# Unsupervised Representation Learning of Speech for Dialect Identification

*Suwon Shon, Wei-Ning Hsu, James Glass*

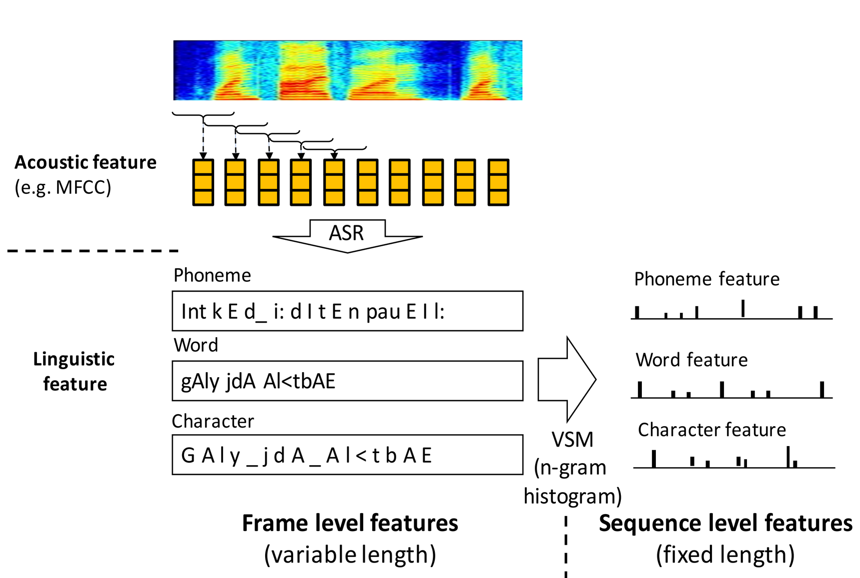
MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA

## Motivation

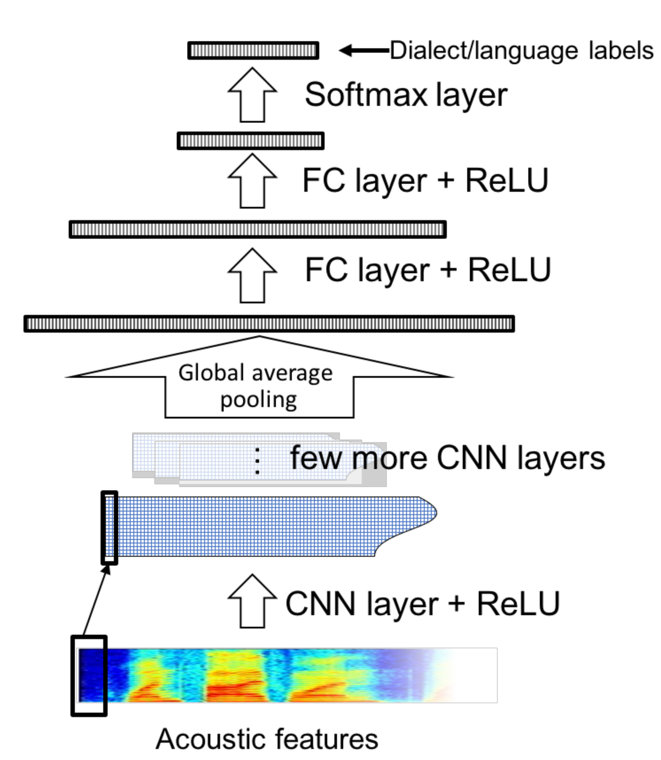
- One of the challenges of processing real-world spoken content, such as media broadcasts, is the potential presence of different dialects of a language in the material
- Dialect identification (DID) can be a useful capability to identify which dialect is being spoken during a recording
- Since DID tasks are data dependent, unsupervised learning from unlabeled datasets is much more important than for other resource-rich tasks
- The Factorized Hierarchical Variational Autoencoder (FHVAE) model can represent static and dynamic generating factors within an utterance
- We argue that language related information like accent, tone, rhythm are mostly encoded in the dynamic generating factors

## Language/Dialect Identification System

- i-vector system
  - Used with MFCCs or Stacked Bottleneck (SBN) Features
  - i-vector extractor generally trained without any supervision
- End-to-End system using CNN/DNN
  - Using MFCCs, logmel-filterbanks or spectrograms
  - Multiple layers of CNNs and a dense layer
  - Global (average) pooling layer to convert the frame level representation to utterance level

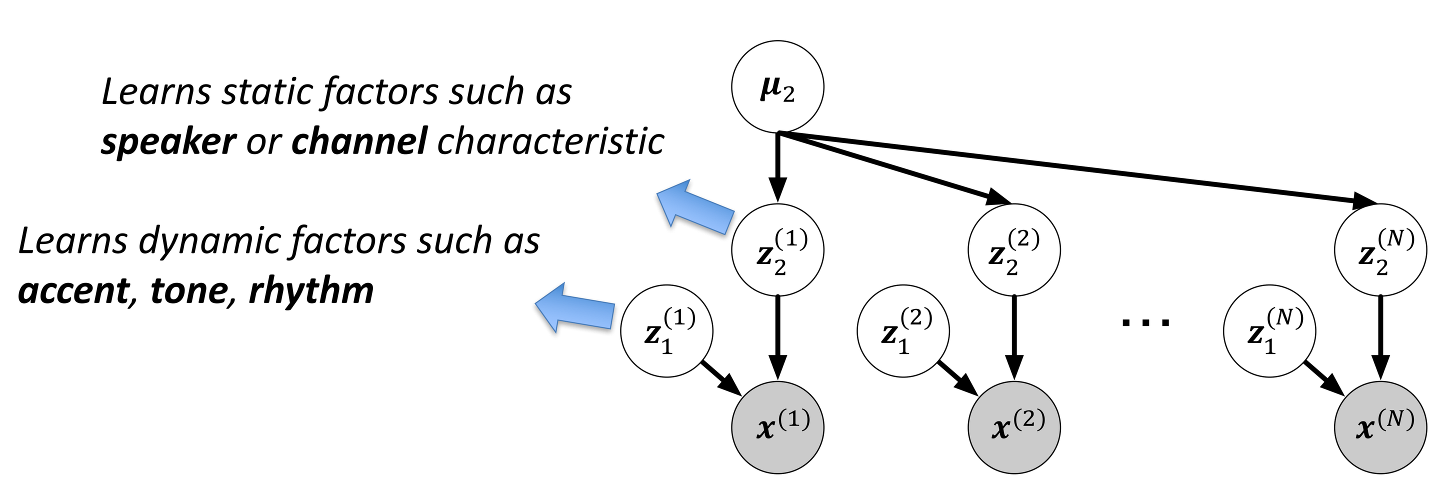



<Feature extraction flow >




<Network architecture for end-to-end system>

## Unsupervised Learning using FHVAE




<Graphical illustration of the FHVAE generative model. Grey nodes denote the observed variables, and white nodes are the latent variables>

- Use FHVAE to learn representation from dialectal speech without supervision
  - FHVAE is a variant of the variational auto-encoder that learns a disentangled representation from sequential data
  - For a given sequence $\boldsymbol{X}=\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ *FHVAE* involves $N$ pairs of sequence-level and segment-level latent variable $z_1$ and $z_2$ as follows :

1. a *s-vector* $\boldsymbol{\mu}_2$ is drawn from $p(\boldsymbol{\mu}_2)=\mathcal{N}(\boldsymbol{\mu}_2|\boldsymbol{0},\sigma^2_{\boldsymbol{\mu}_2}\boldsymbol{I})$.
2. $N$ i.i.d. *latent segment variables* $\boldsymbol{Z}_1=\{\boldsymbol{z}_1^{(n)}\}_{n=1}^{N}$ are drawn from a global prior $p(\boldsymbol{z}_1)=\mathcal{N}(\boldsymbol{z}_1|\boldsymbol{0},\sigma^2_{\boldsymbol{z}_1}\boldsymbol{I})$.
3. $N$ i.i.d. *latent sequence variables* $\boldsymbol{Z}_2=\{\boldsymbol{z}_2^{(n)}\}_{n=1}^{N}$ are drawn from a sequence-dependent prior $p(\boldsymbol{z}_2|\boldsymbol{\mu}_2)=\mathcal{N}(\boldsymbol{z}_2|\boldsymbol{\mu}_2,\sigma^2_{\boldsymbol{z}_2}\boldsymbol{I})$.
4. $N$ i.i.d. sub-sequences $\boldsymbol{X}=\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ are drawn from $p(\boldsymbol{x}|\boldsymbol{z}_1,\boldsymbol{z}_2)=\mathcal{N}(\boldsymbol{x}|f_{\mu_x}(\boldsymbol{z}_1,\boldsymbol{z}_2),diag(f_{\sigma^2_x}(\boldsymbol{z}_1,\boldsymbol{z}_2)))$, where $f_{\mu_x}(\cdot,\cdot)$ and $f_{\sigma^2_x}(\cdot,\cdot)$ are parameterized by a decoder neural network.

  - The model encourages $z_2$ to represent relatively consistent latent factors within a sequence such as channel response, vocal tract characteristics
  - We use $z_1$ as a new feature since phonetic and lexical variability is useful information for dialect identification

## Experiments

### <Dialectal Speech Dataset>

- 5 dialects : Modern Standard Arabic, Egyptian, Levantine, Gulf, North African
- The test domain is different from the training dataset

**Table 1**. MGB-3 Dialectal Arabic Speech Dataset Properties.

| Dataset | Training | Development | Test |
|---|---|---|---|
| Utterances | 13,825 | 1,524 | 1,492 |
| Size | 53.6 hrs | 10 hrs | 10.1 hrs |
| Channel (recording) | Carried out at 16kHz | Downloaded directly from a high-quality video server | |

### <Resource Limitation Impact on Domain Mismatch>

- With a target domain label, the end-to-end system shows impressive performance compared to the i-vector system
- Without a target domain label, both the i-vector and end-to-end discriminative model show similar performance

| System | Accuracy on Test set | |
|---|---|---|
| | If Dev. set is labeled | If Dev. set is unlabeled |
| I-vector | +14% 57.44 | −19% 46.11 +5% |
| End-to-End (MFCC) | **65.55** | −25% **48.86** |
| End-to-End (FBANK) | 64.81 | 47.11 |

<Baseline accuracy on MGB-3 Test set>

### <Resource-Rich Condition>

- Trained FHAVE model with both train and development set
- $z_1$ learns dynamic factors from the input segment and effectively learns dialectal information from speech

| | Accuracy | EER | $C_{avg}$*100 |
|---|---|---|---|
| i-vector +14% | 57.44 | 24.43 | 23.79 |
| End-to-end (MFCC) | 65.55 | 20.24 | 19.92 |
| End-to-end (FBANK) +4% | 64.81 | 20.22 | 19.91 |
| **End-to-end (FHVAE_$z_1$)** | **67.98** | **18.62** | **18.32** |
| End-to-end (FHVAE_$z_2$) | 54.55 | 27.39 | 27.35 |

<With Label>

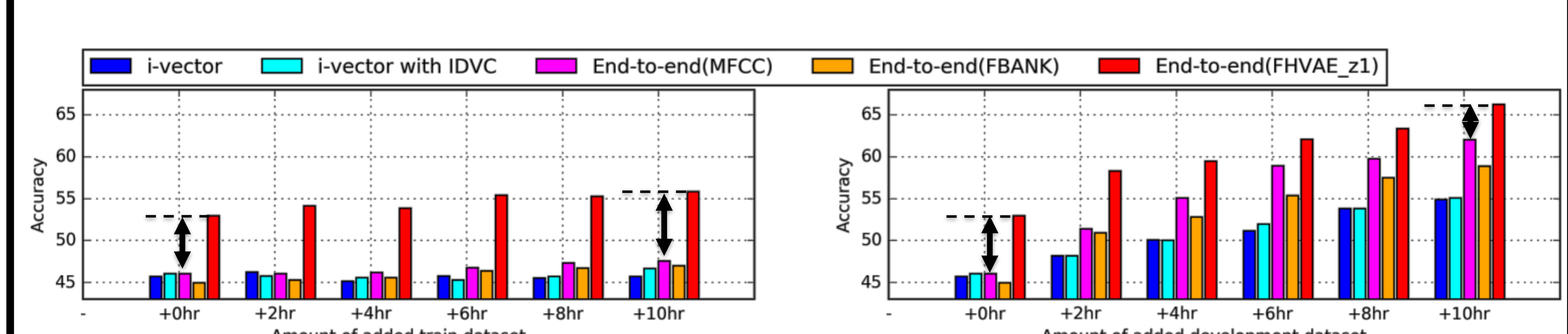
### <Resource-Poor Condition>

- When a label is unavailable, unsupervised learning with FHVAE has much more effective than other approaches
- $z_1$ is more effective than $z_2$ as we expected because it encodes dynamic factors such as accent, tone, rhythm

| | Accuracy | EER | $C_{avg}$*100 |
|---|---|---|---|
| i-vector +5% | 46.11 | 32.77 | 32.08 |
| End-to-end (MFCC) | 48.86 | 29.31 | 28.61 |
| End-to-end (FBANK) +19% | 47.86 | 30.19 | 29.67 |
| **End-to-end (FHVAE_$z_1$)** | **58.16** | **25.40** | **24.66** |
| End-to-end (FHVAE_$z_2$) | 36.36 | 39.00 | 38.32 |

<Without Label>

- When we add a domain mismatched labeled dataset for training, unsupervised learning shows still more effective performance than other approaches
- When we add a domain matched labeled dataset for training, the performance of the two systems (MFCC and FHVAE_z1) get gradually closer




<Efficiency of domain mismatched and matched labeled dataset>

**Table 7**. Performance comparison on MGB-3 test set.

| Resource-poor | Accuracy | EER | $C_{avg}$*100 |
|---|---|---|---|
| End-to-end (uBNF [18] | 56.64 | 27.46 | 26.92 |
| **End-to-end (FHVAE_$z_1$)** | **58.16** | **25.40** | **24.66** |
| Resource-rich | Accuracy | EER | $C_{avg}$*100 |
| End-to-end (uBNF [18]) | 66.24 | 19.98 | 19.63 |
| **End-to-end (FHVAE_$z_1$)** | **67.98** | **18.62** | **18.32** |

<Comparison with another unsupervised learning of speech, uBNF>

## Conclusion

- One of the challenges of processing real-world spoken content, such as media broadcasts, is the potential presence of different dialects of a language in the material
- A major challenge for Dialect identification (DID) is that there is not enough data or labeled data for training
- Unsupervised learning could be a remedy for such low-resource languages
- The proposed FHVAE based unsupervised learning effectively encodes language/dialectal information from speech
- The latent variables learned from FHVAE can substitute MFCCs as robust features that exploit dialectal information from speech