THE TUM SYSTEM FOR THE REVERB CHALLENGE: RECOGNITION OF REVERBERATED SPEECH USING MULTI-CHANNEL CORRELATION SHAPING ENHANCEMENT AND BLSTM RECURRENT NEURAL NETWORKS

Jürgen T. Geiger, Erik Marchi, Björn Schuller [1] and Gerhard Rigoll
Institute for Human-Machine Communication, Technische Universität München, Germany
[1] also with the Department of Computing, Imperial College London, UK
geiger@tum.de

ICASSP 2014 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING

Imperial College London

Introduction

- Multi-channel dereverberation with Correlation Shaping (CS)
- Bidirectional Long Short-Term Memory (LSTM) RNNs for phoneme prediction
- GMM-LSTM double-stream decoding

Proposed System

- **Correlation Shaping**
  - Modifies the correlation structure of the input
  - Reduces long-term correlation in the LP residual

- Multi-input single-output linear filter

$$y(n) = \sum_{m=0}^{M-1} \boldsymbol{g}_m^T(n)\boldsymbol{x}_m(n)$$

- Minimization of weighted MSE

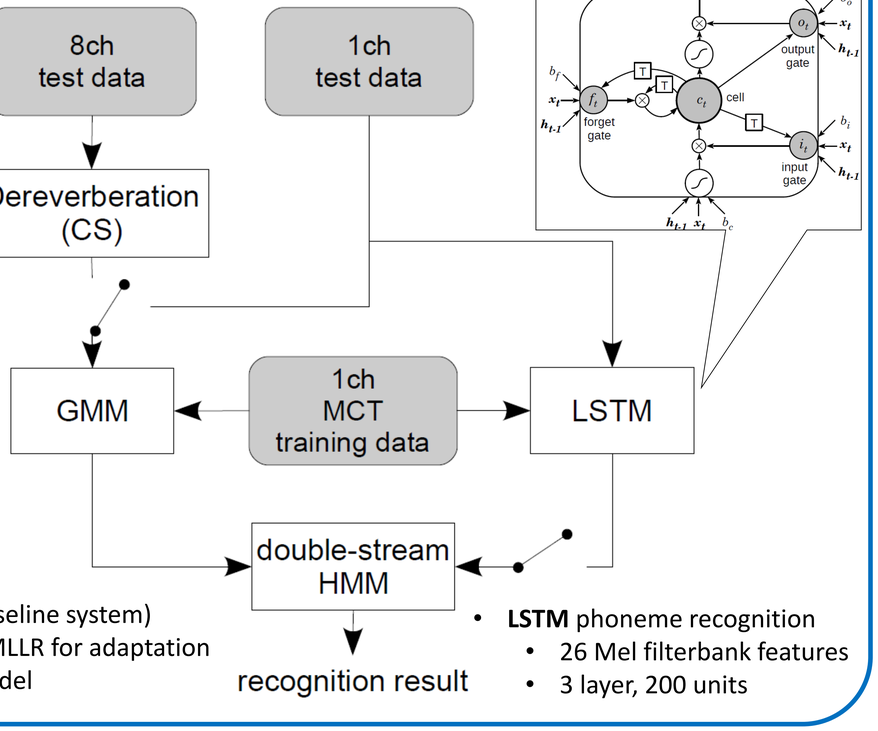
$$e(\tau) = W(\tau)\Big(R_{yy}(\tau) - R_{dd}(\tau)\Big)^2$$

- Filter update equation

$$g_m(l, n+1) = g_m(l, n) - \mu\nabla'_m(l)$$

- Gradient

$$\nabla'_m(l) = \frac{\nabla_m(l)}{\sqrt{\sum_m \sum_l \nabla_m^2(l)}}$$



- **Kaldi GMM** (Similar to baseline system)
  - Basis feature space MLLR for adaptation
  - Trigram language model
- **LSTM** phoneme recognition
  - 26 Mel filterbank features
  - 3 layer, 200 units

Experiments and Results

- Results on test set
- 8 different recording conditions
- 1- and 8-channel audio processing
- CS effectively reduces long-term reverberation energy for higher reverberation times

LSTM training software: http://currennt.sf.net

Kaldi GMM					+CS		+LSTM		+CS, +LSTM	
Adapt	MCT	LM	SIM	REAL	SIM	REAL	SIM	REAL	SIM	REAL
-	-	bg	49.95	88.50	32.88	70.98	36.80	79.81	23.92	62.06
-	•	bg	27.53	53.78	22.06	40.37	20.79	48.97	17.63	37.89
•	•	bg	22.36	46.14	17.75	34.06	17.77	43.83	14.82	34.39
•	•	tg	17.26	39.76	13.20	28.15	13.75	36.78	11.19	28.13

	SIM DATA							REAL DATA		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
System	Near	far	near	far	near	far		near	far	
Baseline	16.23	18.71	20.50	32.47	24.76	38.88	25.25	50.14	47.57	48.85
CS + Baseline	16.03	17.66	17.37	24.10	19.38	29.26	20.62	38.87	38.79	38.83
Kaldi	10.23	12.26	12.96	23.34	15.25	29.53	17.26	40.56	38.96	39.76
CS + Kaldi	10.86	11.50	10.74	15.62	11.40	19.09	13.20	**28.04**	28.26	28.15
Kaldi + LSTM	**8.32**	9.98	10.63	18.66	12.21	22.67	13.75	36.38	37.17	36.78
CS + Kaldi + LSTM	8.50	**9.66**	**9.40**	**13.70**	**9.64**	**16.26**	**11.19**	28.27	**27.99**	**28.13**

Conclusions

- Multi-channel: CS reduction > 25% WER, LSTM reduction ~15% WER
- Single-channel: LSTM reduction ~7% WER (Real Data), ~20% WER (Sim Data)