Speaker Invariant Feature Extraction for Zero-Resource Languages with Adversarial Learning

Taira Tsuchiya, Naohiro Tawara,
Tetsuji Ogawa, Tetsunori Kobayashi

Waseda University

Abstract

- Objective
  - Obtain **speaker invariant features**
  - Apply method to **zero-resource languages**

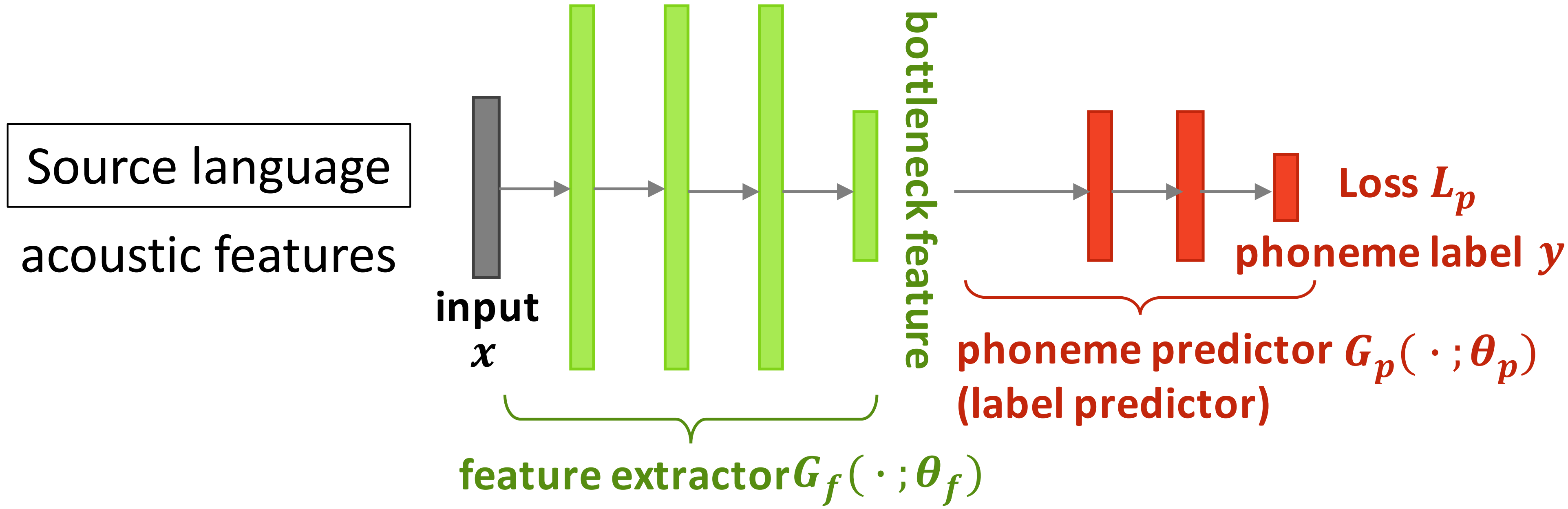
- Approach
  - Introduce "**domain adversarial multi-task learning**" into bottleneck feature extractor

BN feature extraction (for ZR lang) [Renshaw+ 2015]

Step 1: Train acoustic model with **source language**

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$$

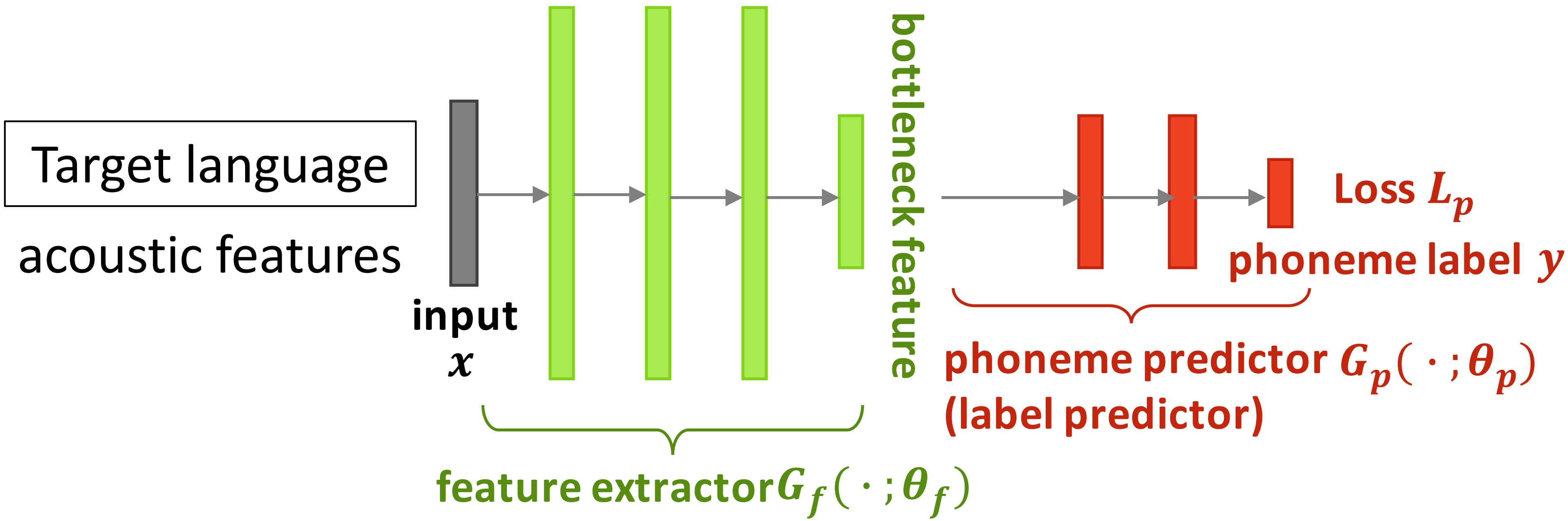
acoustic feature phoneme label



Single-task learning

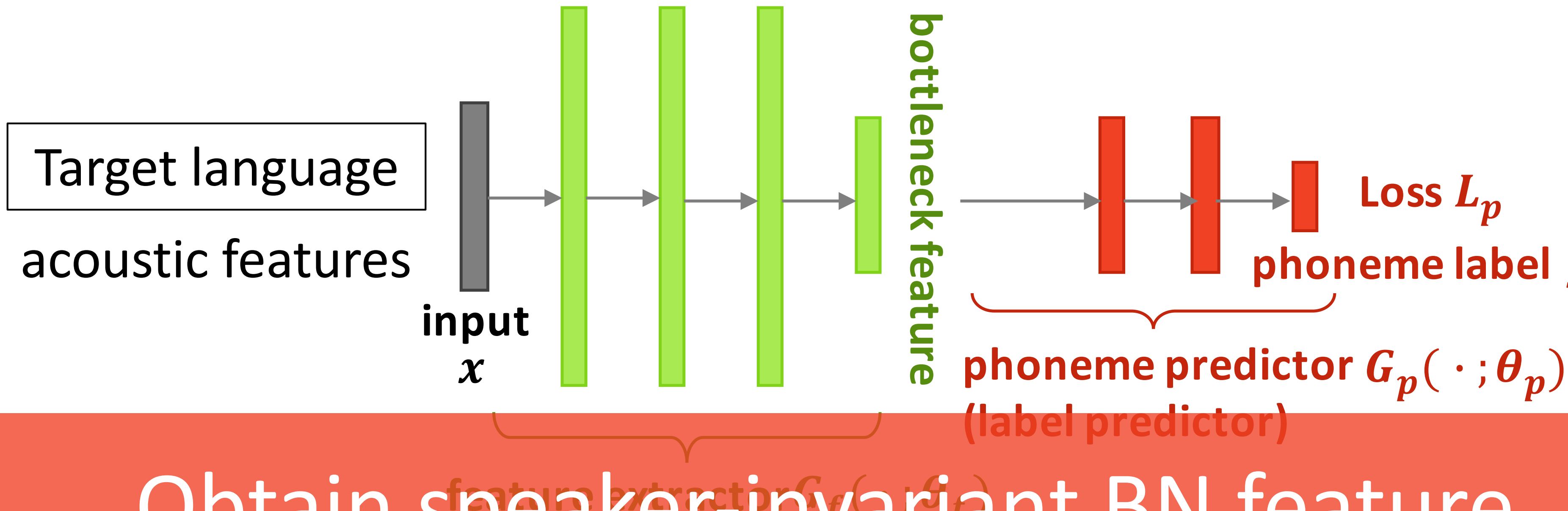
BN feature extraction (for ZR lang) [Renshaw+ 2015]

Step 2: Obtain **bottleneck feature** of **target language**



BN feature extraction (for ZR lang) [Renshaw+ 2015]

Step 2: Obtain **bottleneck feature** of **target language**



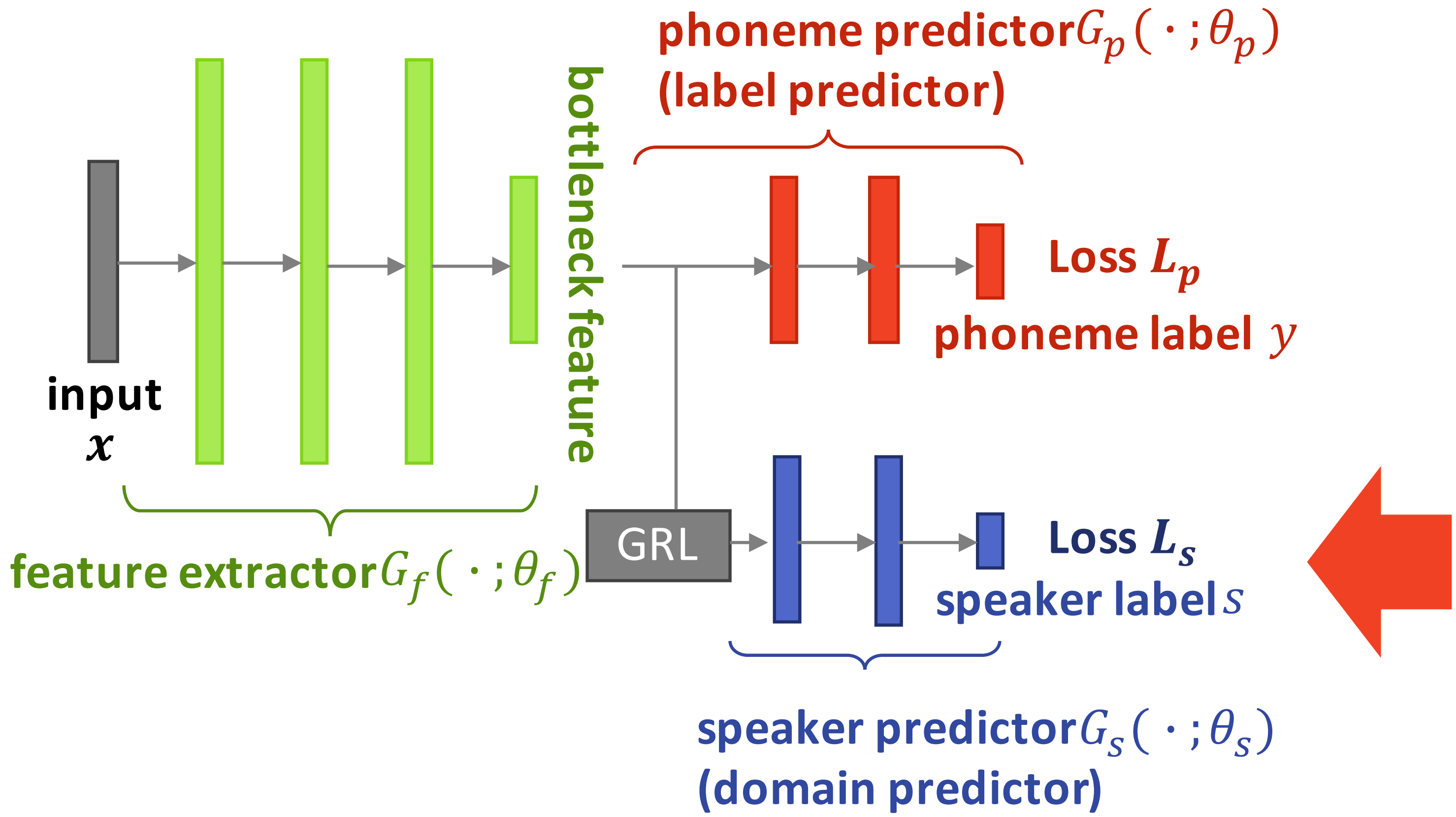
Obtain speaker-invariant BN feature
By Adversarial Multi-task Learning

Our work: problem setting and structure

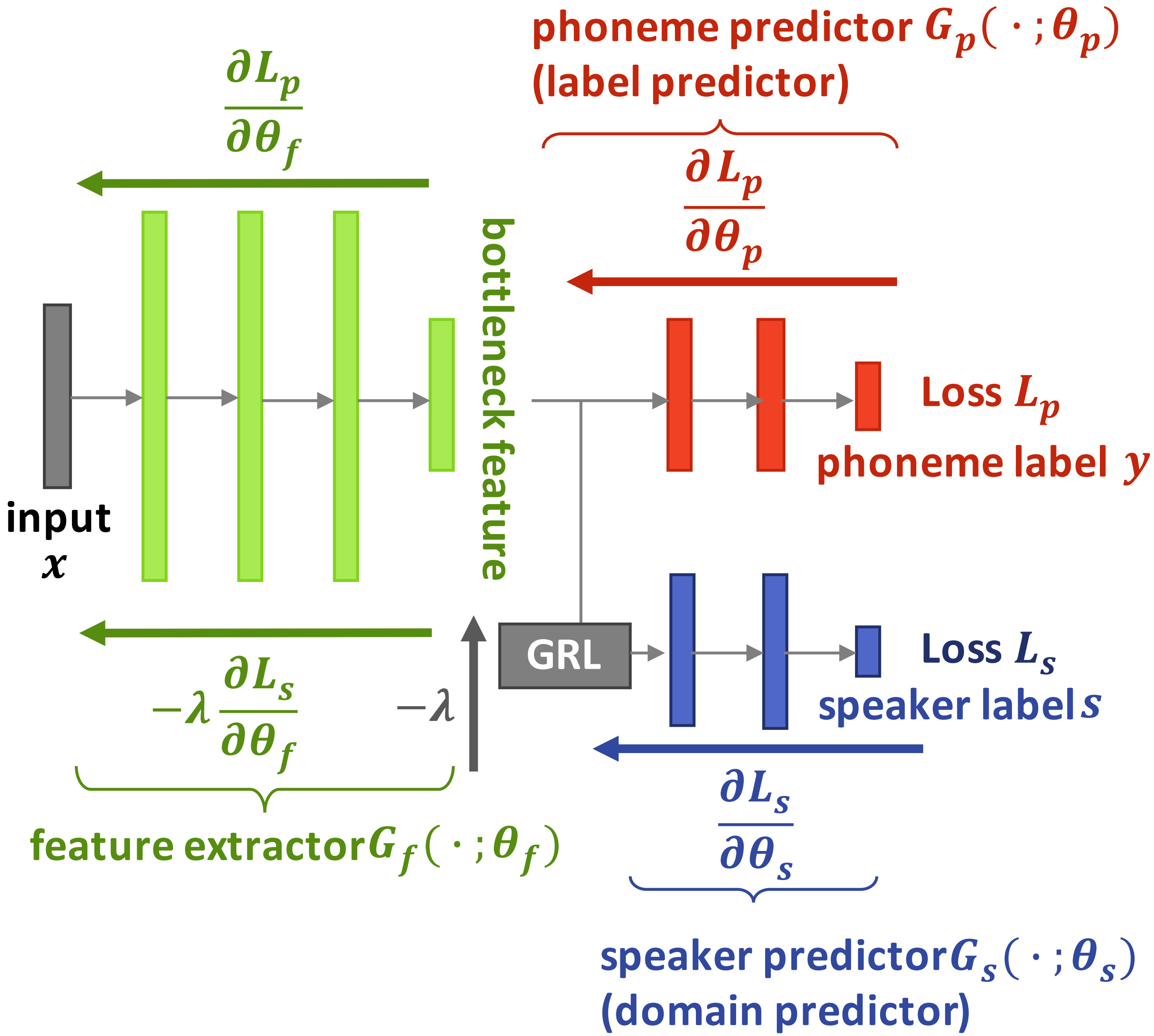
- Resource abundant languages

$$\{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^{N} \qquad s_i \in \{1, \dots, C\}$$ **Speaker labels**

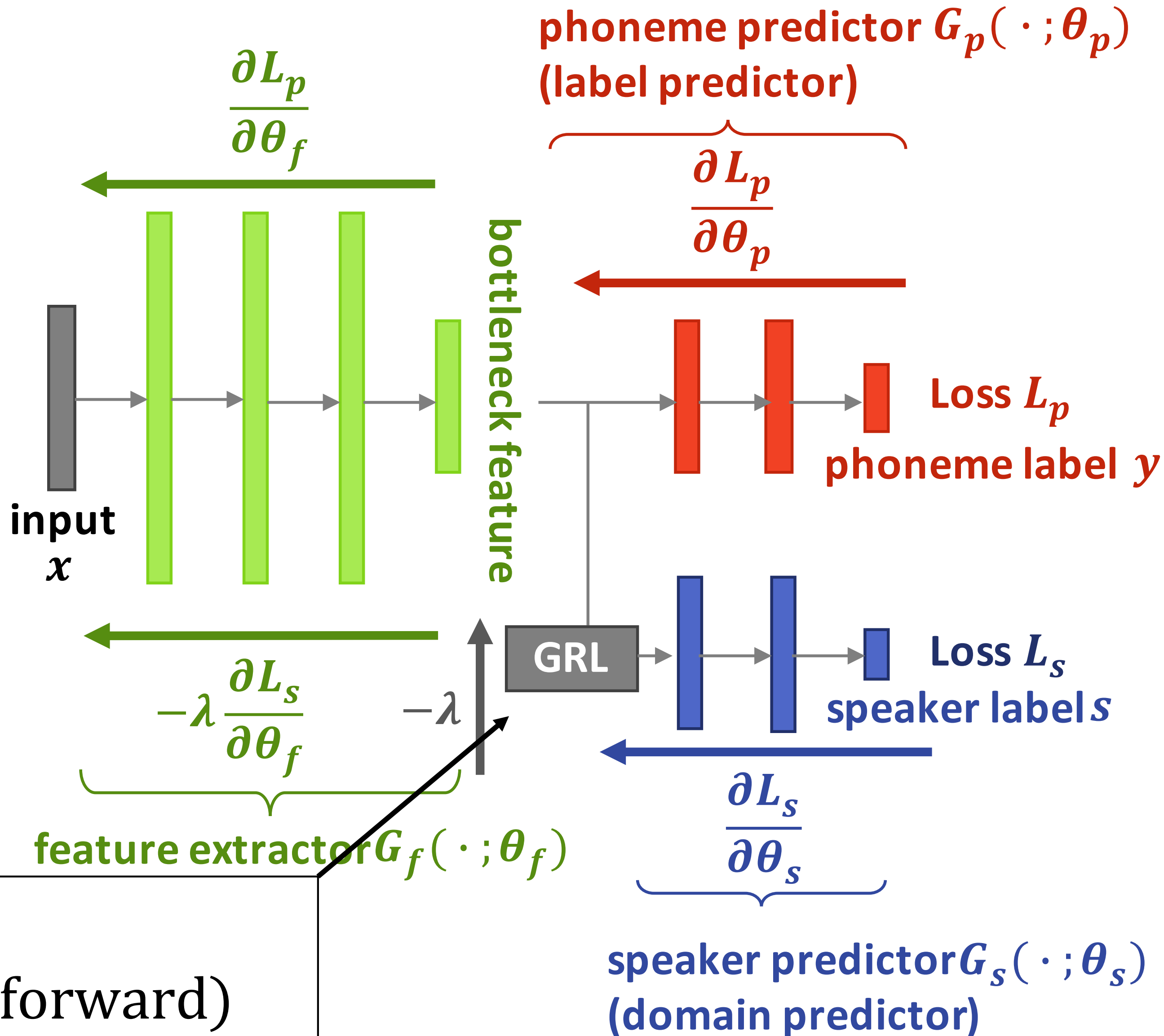
- Structure – insert **speaker predictor**



Adversarial multi-task learning



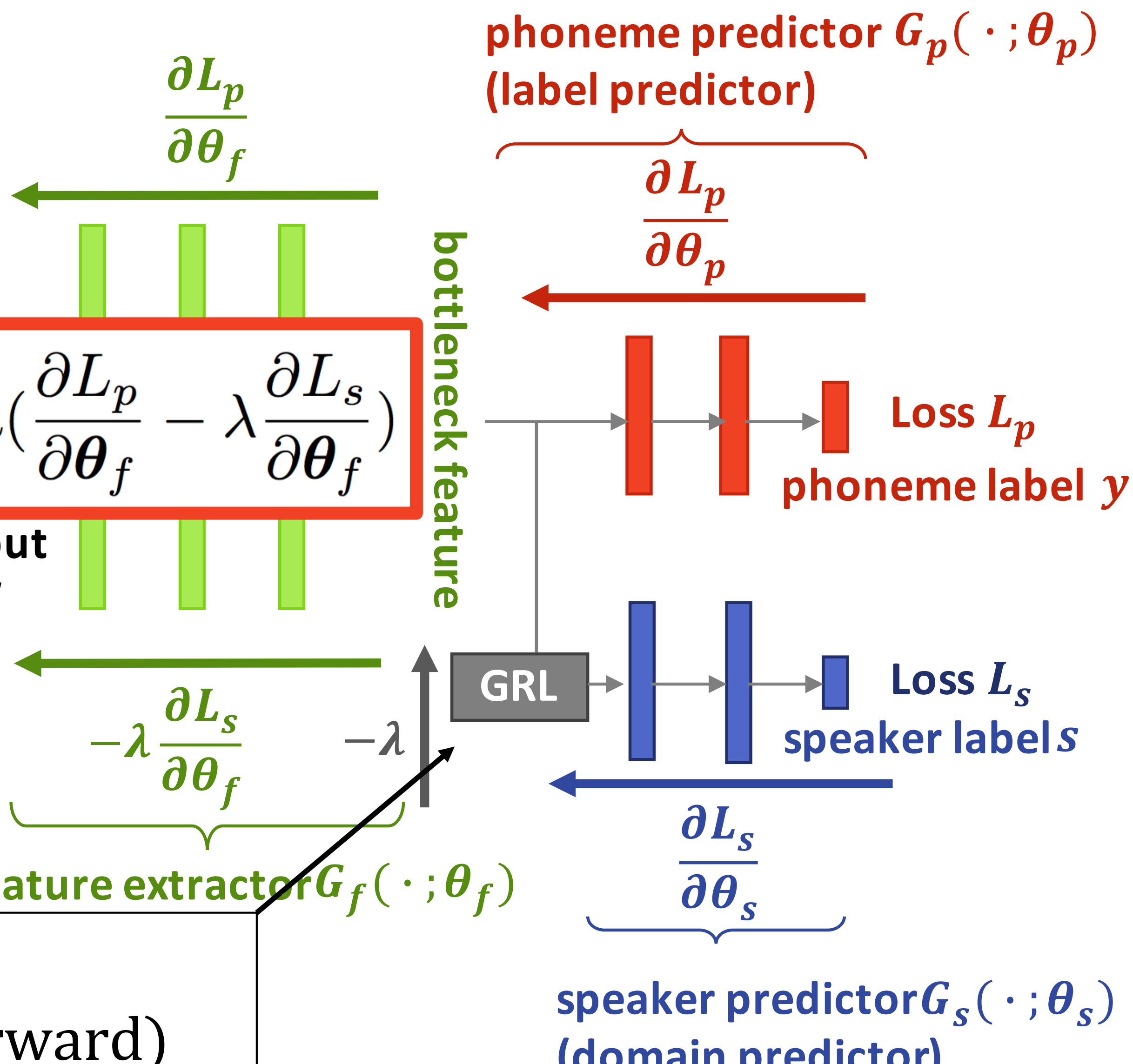
Adversarial multi-task learning



$$GRL(x) = \begin{cases} x \text{ (forward)} \\ -\lambda x \text{ (backward)} \end{cases}$$

GRL := Gradient Reversal Layer

Adversarial multi-task learning

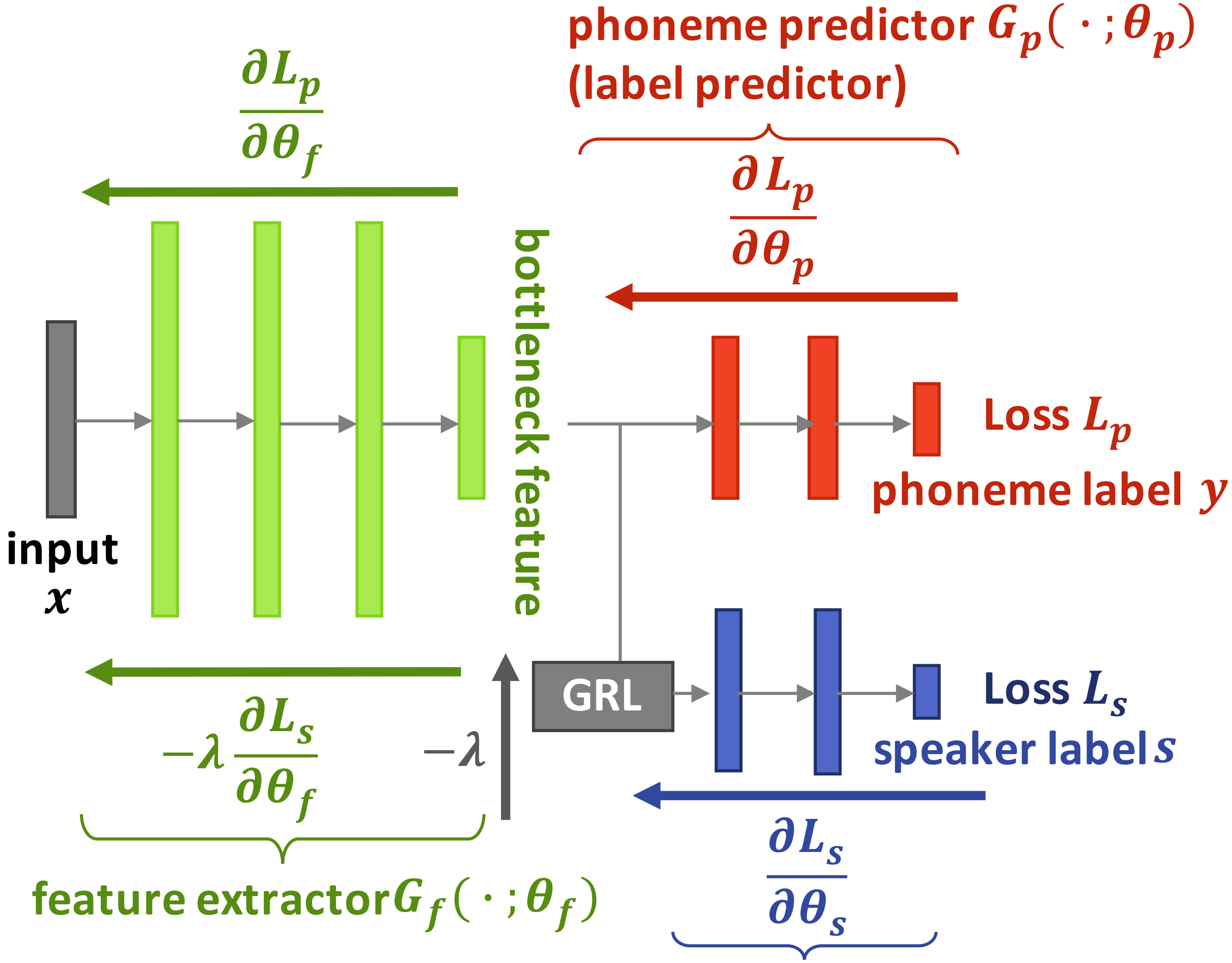


$$\boxed{\boldsymbol{\theta}_f \leftarrow \boldsymbol{\theta}_f - \mu\left(\frac{\partial L_p}{\partial \boldsymbol{\theta}_f} - \lambda \frac{\partial L_s}{\partial \boldsymbol{\theta}_f}\right)}$$

$$GRL(x) = \begin{cases} x \text{ (forward)} \\ -\lambda x \text{ (backward)} \end{cases}$$

GRL := Gradient Reversal Layer

Adversarial multi-task learning



Bottleneck feature: Easy to recognize phonemes,
but **difficult to predict speakers**.

Experiments

- Goal
  - Evaluate **features of zero-resource languages** from **phoneme discriminability** viewpoint

- Compared features
  - Acoustic feature (fMLLR)
  - Bottleneck feature (single-task learning)
  - Bottleneck feature (adversarial multi-task learning)

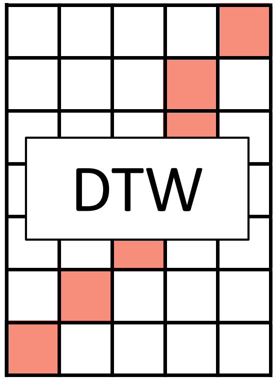
Phoneme discriminability of features

- <u>ABX error rate</u>

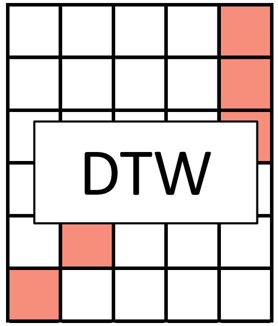
$$a, x \in A, b \in B$$

$$a = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_m]$$

$$b = [\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_n]$$



$d(a, x)$ $d(b, x)$



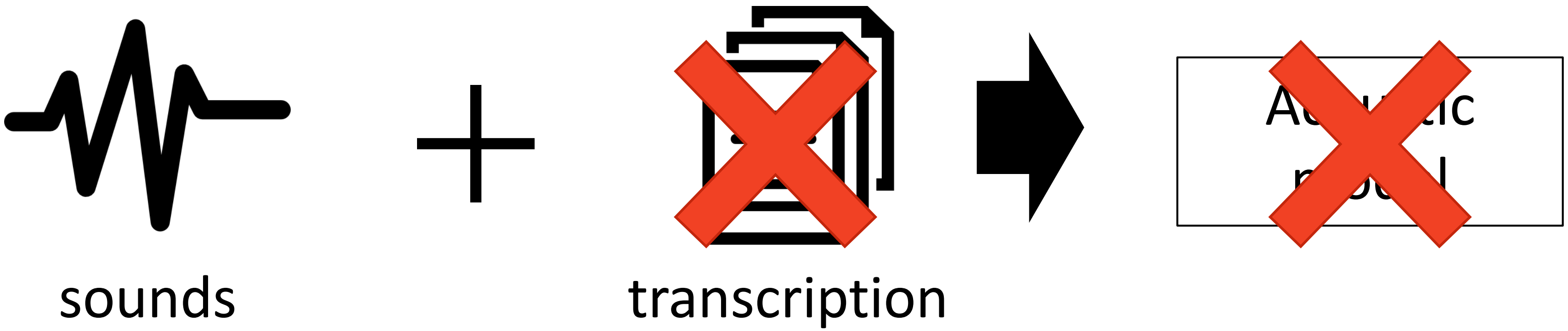
$$x = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_l]$$

Expect: Distance of same phonemes features are smaller than that of different phonemes. $\boxed{d(a, x) < d(b, x)}$

Characteristics of zero-resource lang.

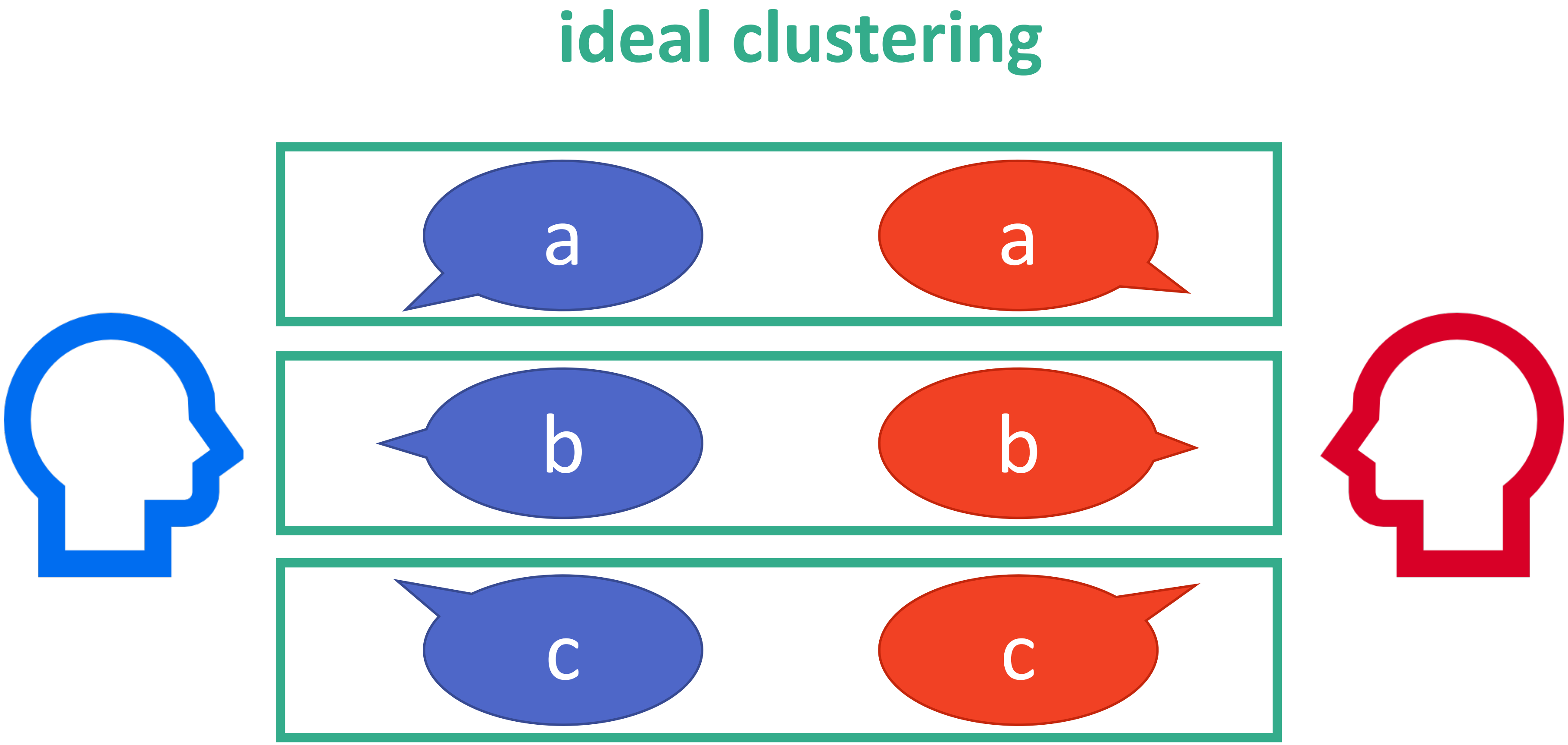
Resource abundant languages (-> source languages)

Zero-resource languages (-> target languages)



In zero-resource lang, transcription of its sound is NOT available. -> **unsupervised** settings !!

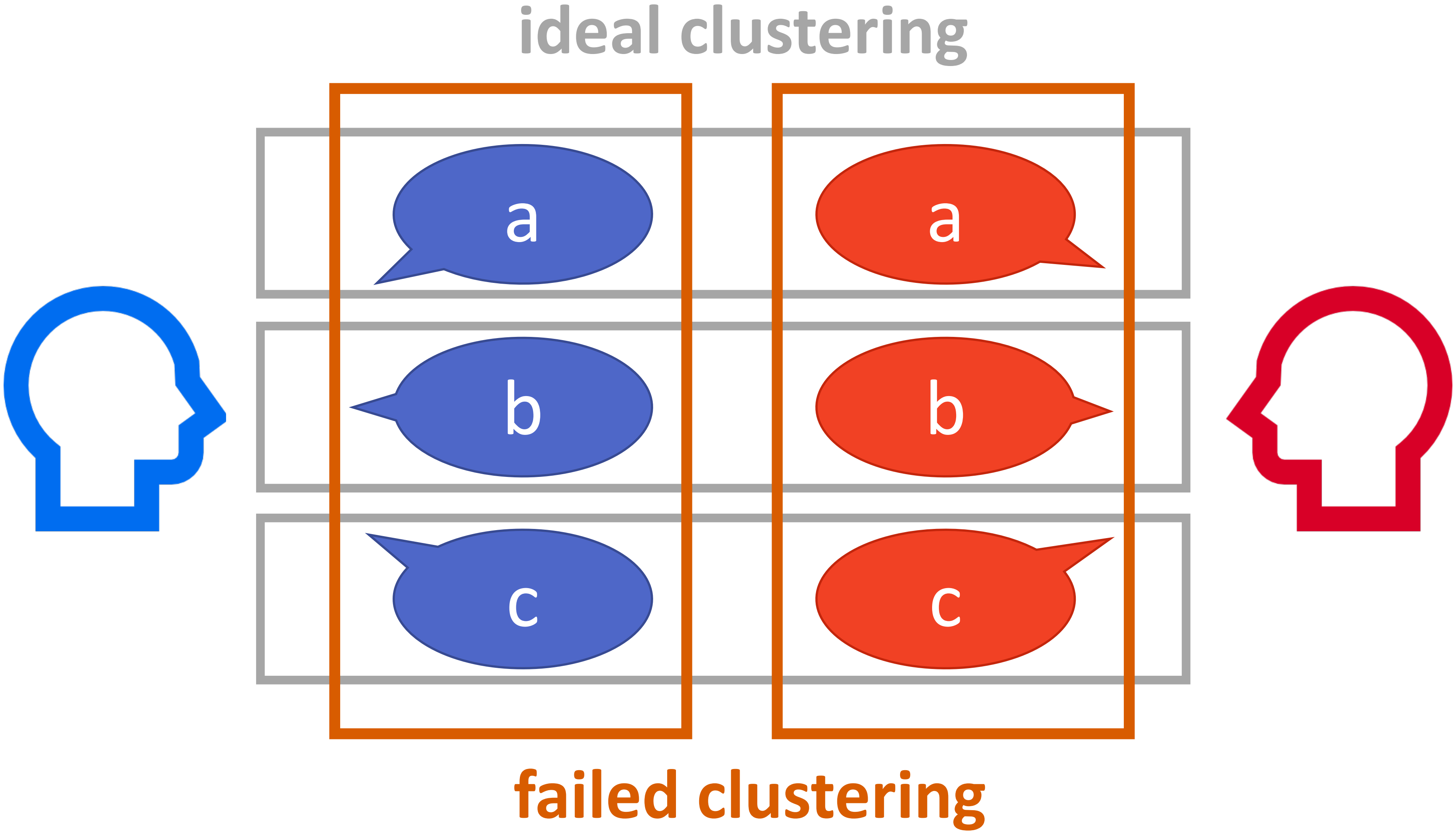
Harmful speaker effect in ZR lang



Phoneme discriminative feature

Clustered by **phonemes** even of different speakers.

Harmful speaker effect in ZR lang



Removing speaker information from features is crucial especially in **zero-resource languages**.

Data

- Dataset
  - Zero Resource Speech Challenge 2017
- Training data (resource abundant (source) language)
  - Lang: English
  - # of speakers : 9 ($C = 9$) → Train models
  - Total length : 35 hours
- Evaluation data (target languages)
  - English -> resource abundant languages
  - French
  - Mandarin

  French and Mandarin: Zero-resource languages → Evaluate ABX error rate

- Input x : fMLLR obtained by **English (resource abundant language)**
- Concatenate with five frames before and after -> 220ms

Experimental settings

- Optimizer : SGD with learning rate adjustment

$$\mu_p = \frac{\mu_0}{(1 + \alpha \cdot p)^\beta} \qquad \mu_0 = 0.01, \alpha = 10, \beta = 0.75$$

- Mini-batch size : 1024

- Dropout
- Batch normalization

Results

- Evaluate different speakers' segments (Across)

	English	French	Mandarin
raw feature	10.83	14.83	10.35
bottleneck feature	7.06	12.10	8.90
bottleneck feature (adversarial)	**6.80**	**11.87**	**8.73**

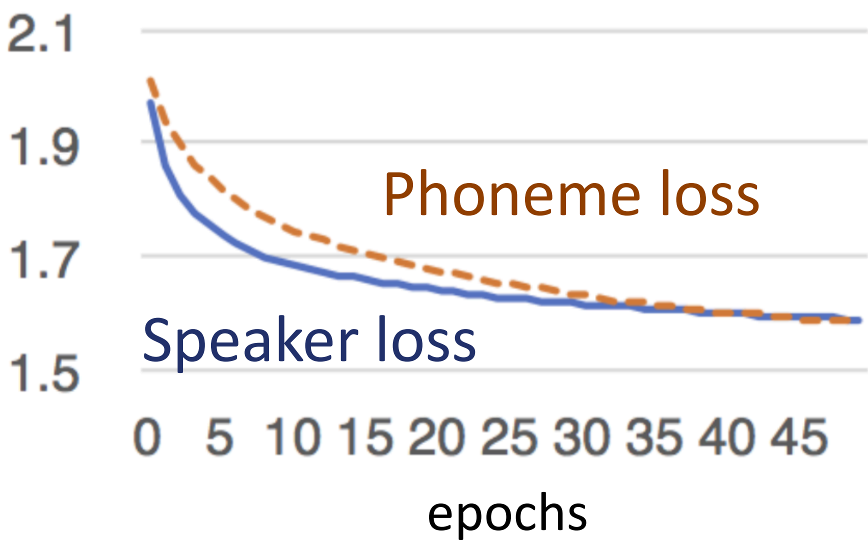
- Evaluate same speaker's segments (Within)

	English	French	Mandarin
raw feature	6.85	8.96	8.74
bottleneck feature	4.96	8.06	8.01
bottleneck feature (adversarial)	**4.71**	**7.59**	**7.82**

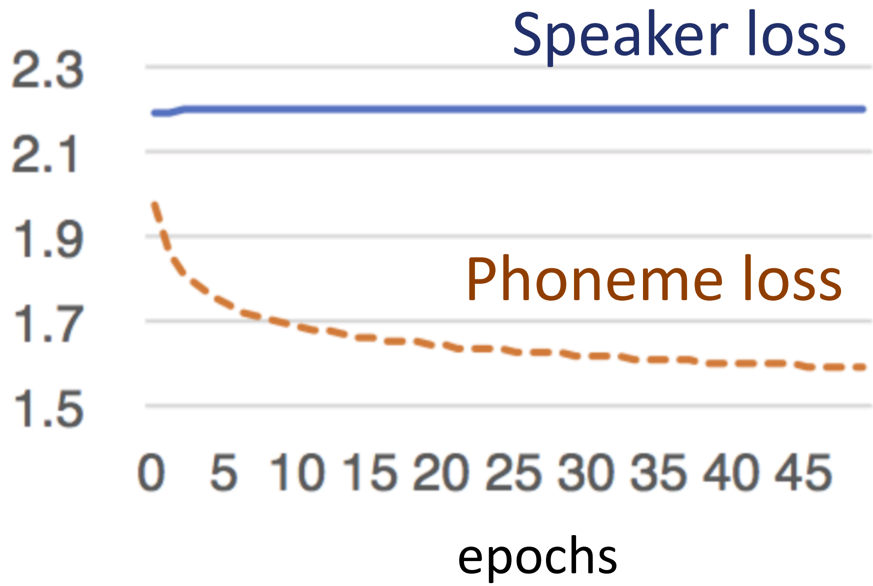
$$a = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_m] \qquad b = [\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_n]$$

$$d(a, x) \qquad x = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_l] \qquad d(b, x)$$

Learning curve – validation loss



Single-task learning



Adversarial multi-task learning

In our method, phoneme loss decreases while speaker loss is kept high.

Conclusion and future work

- Goal
  - Obtain phoneme discriminative features of target languages by suppressing speaker information
- Proposed method
  - Extend bottleneck feature approach
  - Introduce adversarial multi-task learning and explicitly suppress speaker information from BN feature
- Future work
  - 220ms would be not enough to obtain speaker information
  - Introduce more long context information
  - Extend to another kind of networks
  - Unsupervised-learning settings