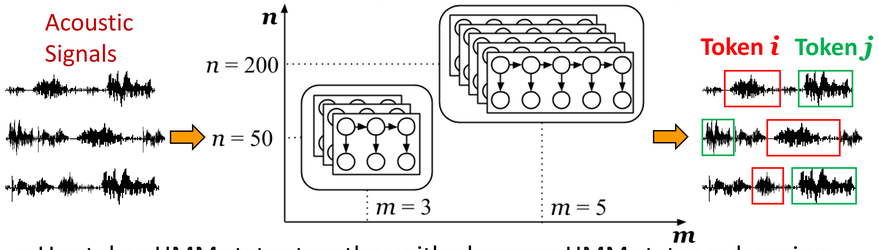
Personalized Acoustic Modeling by Weakly Supervised Multi-task Deep Learning Using Acoustic Tokens Discovered from Unlabeled Data

Cheng-Kuan Wei, Cheng-Tao Chung, Hung-Yi Lee, and Lin-Shan Lee
National Taiwan University

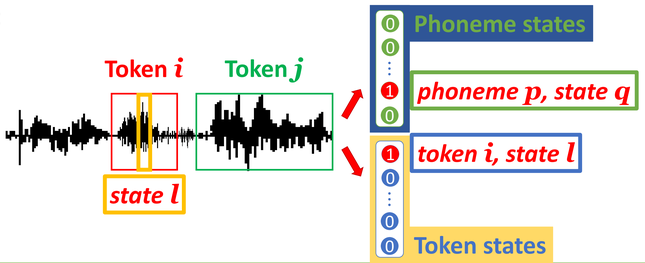
1. Introduction

- Motivation: with the popularity of smartphones,
  - Easier to collect huge personal audio data which can be used
  - Little transcribed data as references for learning
- Proposed: **phoneme-token deep neural network (PTDNN)**
  - Jointly trained by **unsupervised acoustic tokens + transcribed phonemes**
  - "Weakly supervised" scenario
    - **Vast unlabeled data** & **limited transcribed data**

- Unsupervised Acoustic Tokens
  - Acoustically similar signal patterns automatically discovered from a large unlabeled corpus [1]
  - Specified by **number of token HMM states m** (token length or temporal granularity) and **number of distinct tokens n** (phonetic granularity)
  - Offer high level information regarding how each signal segment sounds like

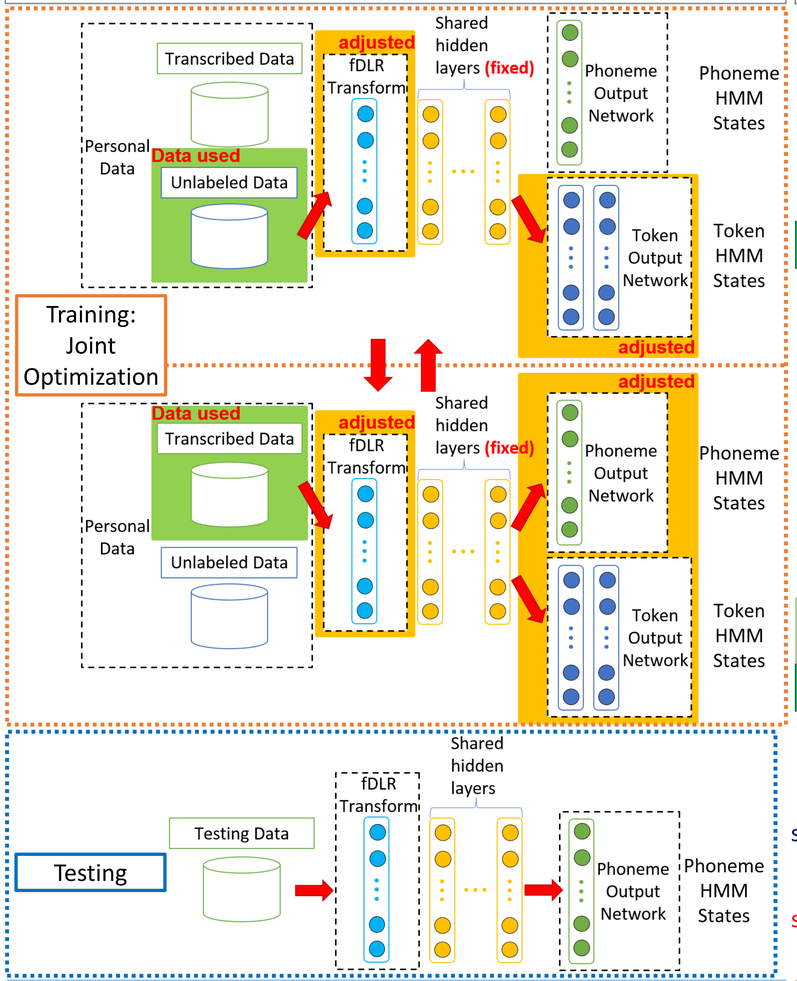


  - Use token HMM states together with phoneme HMM states as learning targets:



[1] C.-T. Chung, C.-a. Chan, and L.-s. Lee "Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity"

2. Proposed Approach

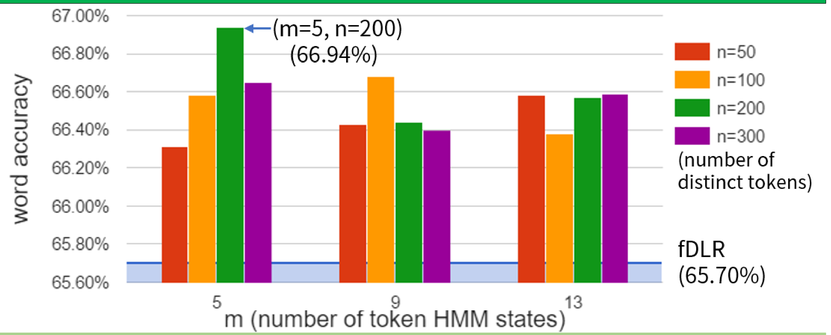


- Multiple sets of acoustic tokens (with different values of m and n) can be learned jointly by **adding more targets** and **output networks**

3. Experiments & Analysis

- SI training data: **31.8 hours Chinese** + **29.7 hours English read speech**
- Adaptation data: **Facebook posts produced by 5 male, 5 female**
  - Chinese-English Bilingual (4.1%: English words)
  - **1000 utterances** for each of **10 speakers**
  - **500 Adapt. ($\mathbf{M}$ transcribed + $(\mathbf{500}-\mathbf{M})$ unlabeled), 250 Valid., 250 Testing**
- **Personal acoustic token sets**: train on **adaptation set of each speaker**

① A single set of tokens with different choices of m and n



- **50 transcribed + 450 unlabeled utterances (from adaptation set)**
- **Robust** to the choice of m **and** n

② More or Less Transcribed Data and More Token Sets

	Word accuracy	Number of transcribed utterances		
	Models	10	50	100
(A)	SI (DNN-HMM)	57.45%		
supervised (B-1)	fDLR	60.08%	65.70%	67.73%
supervised (B-2)	Speaker code	60.34%	65.92%	67.89%
weakly supervised (C)	Lightly supervised adaptation	61.45%	62.10%	63.18%
weakly supervised (D-1)	Proposed, $m = 5$, $n = 200$	63.05%	66.94%	68.65%
weakly supervised (D-2)	Proposed with 4 token sets	**64.89%**	**67.15%**	**68.75%**

(C) $(\mathbf{500}-\mathbf{M})$ unlabeled data transcribed by SI model and used in adaption

- **More improvements** when given **fewer transcribed utterances**
- **Different acoustic token sets** was **slightly complementary**