DOMAIN AND SPEAKER ADAPTATION FOR CORTANA SPEECH RECOGNITION

Yong Zhao, Jinyu Li, Shixiong Zhang, Liping Chen, and Yifan Gong

Microsoft AI and Research

Introduction

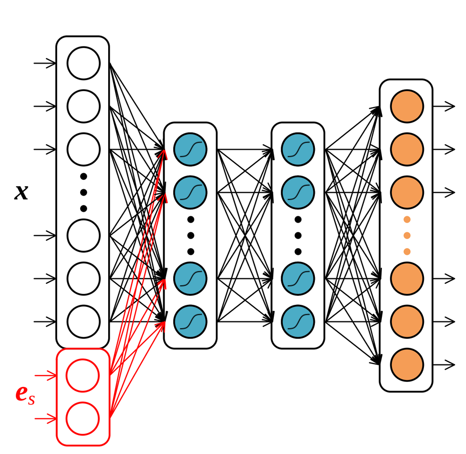
- Voice assistant represents one of the most popular and important scenarios for speech recognition
- We propose a deep adaptation framework that adapts a generic acoustic model towards Cortana assistant
  - Anchor word speaker embedding
  - Adapting multiple layers in both weight matrices and biases
  - Prior interpolation
- The proposed system yields 32% WERR over the SI baseline

Prior Works and Motivation

I-vector Based Adaptation on 1-st layer

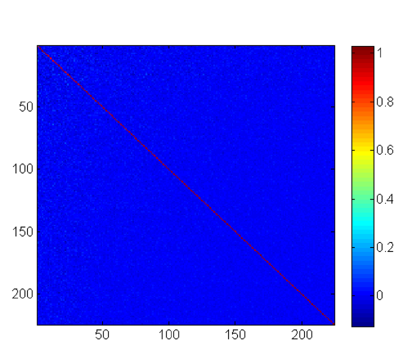
$$z_s^1 = W^1 x + V e_s + b^1$$

- Pros
  - No need to retrain model parameters
  - No need to store speaker profile, if i-vector is estimated per utterance
- Cons
  - I-vector estimation over a short utterance is very noisy
  - Recognition starts until the utterance is finished



Low-Rank Plus Diagonal (LRPD) Adaptation

- A transformation matrix is very close to an identity matrix



- LRPD adaptation

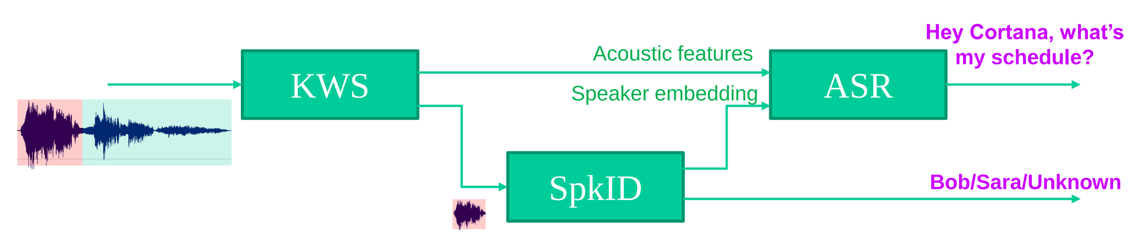
  $$W_{s,k\times k} \approx I_{s,k\times k} + P_{s,k\times c} Q_{s,c\times k}$$

  - # of SD parameters: $2ck + k$
- Extended LRPD adaptation

  $$W_{s,k\times k} \approx I_{s,k\times k} + P_{k\times c} T_{s,c\times c} Q_{c\times k}$$

  - # of SD parameters: $c^2 + c$

Adaptation Using Anchor Embedding



- Given an L-layer SI DNN

  $$h^l = \sigma(z^l) = \sigma(W^l h^{l-1} + b^l)$$

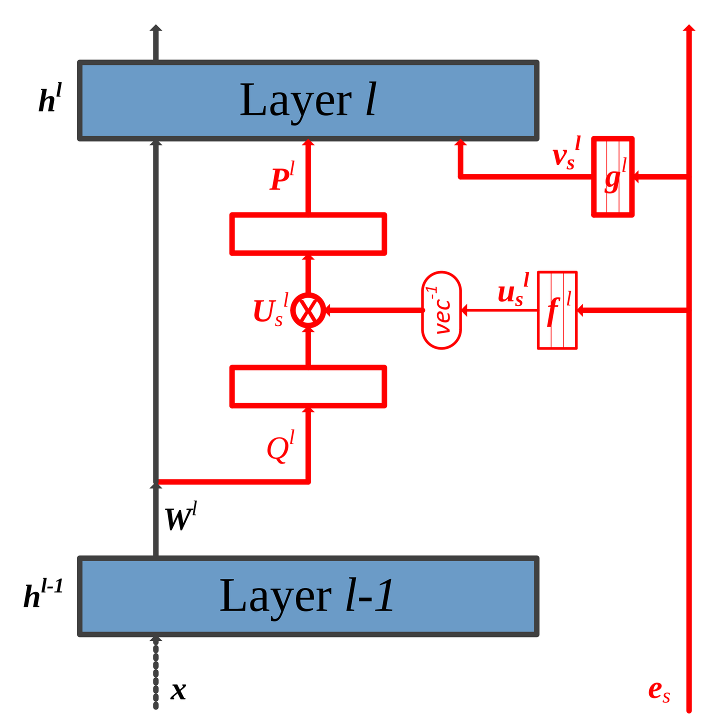
- Deeply adapts multiple layers in both weight matrices and biases

  $$z_s^l = \underbrace{\left(I + P^l U_s^l Q^l\right) W^l}_{\text{adapting weight matrix}} h^{l-1} + \underbrace{v_s^l}_{\text{adapting bias}} + b^l$$

- Extracts speaker embedding e_s from anchor segments: i-vector and d-vector
- Maps e_s to the SD vector u_s^l and v_s^l through auxiliary networks, and reshapes u_s^l into a matrix U_s^l

  $$U_s^l = vec^{-1}(u_s^l) = vec^{-1}(f^l(e_s))$$
  $$v_s^l = g^l(e_s)$$

- Absorbs the transformations of both the weight matrix and bias into one term



Adaptation with Prior Interpolation

- Biased distribution of Hey Cortana data: most queries begin with "Hey Cortana"
- Conceptually, prior distribution should be updated along with the DNN model adaptation

  $$p(x_t|q_t) \propto p(q_t|x_t)/p(q_t)$$

- Prior interpolation between the source domain and the target domain

  $$\hat{p}(q_t) = (1-\rho)\tilde{p}(q_t) + \rho p^{SI}(q_t)$$

- Analogous to Kullback-Leibler (KL) regularization

Experiments and Results

Experimental Data

- Train: 3400hr multi-style US English data
- Dev: 220hr Cortana desktop data that begin with "Hey Cortana" within the 3400hr data
- Eval: Hey Cortana desktop test set

Experimental Setup

- SI DNN
  - 8 layers: 2,048*6:4,096:9,801
  - Input: 11 frames of 80-dim $\mathrm{LFB} + \Delta + \Delta^2$
- Anchor embedding (100 dim)
  - I-vector: 39-dim MFCC, 512 mixture UBM
  - D-vector: 5-layer DNN, 1024*4:100:8398
- Auxiliary networks
  - Two sigmoid layers each with 100 nodes followed by a linear layer
  - Reshape u_s^l into U_s^l of size 10×10

Baseline Performance

Table 1: WER (%) of the SI and SAT models on the Hey Cortana desktop test set.

Model	WER (%)
3400hr SI	16.36
220hr SI	20.20
220hr SAT, L1 ivec bias	17.49
220hr SAT, L1 dvec bias	16.64

Speaker Adaptation Using Anchor Embedding

- Adapting multiple layers (L1-*) outperforms adapting a single layer
- Adapting the weight matrix only of multiple layers suffices to yield optimal performance

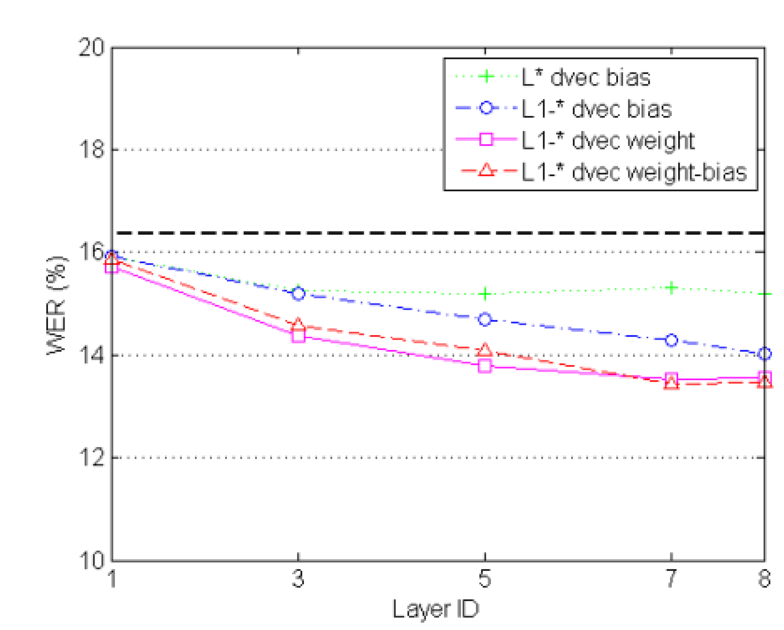


Fig. 2: WERs (%) for adapting biases and weight matrices of single layer and multiple layers from bottom to top using the anchor d-vectors. The dashed level line is the 3400hr SI baseline.

Adaptation with Prior Interpolation

- Directly adapting the softmax layer (L8) with prior interpolation ($\rho = 0.5$) yields 20% WERR
- Adapting the softmax layer without changing the priors ($\rho = 1$) yields only 3% WERR
- Adapting the priors alone yields 4% WERR

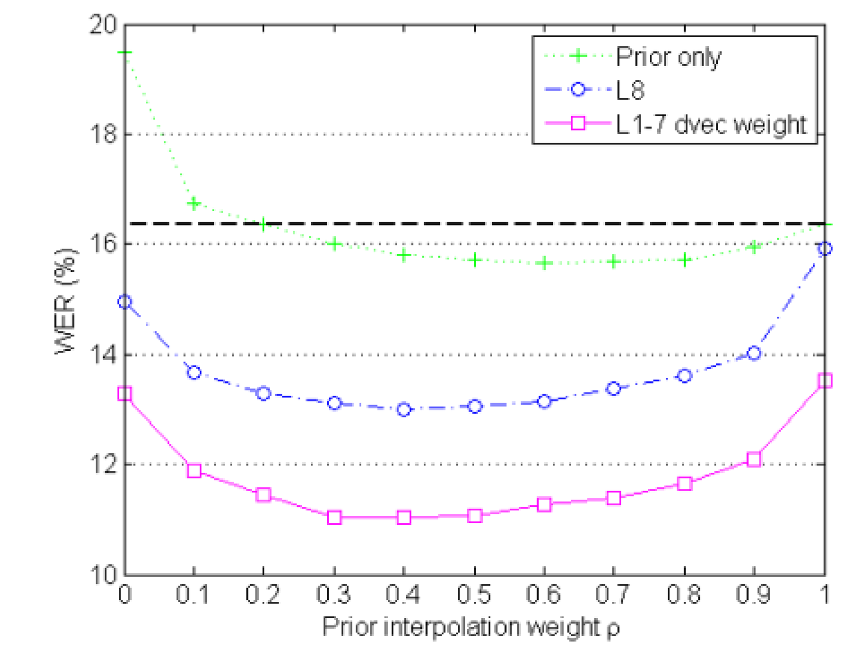


Fig. 3: WERs (%) against prior interpolation weights ρ for three adaptation models. The dashed level line is the SI baseline.

Anchor-Based Adaptation with Prior Interpolation

- Anchor-based model (L1-7 dvec weight, $\rho = 0.5$) yields 32% WERR
- Adaptation using the anchor d-vector outperforms using the anchor i-vector

Table 4: WERs (%) for the anchor-based speaker adaptation using i-vector and d-vector embeddings with prior interpolation $\rho = 0.5$.

Model	ivec	dvec
L1-7 weight	11.66	11.06
L8 + L1-7 weight	11.42	11.02

Analysis of Recognition Results

Table 2: Top 10 word count changes from the SI model to the model with layer L8 updated ($\rho = 1$) on Hey Cortana desktop test set.

	Ref	SI	L8 ($\rho = 1$)	L8 ($\rho = 0.5$)
WER (%)	–	16.36	15.92	13.06
have	61	575	111	117
hello	9	492	42	50
what	591	890	1067	764
hey	6212	6132	6306	6230
caught	0	1	102	23
nah	0	42	133	4
ne	0	15	98	3
the	1081	1203	1279	1184
what's	498	561	609	562
anna	0	13	55	5