Discriminatively Trained Joint Speaker and Environment Representations for Adaptation of Deep Neural Network Acoustic Models

Maofan Yin[1], Sunil Sivadas[2], Kai Yu[1], Bin Ma[2]

[1]SpeechLab., Shanghai Jiao Tong University, China
[2]Human Language Technology Lab., Institute for Infocomm Research, Singapore

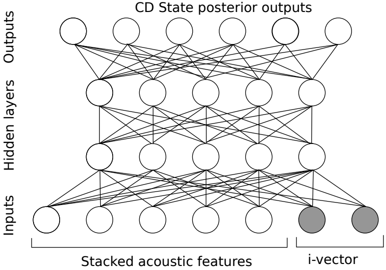
March 22, 2016

Joint Speaker-Environment Representation

Introduction

- DNN: there have been tremendous advances in the accuracy of large vocabulary speech recognition systems
- The performance improvements are largely limited to clean and moderately noisy test conditions
- Solution: normalization of speaker and environment variability
  - fMLLR: feature-transform-based
  - CAT (Cluster Adaptive Training): structured-model-based
  - Multi-condition training: data-based
  - Augmentation the DNN input with auxiliary features (*)

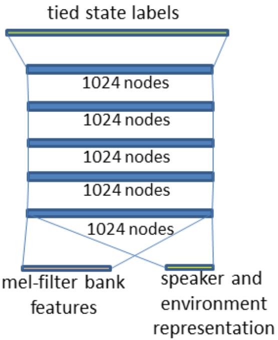
- Concatenate i-vector for a given speaker (IBM) / utterance (Google) to every frame
- Concatenate noise estimation for each utterance to every frame (Microsoft, Cambridge)



Joint Speaker-Environment Representation

Introduction

- (*) Use the i-vector/noise-spectrum as a non-phonetic representation to augment input
- These coarse representations could be finer $\Longrightarrow$ "JSER"



Joint Speaker-Environment Representation

An overview of the idea

- Derive a joint representation of speaker and environment that can be used to augment DNN input
- Use **noisy i-vectors** as input to train the DNN that estimates the **Joint Speaker and Environment Representation** (JSER)

Joint Speaker-Environment Representation

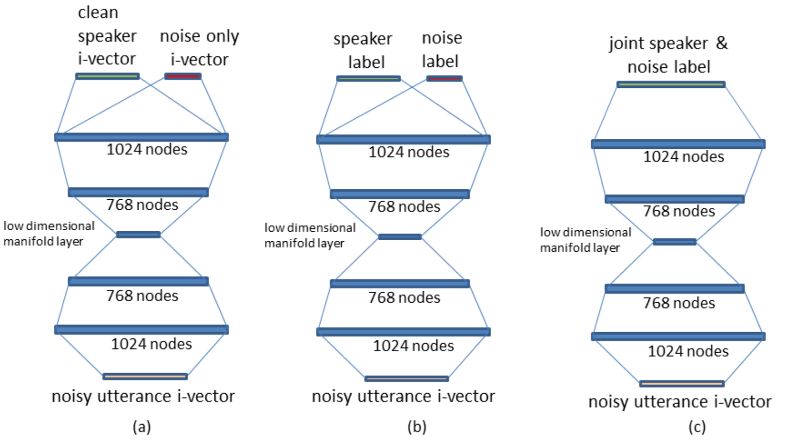
Why i-vectors?

- i-vector is a low-dimensional representation of the acoustic variability related to:
  - Speakers
  - Environment
  - Dialects

  etc., rather than phonetic variability

Joint Speaker-Environment Representation

The proposed model: MTL-MSE-JSER, MTL-CE-JSER, JTL-CE-JSER



Joint Speaker-Environment Representation

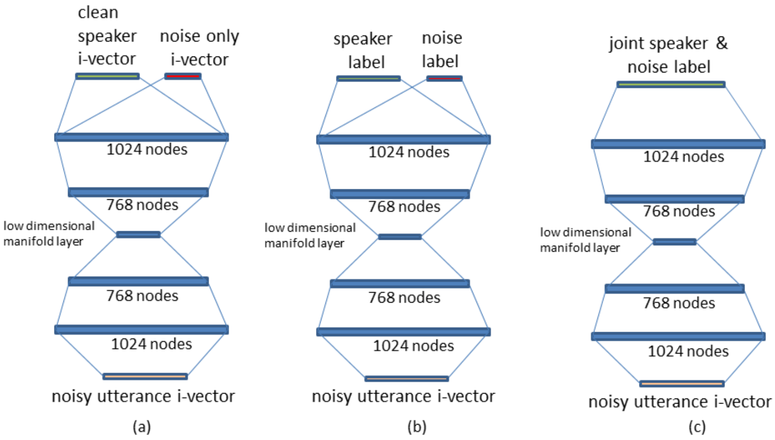
Experiments: setup

- Experiments were conducted on corrupted WSJ databases
  - 84 speaker WSJ0 subset for training the acoustic model
  - WSJ0 + WSJ1 for training the Joint Speaker and Environment Representation (JSER) transforms
  - 8 different types of noise were added to the clean waveforms at different SNRs

- Two different noise corrupted databases
  - i-vector extraction
    - 283 speakers
    - 8 noise types $\times$ 8 SNRs $\Longrightarrow$ 64 times the size of clean database
  - Acoustic model training (TBC)

Joint Speaker-Environment Representation

The JSER model, revisited



Joint Speaker-Environment Representation

Experiments: JSER prediction accuracy

Multi-Task Learning	Speaker		Environment	
	Train	CV	Train	CV
MTL-MSE-JSER (60)	0.0501	0.0633	0.0931	0.1337
MTL-CE-JSER (60)	99.28	97.39	93.94	89.50
Joint-Task Learning	**Spk. $\times$ Env.**			
	Train		CV	
JTL-CE-JSER (60)	93.02		80.62	

Table: Speaker and noise classification performance of JSER-DNNs. For MTL-MSE-JSER, the numbers are MSE values and for the rest they are classification accuracies in percentage. The number in brackets is the dimensionality of the bottleneck layer.

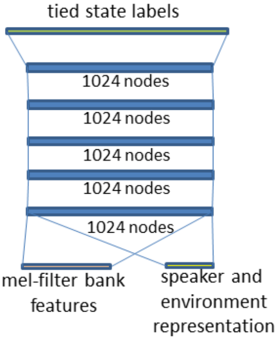
Joint Speaker-Environment Representation

Experiments: acoustic model training

- Two different noise corrupted databases
  - i-vector extraction (done)
  - Acoustic model training
    - 84 speakers
    - (8 noise types + clean) at random SNRs
    - Multi-condition training set (the same size of clean set)

Joint Speaker-Environment Representation

The acoustic model, revisited



- Experiments were conducted on a corrupted WSJ database
  - Evaluation set
    - Corrupted *eval92, dev93, eval93* 5k closed vocabulary test sets
    - The same 8 noise types at random SNRs
    - Trigram language model was used in decoding

Joint Speaker-Environment Representation

Experiments: evaluation

	dev93	eval92	eval93
multi-condition	14.08	8.31	11.14
i-vector (25)	13.90	**7.73**	11.40
i-vector (100)	14.38	8.09	11.22
MTL-MSE-JSER (60)	13.72	8.07	11.06
MTL-CE-JSER (60)	**13.34**	8.37	**9.89**
JTL-CE-JSER (60)	15.36	9.47	11.89

Table: Word error rates for various speaker and environment representations. The number in brackets is the dimensionality of the representation.

Joint Speaker-Environment Representation

Analysis

- MTL-MSE-JSER outperforms the 100-dimensional baseline and multi-condition baseline in all 3 test sets
- MTL-CE-JSER is even better on *dev93* and *eval93*
- MTL-CE-JSER has much better WERs on *dev93* and *eval93* than 25-dim i-vector $\Longrightarrow$ the best in terms of averaged WER:
  - MTL-CE-JSER: 10.53%
  - 25-dim baseline: 11.01%
- JTL-CE-JSER causes degradation on all test set

Joint Speaker-Environment Representation

MTL training results, revisited

Multi-Task Learning	Speaker		Environment	
	Train	CV	Train	CV
MTL-MSE-JSER (60)	0.0501	0.0633	0.0931	0.1337
MTL-CE-JSER (60)	99.28	97.39	93.94	89.50
Joint-Task Learning	Spk. × Env.			
	Train		CV	
JTL-CE-JSER (60)	93.02		80.62	

Table: Speaker and noise classification performance of JSER-DNNs. For MTL-MSE-JSER, the numbers are MSE values and for the rest they are classification accuracies in percentage. The number in brackets is the dimensionality of the bottleneck layer.

- Presented 3 novel methods for training discriminative joint speaker-environment representations from i-vectors
  - Investigated multi-task learning to learn the mapping from **noisy utterance i-vectors** to:
    - **Clean speaker i-vectors** and **pure noise i-vectors** (MSE)
    - **Speaker labels** and **noise labels** (CE)
    - Joint **speaker-noise** labels (CE)
- The representations are the activation of the linear bottleneck layer
- Appending representations at the input of acoustic model $\Longrightarrow$ promising (except JTL-CE-JSER)

- Explore additional auxiliary tasks
- Application to noise robust speaker verification
- Address the issue: in some settings, the frame accuracy has a huge gain, but it does not translate into WER (. . . may try an end-to-end NN?)

Thank you!

Q & A

Joint Speaker-Environment Representation

Why i-vectors?

- Given a Gaussian Mixture Model (GMM), the corresponding speaker-specific mean super-vector $M(s)$, for speaker s, can be approximated as:

$$M(s) = m + Tw(s)$$

- m is the mean super-vector from the GMM-UBM
- T is the low-rank total variability matrix
- $w(s)$ is the low-dimensional i-vector for speaker s

Joint Speaker-Environment Representation

Experiments: setup

- Experiments were conducted on a corrupted WSJ database
  - 84 speaker WSJ0 subset for training the acoustic model
  - WSJ0 + WSJ1 for training the Joint Speaker and Environment Representation (JSER) transforms
  - 8 different types of noise were added to the clean waveforms at different SNRs
    - Restaurant, street, supermarket, food-court, living room, mall, taxi and gym
    - Noise recording was about half an hour long
    - Mixed with a random noise segment equal to the duration of the waveform

Joint Speaker-Environment Representation

Experiments: i-vector extraction

- Two different noise corrupted databases
  - i-vector extraction
    - 283 speakers
    - 8 different SNRs: 5dB to 20dB in steps of 2dB
    - 8 noise types $\times$ 8 SNRs $\Longrightarrow$ 64 times the size of clean database
    - Pure noise i-vectors: long noise recordings randomly segmented into many 20-second chunks and MFCC features were extracted
    - For each utterance i: $\{w(i), w(s_i), w(n_i)\}$ and $\{w(i), s_i, n_i\}$.
  - Acoustic model training (TBC)

Joint Speaker-Environment Representation

Experiments: acoustic model training

- Two different noise corrupted databases
  - i-vector extraction (done)
  - Acoustic model training
    - 84 speakers
    - (8 noise types + clean) at random SNRs between 10dB and 20dB
    - Multi-condition training set (the same scale of clean set)
    - 13 MFCC, Δ and $\Delta\Delta$ features normalized by mean and variance over the utterance
    - 11 frames of temporal context
    - Tied-state labels are from MMI trained GMM-HMM

- Experiments were conducted on a corrupted WSJ database
  - Evaluation set
    - Corrupted *eval92, dev93, eval93* 5k closed vocabulary test sets
    - The same 8 noise types at random SNR from 5 dB to 20 dB
    - Trigram language model was used in decoding

Joint Speaker-Environment Representation

Analysis

- Utterance-level i-vector adaptation instead of speaker-level adaptation
- 25-dimensional i-vector setting is better than 100-dimensional one