Asynchronous Tracking of Background Noise for Speech Recognition and Content Analysis

Oscar Saz, Mortaza Doulaty,
Thomas Hain

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

The University Of Sheffield.

University of Sheffield

May 2014

Overview

Natural Speech Technology

- Introduction
- Asynchronous adaptation and factorisation
- Applications in speech recognition
- Applications in content analysis
- Conclusions

Introduction: NST perspective

Natural Speech Technology

- Canonical acoustic modelling aims to develop modular and robust models for speech recognition in real scenarios
  - Complex and non–stationary acoustic backgrounds are a major difficulty that appear in certain tasks
    - Broadcasts and media data
- Going beyond ASR, a deeper understanding of acoustic backgrounds has further applications in natural speech technology

Asynchronous adaptation and factorisation

Natural Speech Technology

- In [Saz and Hain, 2013], we proposed an asynchronous use of CMLLR transformations trained from multiple backgrounds
  - Combined with speaker transformations: Factorisation
  - Similar approach to [Seltzer and Acero, 2011]
- Mathematically expressed as:

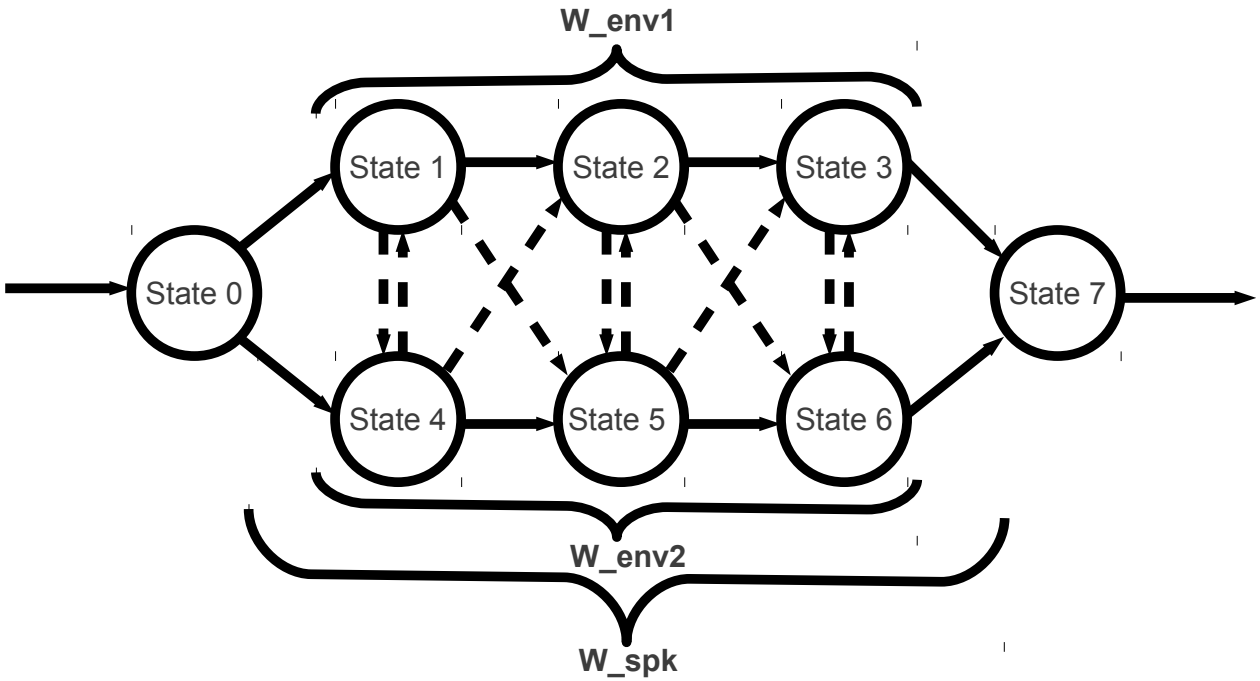
$$y(t) = A_{spk}(A_{bck}(t)x(t) + b_{bck}(t)) + b_{spk}$$

$$W_{bck}(t) = \begin{cases} W_{bck1} & \text{if } x(t) \in bck1 \\ W_{bck2} & \text{if } x(t) \in bck2 \end{cases}$$

Asynchronous adaptation and factorisation

Natural Speech Technology

- Modifying the HMM topology provides an implementation where the ML criterion is used to find the best path



Evaluation in ASR: Artificial data

Natural Speech Technology

- WSJCAM0
  - Controlled scenario/read speech
- Diverse backgrounds
  - Used in [Saz and Hain, 2014] for evaluation of the use of metadata in ASR (see poster)

Clean	33%		
Music	33%	Orchestral	50%
		Pop–contemporary	50%
Noise	33%	Traffic	25%
		Restaurant	25%
		Applause	25%
		Outdoors	25%

Evaluation in ASR: Artificial data

Natural Speech Technology

- Results with clean models on recognition of diverse data
  - Evaluation on 5K and 20K standard WSJ development sets

Condition	WER	Improv
Baseline (HMM-ML-PLP)	33.1%	-
Speaker adaptation	24.6%	25.7%

Evaluation in ASR: Artificial data

Natural Speech Technology

- Results with clean models on recognition of diverse data
  - Evaluation on 5K and 20K standard WSJ development sets

Condition	WER	Improv
Baseline (HMM-ML-PLP)	33.1%	-
Speaker adaptation	24.6%	25.7%
Synchronous background adaptation	28.2%	14.8%
+ Speaker adaptation	23.4%	29.3%

Evaluation in ASR: Artificial data

Natural Speech Technology

- Results with clean models on recognition of diverse data
  - Evaluation on 5K and 20K standard WSJ development sets

Condition	WER	Improv
Baseline (HMM-ML-PLP)	33.1%	-
Speaker adaptation	24.6%	25.7%
Synchronous background adaptation	28.2%	14.8%
+ Speaker adaptation	23.4%	29.3%
Asynchronous background adaptation	24.3%	26.5%
+ Speaker adaptation	20.7%	37.5%

Evaluation in ASR: Media data

Natural Speech Technology

- BBC media
  - Task based on actual broadcasts from one week in 2008, covering all programmes
- System description
  - 100 hours of training data
  - 20 hours of test data
  - Noisy reference based on subtitles
  - Language model including large subtitling data

Evaluation in ASR: Media data

Natural Speech Technology

- Results in a 6–hour subset of the test set
  - Guarantee reference subtitles match the ground truth

Condition	WER	Improv
Baseline (HMM-ML-PLP-HLDA)	29.9%	-
Baseline (HMM-ML-Bottlenecks)	24.0%	-
Speaker adaptation	27.9%	6.7%

Evaluation in ASR: Media data

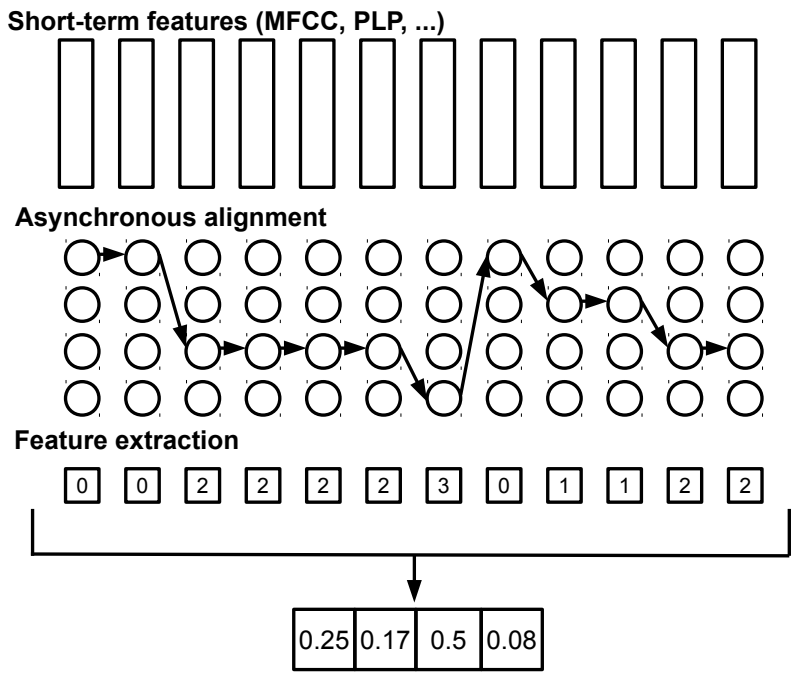
Natural Speech Technology

- Results in a 6–hour subset of the test set
  - Guarantee reference subtitles match the ground truth

Condition	WER	Improv
Baseline (HMM-ML-PLP-HLDA)	29.9%	-
Baseline (HMM-ML-Bottlenecks)	24.0%	-
Speaker adaptation	27.9%	6.7%
Asynchronous background adaptation	28.9%	3.3%
+ Speaker adaptation	27.4%	8.0%

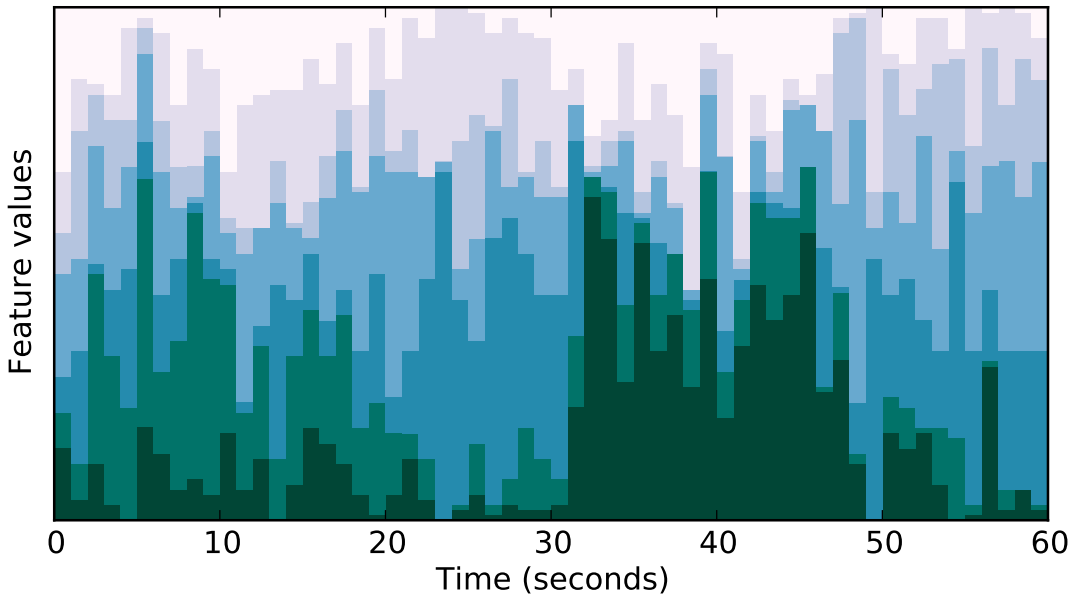
Tracking of background

- Asynchronous background adaptation yields the sequence of most likely background noise for each frame
  - Convert this to a set of long–term background–tracking features



Tracking of background

- Features that can be used to visualise the background conditions of audio data
  - Full visualisation in the demo



Tracking of background

Natural Speech Technology

- We evaluated these features in the automatic identification of the genre of broadcast shows
  - Compared short–term spectral features to our long–term background features with different classifiers
- State–of–the–art systems reported in [Ekenel and Semela, 2013] indicate accuracies of 50% to 80% using MFCC–GMM systems.
  - More complex sets based on heterogeneous data report lower results.

Evaluation in genre identification

Natural Speech Technology

- BBC broadcasts, with 8 genres:
  - Advice: Consumer and DIY shows.
  - Children's: Including cartoons and educational shows.
  - Comedy: Sit–coms and light entertainment shows.
  - Competition: Quiz shows and other contest shows.
  - Documentary: Including fly–on–the–wall shows.
  - Drama: Soap operas and other serialised dramas.
  - Events: Live events, sports and concerts.
  - News: Broadcast news and current affair shows.
- 285 shows for training and 47 shows for testing

Evaluation in genre identification

Natural Speech Technology

- Results in identification accuracy

Features	Classifier	Accuracy
Baseline (PLP)	GMM	62%
Background	GMM	72%

- GMM size evaluated from 8 to 1024 gaussians
- PLP require more gaussians (1024) than background features (64)

Evaluation in genre identification

Natural Speech Technology

- Results in identification accuracy

Features	Classifier	Accuracy
Baseline (PLP)	GMM	62%
Background	GMM	72%
	HMM	79%
	SVM	81%

- HMM with 8 states and 32 gaussians per state
- SVM based on MAP–adapted mean supervectors of 16 gaussians

Evaluation in genre identification

Natural Speech Technology

- Results in identification accuracy

Features	Classifier	Accuracy
Baseline (PLP)	GMM	62%
Background	GMM	72%
	HMM	79%
	SVM	81%
	HMM & SVM	83%

- System combination from the best HMM and SVM systems
- Decision based on the best confidence measure provided

Conclusions

Natural Speech Technology

- Framework to deal with non–stationary and complex backgrounds in speech recognition
  - Allows use on factorisation approaches
  - Performance gains in controlled and real scenarios
- Integrates background modelling into the recognition process
  - No need for separate systems
  - Optimised for best recognition performance via ML
  - Provides further functionality

References

1 Saz, O., and Hain, T., "Asynchronous Factorisation of Speaker and Background with Feature Transforms in Speech Recognition", Interspeech 2013, Lyon, France.

2 Seltzer, M. L. and Acero, A., "Separating Speaker and Environmental Variability Using Factored Transforms", Interspeech 2011, Florence, Italy.

3 Saz O., and Hain, T., "Using contextual information in Joint Factor Eigenspace MLLR for speech recognition in diverse scenarios", ICASSP 2014, Florence, Italy

4 Ekenel, H.K., and Semela, T., "Multimodal Genre Classification of TV Programs and YouTube Videos", Multimedia Tools and Applications, 63, pp. 547567, 2013.

Asynchronous Tracking of Background Noise for Speech Recognition and Content Analysis

Oscar Saz, Mortaza Doulaty,
Thomas Hain

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

The University Of Sheffield.

University of Sheffield

May 2014