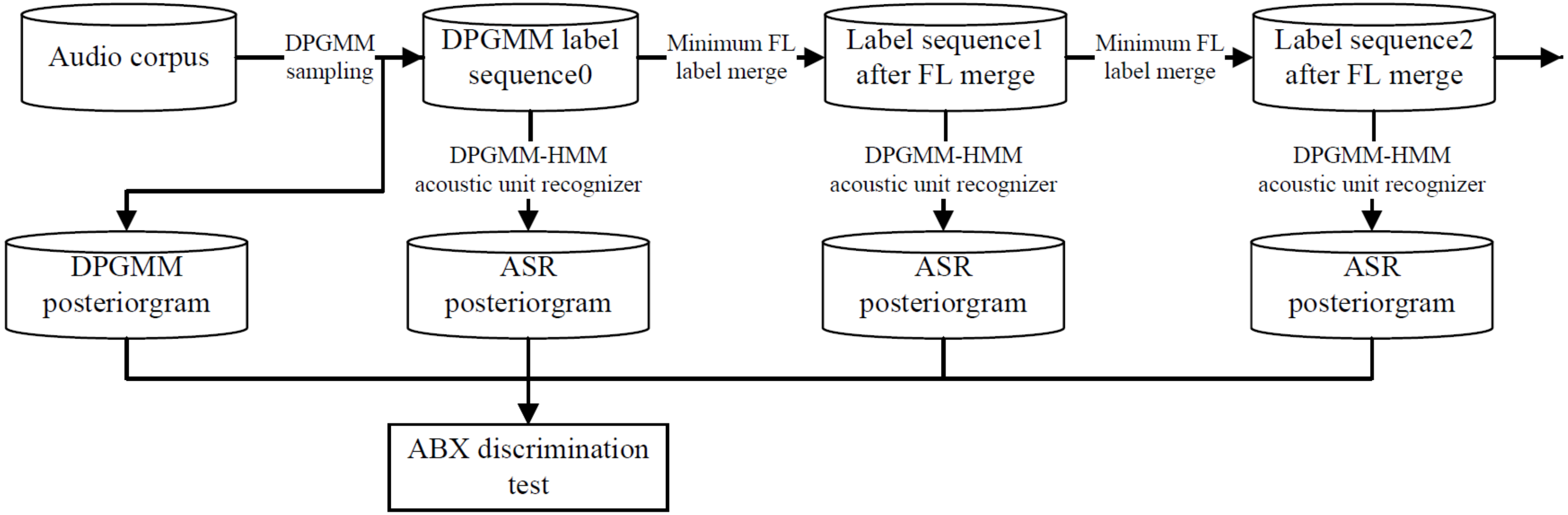


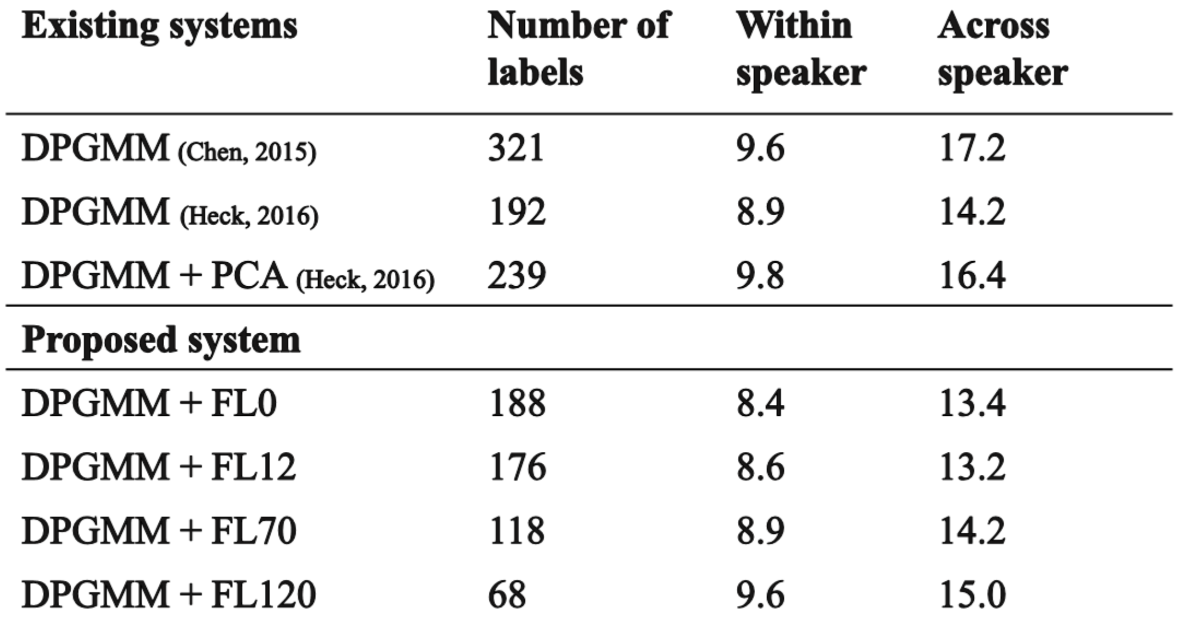
Figure 1: *System to optimize DPGMM based on functional load.*

Experiment & Result

Experiment and result

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

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
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

Conclusion

- DPGMM is weak in context modeling and sensitive to acoustics
- We enhance the contextual modeling of DPGMM labels by minimum functional criteria
- Result shows we can get posterigram of much lower dimension with similar ABX error

Thank you for listening