DIVERGENCE ESTIMATION BASED ON DEEP NEURAL NETWORKS AND ITS USE FOR LANGUAGE IDENTIFICATION

Yosuke Kashiwagi, Congying Zhang, Daisuke Saito, Nobuaki Minematsu (The University of Tokyo)

Introduction

- Statistical divergence between distributions, such as Kullback-Leibler divergence or Bhattacharyya Divergence has been widely used.

Bhattacharyya Divergence

$$BD(a,b) = -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x}$$

- Since statistical divergence is defined as a functional of two probability density functions, a parametric form of the distribution is required.

$$BD(a,b) = \frac{1}{8}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)})^\top \Sigma^{-1}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)}) + \frac{1}{2}\ln\left(\frac{\det \Sigma}{\sqrt{\det \Sigma^{(a)} \det \Sigma^{(b)}}}\right)$$
$$\Sigma = \frac{\Sigma^{(a)} + \Sigma^{(b)}}{2}$$

- However, the "true" distribution can have a complex shape.
  - To increase estimation accuracy, more complex models are applied.
    - Gaussian Mixture Model based approximation [J. R. Hershey, 2007]
    - log-linear model based approximation [Heigold, 2011] [J. Li, 2014]
- We propose a new discriminative technique to estimate the statistical divergence not using generative parameters explicitly.
  - Flexibility of Deep Neural Network (DNNs) is effectively introduced to estimate the statistical divergence.

Proposed approach

- When DNN-based models are available, they can directly calculate posterior probabilities.
  - Applying Bayes' theorem, the Bhattacharyya Divergence is represented as a functional of the posterior probabilities as

$$\begin{aligned}BD(a,b) &= -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \\ &= -\ln \int p(\boldsymbol{x})\sqrt{p(y=a|\boldsymbol{x})p(y=b|\boldsymbol{x})}d\boldsymbol{x} \\ &\quad + \frac{1}{2}\ln p(y=a) \\ &\quad + \frac{1}{2}\ln p(y=b)\end{aligned}$$

calculated using DNNs

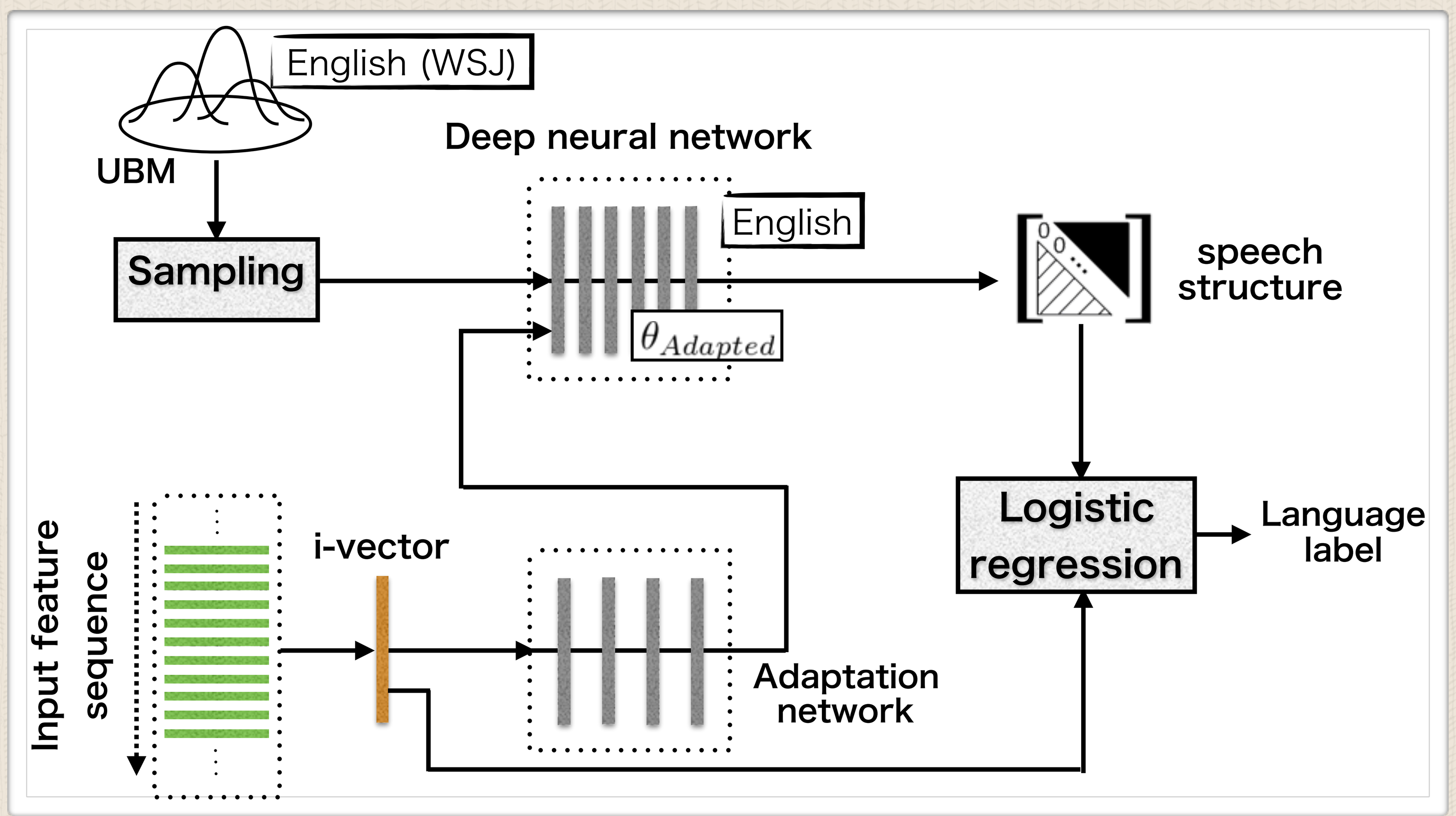
  - using sampling approach

$$BD(a,b) = -\ln \frac{1}{L}\sum_l \sqrt{p(y_l=a|\boldsymbol{x}_l,\theta)p(y_l=b|\boldsymbol{x}_l,\theta)} + \frac{1}{2}\ln\frac{1}{L}\sum_l p(y_l=a) + \frac{1}{2}\ln\frac{1}{L}\sum_l p(y_l=b)$$

- In discriminative model, such as Deep neural networks (DNNs), the parametric form of the feature distribution is not explicitly assumed.
- Discriminative model can characterize the feature distribution more flexibly.

Use for language identification

- Overview



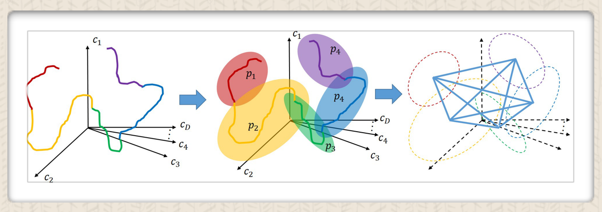
1. The system assumes at first that all the input utterances as English.
2. DNNs, which are used as English phoneme posterior estimator, are adapted to the input utterance using i-vector [Y. Miao, 2014].
3. Calculates the BD between every possible pair of the132 English phoneme states using sampling method. It is called "structural features."
4. They will be used as additional features of logistic regression.

- If we use a set of feature samples collected only from the observed utterance, however, it is intractable to calculate the summation over the entire feature space.
- To address this problem, we use Universal Background Models (UBMs) for sampling.

$$BD(a,b) = -\ln\frac{1}{N}\sum_n \sqrt{p(y_n=a|\boldsymbol{x}_n,\theta_{Adapted})p(y_n=b|\boldsymbol{x}_n,\theta_{Adapted})} + \frac{1}{2}\ln\frac{1}{L}\sum_l p(y_l=a) + \frac{1}{2}\ln\frac{1}{L}\sum_l p(y_l=b)$$

approximate as constants

- If we use Gaussian distribution to model the feature distribution of each phoneme state, the Bhattacharyya Divergence is invariant to any static affine transformation. This means that the structural features are robust to speaker differences.

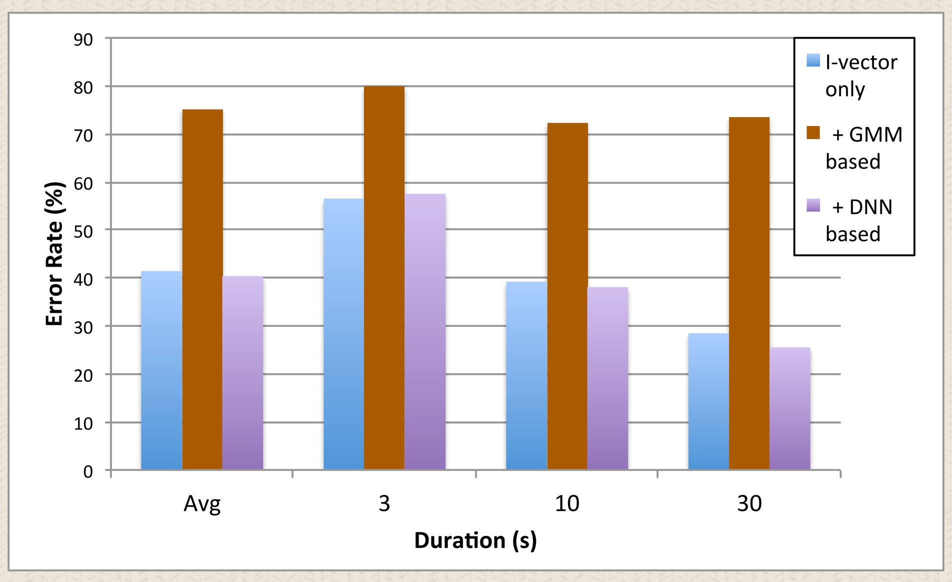


- Even if all the kinds of phonemes are not observed in input utterances, our system can calculate the divergence related to those unobserved phonemes using only from i-vector.

Experiments

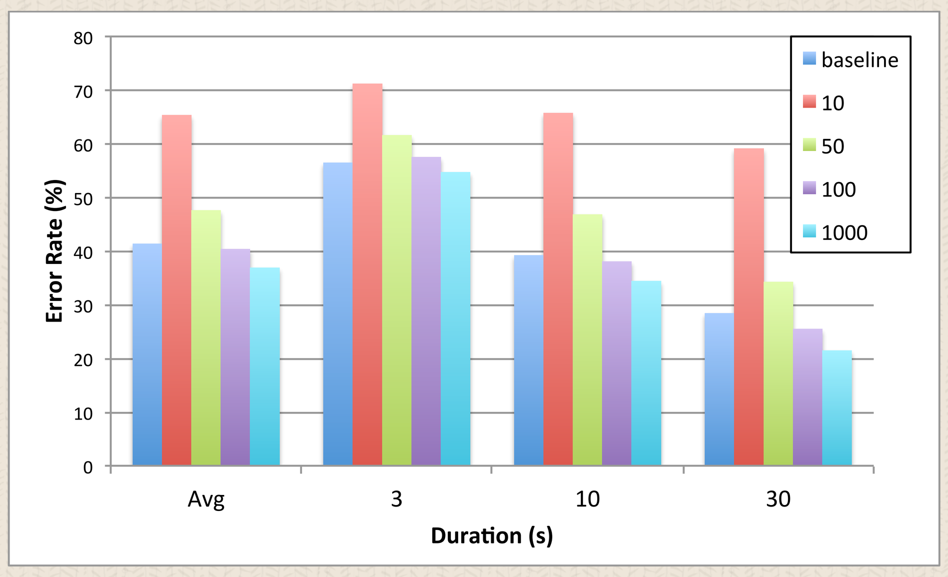
- Evaluation data
  - NIST LRE test set (contains 3, 10, 30 seconds utterances)
- i-vector
  - database: NIST LRE 2003, 2005, 2007 training
  - input features: MFCC (6 dim.) + power
  - dimension: 600
- DNNs
  - database: WSJ (English data)
  - input features: MFCC (12 dim.) + C0 + neighboring 10 frames
  - network: 6 hidden layers and each layer has 1024 nodes
  - output labels: monophone states (132 dim.)
- Adaptation network
  - database: WSJ (English data)
  - network: 4 hidden layers and each layer has 1024 nodes
- Universal background model for sampling
  - database: WSJ (English data)
  - the number of mixtures is 1024

- Comparison among our proposed system (100 frames for sampling) and the two baseline systems in terms of error rates



- "i-vector only" only used i-vector as input of logistic regression
- "GMM-based" calculated the Bhattacharyya divergence from MAP-adapted UBM.

- Error rates as a function of the duration of input utterances for different numbers of sampled frames



- If the number of sampled frames is small, the performance of our approach is lower.
- However, our proposed approach becomes effective when the number increases up to 1,000.