Using Contextual Information in Joint Factor Eigenspace MLLR for Speech Recognition in Diverse Scenarios

Oscar Saz and Thomas Hain
Speech and Hearing Research Group, The University of Sheffield, Sheffield, UK
O.SazTorralba@sheffield.ac.uk,T.Hain@sheffield.ac.uk

Introduction

This work[1] presents a new approach for rapid adaptation in the presence of highly diverse scenarios that takes advantage of contextual information describing the input signals.

- Context information and metadata are common for augmenting audio–visual data:
  - Describing the speakers, type of recording, background condition, etc
  - In multiple forms: Textual description, tags, etc
- **Research Question: Can we achieve fast and robust adaptation using only metadata information?**
  - Modelling of metadata in an Eigenspace MLLR framework
  - Use of factorisation to separate influence of speaker and background information

Eigenspace MLLR

- Eigenspace MLLR was proposed to model speaker variability in ASR in cases when little adaptation data exists[a]
  - PCA is used to learn eigenvectors (W_m) from MLLR transformations that characterise the acoustic space
  - Maximum Likelihood Eigen Decomposition (MLED) is used to estimate the combination coefficients (α_m) to obtain the best MLLR estimation ($\hat{W}$) for an input speech signal
  - Requires an initial decoding to obtain an estimated transcript for MLED

$$\hat{W} = \sum_{m=0}^{M} \alpha_m \cdot W_m$$

- Further work has also applied Eigenspace MLLR for capturing background variability[b]
- Similar to CAT adaptation[c] but using an eigenspace approach

[a]R. Kuhn et al: "Eigenvoices for speaker adaptation", ICSLP, Sydney, Australia, 1998.
[b]K. Chen et al: "Fast speaker adaptation using eigenspace–based Maximum Likelihood Linear Regression", ICSLP, Beijing, China, 2000.
[c]M.J.F. Gales, "Transformation smoothing for speaker and environmental adaptation", Eurospeech, Rhodes, Greece, 1998.

JFEMLLR: Joint Factor EigenMLLR

- **JFEMLLR expands EigenMLLR with a joint factor approach**, using three terms
  - **A global mean** (W_0) of the training acoustic space
  - **Speaker–based eigenvectors** (W_p^{spk})
  - **Background–based eigenvectors** (W_n^{bgd})

$$\hat{W} = W_0 + \sum_{n=1}^{N} \beta_n \cdot W_n^{bgd} + \sum_{p=1}^{P} \gamma_p \cdot W_p^{spk}$$

- Training setup:
  - Do PCA on background–dependent MLLR transformations to learn W_n^{bgd}
  - Substract $\sum_{n=1}^{N} \beta_n \cdot W_n^{bgd}$ from the training utterances
  - Do PCA on speaker–dependent MLLR transformations to learn W_p^{spk}
- Decoding setup:
  - Use MLED to estimate β_n and γ_p separately on the input utterance
  - Apply estimated MLLR transform $\hat{W}$ to the utterance

Introducing contextual information

- Contextual information will be defined as a set of **discrete tags used to describe properties of a speech signal**.
  - A tag cloud of T tags ($Tags = \{Tag^1, Tag^2, ..., Tag^T\}$) can be set for any input signal, describing different characteristics of the audio.
  - The distribution of the eigenspace coefficients (ϕ) for a given tag (Tag^t) is modelled as a GMM learnt from train data

$$P(\phi|Tag^t) = \sum_{g=1}^{G} c_g^{Tag^t} \cdot N(\phi; \mu_g^{Tag^t}, \Sigma_g^{Tag^t})$$

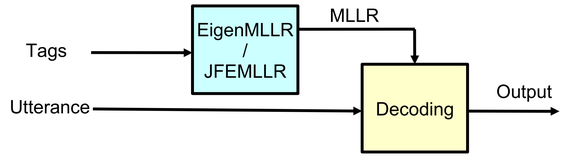
  - For an input test speech signal with a set of $Tags$, the best set of coefficients will be given by the probability

$$\hat{\phi} = \arg\max_{\phi} P(Tags|\phi) \approx \arg\max_{\phi} \sum_{t=1}^{T} P(\phi|Tag^t)$$

  - Maximising this probability, leads to an equation system from where to calculate ϕ

$$\sum_{t=1}^{T}\sum_{g=1}^{G}\left[(\Sigma_g^{Tag^t})^{-1}\mu_g^{Tag^t}\right] = \left[\sum_{t=1}^{T}\sum_{g=1}^{G}(\Sigma_g^{Tag^t})^{-1}\right] \cdot \phi$$

- This provides a **1–step system, opposed to common 2–step adaptation systems**



Experimental setup

- **WSJCAM0**: British English read speech
- We created an artificial dataset with ***Diverse* backgrounds**:
  - Channel: Close–talk microphone (50%) or table–top microphone (50%).
  - Noise: Clean (33%) or music (33%), divided equally in orchestral and popular contemporary, or noise (33%), divided equally in traffic, outdoors, cocktail party and applause.
  - Signal–to–Noise Ratio (SNR): Uniform distribution from 5 to 15dB if noise is present.
- **Metadata available**:
  - Background–based: **Channel, noise and SNR**
  - Speaker–based: **Age, gender and accent**

Baseline system

- ASR setup: 39 PLP features (static+$\triangle$+$\triangle\triangle$), HMM-GMM triphone models, ML training, standard WSJ language models, evaluation in WER.

Train	Test	5K set	20K set	Total
Clean	Clean	5.8%	13.0%	9.5%
Clean	Diverse	27.0%	39.1%	33.1%
Diverse	Diverse	14.9%	25.5%	20.3%

Results

2–pass adaptation systems

- Baseline adaptation results with MLLR and Eigenspace MLLR

Adaptation	5K set	20K set	Total
Unsupervised speaker MLLR	19.3%	29.8%	24.6%
Supervised background MLLR	18.5%	29.9%	24.2%
Eigenspace MLLR (30 eigenbasis)	17.0%	28.5%	22.8%

- Joint Factor Eigenspace MLLR results

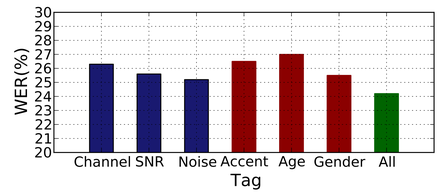
Background	Speaker	5K set	20K set	Total
15 basis	–	18.1%	29.9%	24.1%
–	15 basis	18.8%	31.1%	25.0%
15 basis	15 basis	17.3%	28.6%	23.0%

Tag–based 1–pass adaptation systems

Background	Speaker	5K set	20K set	Total
Eigenspace MLLR (30 basis)		18.9%	30.8%	24.9%
JFEMLLR (15)	–	18.6%	30.4%	24.6%
–	JFEMLLR (15)	19.8%	32.0%	26.0%
JFEMLLR (15)	JFEMLLR (15)	18.2%	30.1%	24.2%

Influence of limiting the amount of tags

- Results using single tags as input show the information carried by each tag



Influence of the amount of training data

- New situation with only **20% of training data** for the *Diverse* condition

Condition	5K set	20K set	Total
Diverse ML models	19.7%	32.8%	26.3%
Supervised background MLLR	19.3%	31.3%	25.4%
JFEMLLR with tag models	18.1%	30.2%	24.2%

Conclusions

- Framework for joint factorisation with Eigenspace MLLR (JFEMLLR) has been proposed
- **Instant adaptation achieved using metadata**
  - JFEMLLR deals separately with speaker and background variability
  - Robust to metadata noise and data sparsity
- Future work
  - Use with naturally occurring metadata: text, images, etc

[1]Supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).