TOSHIBA

Leading Innovation >>>

Speaker adaptive training in deep neural networks using speaker dependent bottleneck features

Rama Doddipatla [rama.doddipatla@crl.toshiba.co.uk]

Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK.

Overview

- Improve the DNN acoustic model using **speaker adaptive training** (SAT).
  - Transforming input features with VTLN, CMLLR or appending speaker information in the form of speaker codes fall into the frame work of SAT.
- This work focuses on **tuning the weights of DNN** to implement SAT.
- **Proposed approach** follows a two-stage architecture to implement SAT.
  - **Stage-1** is a bottleneck (BN) feature extractor, where the weights of the BN layer are adjusted using speaker specific data while keeping the weights in rest of the layers fixed.
  - **Stage-2** is the SAT-DNN model trained using the **speaker dependent bottleneck** (SDBN) features from stage-1.
- **Unsupervised adaptation** using SAT on Aurora4 task provides:
  - 8.6% WERR* on DNN trained with Mel filter-bank (FBANK) features.
  - 10.3% WERR* on DNN trained with CMLLR-FBANK.
- **Supervised adaptation** using one minute of audio improves the performance when compared with the performance of baseline DNN.

*WERR - Relative word error rate

Experimental Setup

- **Corpus**: Aurora4
  - Train: 7138 Utterances, 83 speakers, multi-condition.
  - Test: 4620 Utterances, 8 speakers per test condition, 14 test conditions.
  - Each test speaker has 40 utterances (approx. 5 min of audio).
- Mel-filter bank features - 40 dimensions (No-LDA)
- Bi-gram language model (Vocabulary - 5K).
- **Conventional DNN**: 2048 (hid-dim) x 7 (layers)
- **Bottleneck DNN**: 512 (hid-dim) x 3 (layers), **BN-dim** : 75
- **SI/SAT-DNN**: 2048 (hid-dim) x 3 (layers)
- **D-vector** is obtained by averaging the BN features of a DNN trained using speaker labels as targets over an utterance.
  - **DNN** : 1024 (hid-dim) x 2 (layers), **BN-dim** : 40
- CMLLR transforms are estimated from the SAT-GMM model.

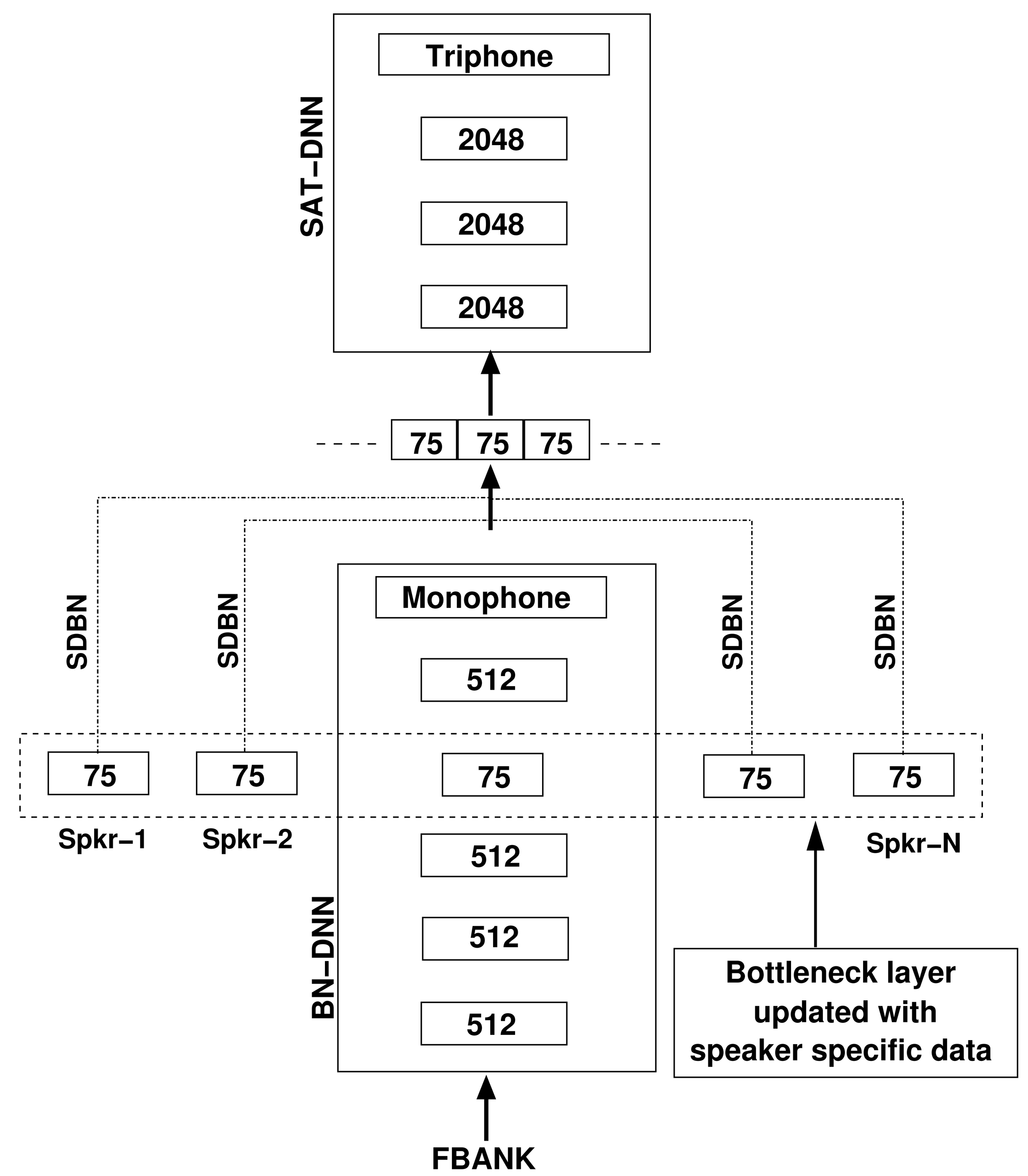
Results: Conventional Vs Two-stage DNN

%WER	Conventional	Two-stage
FBANK	14.6	14.5
+ D-vec	13.9	13.9
+ CMLLR	12.6	12.6
+CMLLR + D-vec	12.3	11.9

- Two-stage DNN seems to perform similar to the conventional DNN.
- Appending speaker information in the form of D-vectors or transforming the features with CMLLR improve the performance.

Proposed approach: SAT using a two-stage DNN

The proposed approach uses a **two-stage architecture** as illustrated below:



Steps in Training

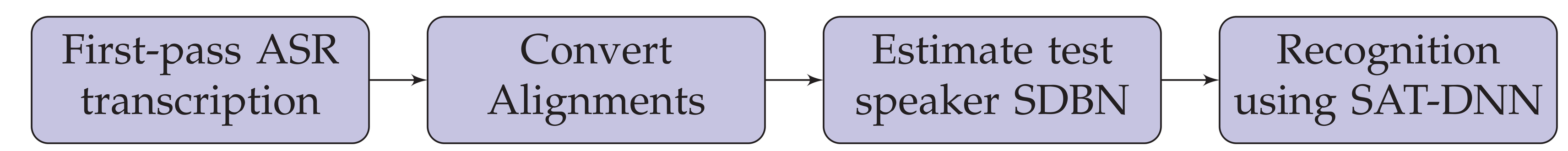
Stage-1 (BN-DNN)

- Train a bottleneck(BN) DNN using monophone targets and FBANK features.
- **SDBN**: Update the weights of the BN layer using speaker specific data, keeping the weights in the rest of the layers fixed.

Stage-2 (SAT-DNN)

- Using speaker dependent (SD) BN features, train the second stage DNN using triphone targets.
- Since the input features are speaker dependent, the second stage model is trained in the SAT frame work.

Steps in Recognition



- Obtain first pass-transcription using the speaker independent (SI) model.
- Tune weights of the BN layer using data from the test speaker and alignments from previous step to derive SDBN features.
- Perform recognition using the SAT-DNN model.

Using monophone alignments reduces the problem of data sparsity and improves robustness to transcription errors.

Results: Appending D-vectors

%WER	FBANK	BN
+ D-vec	13.9	13.8
+ CMLLR + D-vec	11.9	12.0

- D-vectors seem to provide similar gains irrespective of the position where they are introduced into the training, i.e either with FBANK or with BN features.
- Appending D-vectors, both with FBANK and BN features did not provide any gains in performance.

Results: Unsupervised adaptation of the proposed SAT

%WER	Baseline	+ SAT-DNN	%WERR
FBANK	14.5	13.2	8.9
+ D-vec	13.9	12.7	8.6
+ CMLLR	12.6	11.3	10.3
+ CMLLR + D-vec	11.9	11.2	5.9

- The proposed SAT consistently improves the performance over the baseline.
- Best gain in performance is obtained when SAT is applied on top of DNN trained with CMLLR-FBANK features.
- Performance of SAT saturates when applied on top of CMLLR-FBANK+D-vector system.

Results: Supervised adaptation of the proposed SAT

%WER	Baseline	+10	+20	+30	+40
FBANK	14.5	13.4	12.7	12.3	11.9
+ D-vec	13.9	13.1	12.1	11.9	11.6
+ CMLLR	12.6	11.5	11.1	10.8	10.4
+ CMLLR + D-vec	11.9	11.4	10.8	10.5	10.4

- Using as little as 10 utterances (approx. 1 min of audio) for supervised adaptation already improves the performance when compared with baseline.
- Performance improves as the data from a specific speaker increases.
- Less data is required to achieve similar performance of FBANK by applying CMLLR or appending D-vectors.

Conclusion

- Presented an approach to perform SAT in DNNs using a 2-stage architecture.
  - First stage is a BN-DNN, used for deriving SDBN features.
  - Second stage is the SAT-DNN model trained using SDBN features.
- Unsupervised adaptation using SAT on CMLLR-FBANK DNN provided the best performance (10.3% WERR).
- Supervised adaptation using one minute of audio improved the performance when compared with the performance of baseline DNN.