CHANNEL SELECTION FOR DISTANT AUTOMATIC SPEECH RECOGNITION

on the CHiME-5 dataset

Hannes Unterholzner, BSc

Supervisor:
Assoc.Prof. Dipl.-Ing. Dr. Franz Pernkopf

Graz, March 14th, 2019

Outline

- Background: CHiME-5 challenge
- Baseline
- Oracle
- Features
- Channel Selection Results
- Conclusion and Future Work

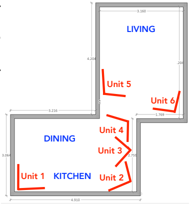
CHiME-5 Challenge and Dataset

Challenge:

- **Topic**: Distant multi-microphone conversational speech recognition in everyday home environments
- **Baseline:** GMM-HMM, DNN-HMM, End-to-End

Baseline	Dev (Kinect)	Dev (Binaural)
GMM-HMM	91.0	71.9
DNN-HMM	**82.5**	48.9
E2E	94.7	67.2

- **Floor plan:** Conventional and open-space apartments (e.g. session S09)



Dataset:

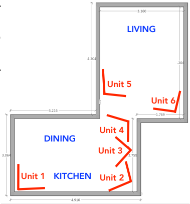
- 20 sessions duration of $\sim$ 2h, 4 participants, three rooms (kitchen, dining, living), 6 Kinect arrays, 4 binaural mic's $\rightarrow \underbrace{(6\times 4)}_{\text{for train/dev/eval}} + \underbrace{(4\times 2)}_{\text{train,dev transcript.}} = 32$ ch.
- **Characteristics:** noise, far-field recordings, simultaneous and spontaneous speech, deviations within/among session/s
- **Simultaneous speech (dev):**

Challenge:

- **Topic**: Distant multi-microphone conversational speech recognition in everyday home environments
- **Baseline:** GMM-HMM, DNN-HMM, End-to-End

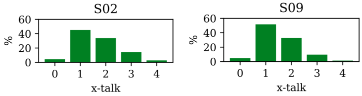
Baseline	Dev (Kinect)	Dev (Binaural)
GMM-HMM	91.0	71.9
DNN-HMM	82.5	48.9
E2E	94.7	67.2

- **Floor plan:** Conventional and open-space apartments (e.g. session S09)

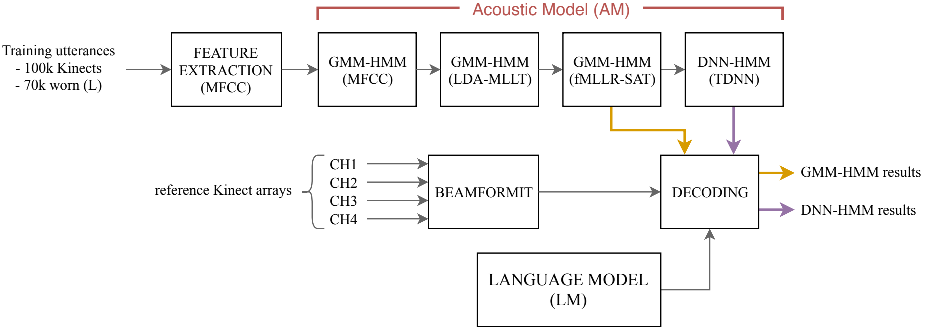


Dataset:

- 20 sessions duration of $\sim$ 2h, 4 participants, three rooms (kitchen, dining, living), 6 Kinect arrays, 4 binaural mic's $\rightarrow \underbrace{(6 \times 4)}_{\text{for train/dev/eval}} + \underbrace{(4 \times 2)}_{\text{train,dev transcript.}} = 32$ ch.
- **Characteristics:** noise, far-field recordings, simultaneous and spontaneous speech, deviations within/among session/s
- **Simultaneous speech (dev):**



DNN-HMM Baseline System



Three stages:

- Array synchronisation (correct clock drifts)
- Speech enhancement (beamforming)
- ASR system
  - several AM retraining stages
  - data, feature and model transformations

DNN-HMM BL:

$WER = 82.5\%$

WER [%] performance of the dev-set among channels (variance, gain):

- Ref. Kinect channels – U_ref (4): $min = 82.36\%$, $max = 82.72\% \rightarrow$ 0.36%/0.26%
- Beamformed Kinects – U+Bflt (5): $min = 82.61\%$, $max = 85.32\% \rightarrow$ 2.74%/−0.09%
- Kinects channels – U (20): $min = 83.39\%$, $max = 85.68\% \rightarrow$ 2.29%/−0.87%

On utterance-level → Oracle WER [%] results:

Channels	Dev S02	Dev S09	Dev Overall
Baseline: U_ref + Bflt (1)	83.4	81.1	82.5
U_ref (4)	76.1	72.8	74.8
U + Bflt (5)	70.8	68.2	69.3
U (20)	66.3	63.3	65.1
U + Bflt, U (25)	65.5	62.3	64.3
U_ref, U + Bflt, U (29)	64.6	62.2	63.6

Performance gain: 18.9%

20 single ch. (WER/ranks):

WER [%] performance of the dev-set among channels (variance, gain):

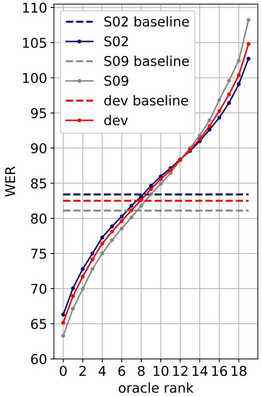
- Ref. Kinect channels - U_ref (4): $min = 82.36\%$, $max = 82.72\% \rightarrow 0.36\%/0.26\%$
- Beamformed Kinects - U+Bflt (5): $min = 82.61\%$, $max = 85.32\% \rightarrow 2.74\%/-0.09\%$
- Kinects channels - U (20): $min = 83.39\%$, $max = 85.68\% \rightarrow 2.29\%/-0.87\%$

On utterance-level → Oracle WER [%] results:

Channels	Dev S02	Dev S09	Dev Overall
Baseline: U_ref + Bflt (1)	83.4	81.1	82.5
U_ref (4)	76.1	72.8	74.8
U + Bflt (5)	70.8	68.2	69.3
U (20)	66.3	63.3	65.1
U + Bflt, U (25)	65.5	62.3	64.3
U_ref, U + Bflt, U (29)	64.6	62.2	63.6

Performance gain: 18.9%

20 single ch. (WER/ranks):



Channel selection:

- **Method:** Deep Neural Network to classify "oracle channels"
- **Labels:** Oracle results $\rightarrow$ multi-label, multi-class problem
- **Features:** Signal-based and/or decoder-based features correlating with oracle results

Signal-based features:

- Signal energy:

$$x_m^u[n] = \frac{1}{N_e - N_s + 1} \sum_{n=N_s}^{N_e} |s_m^u[n]|^2$$

- Peak of GCC-PHAT:

$$\hat{R}_{i,ref}(d) = \mathcal{F}^{-1}\left(\frac{X_i(f)X_{ref}^*(f)}{|X_i(f)X_{ref}^*(f)|}\right)$$

- Envelope variance:

$$C^* = \underset{m}{\operatorname{argmax}} \sum_k w_m[k] \frac{V_m[k]}{\max_m(V_m[k])}$$

- Mel-filterbank

Decoder-based features:

- Average posterior entropy:

$$H_t^m = -\sum_S \boldsymbol{p}_t^m \cdot log_2\left(\boldsymbol{p}_t^m\right)$$

$$H_{avg}^m = \frac{1}{T}\sum_{t=0}^{T} H_t^m$$

- Average posterior moments: mean, variance, skewness, kurtosis

Channel selection:

- **Method:** Deep Neural Network to classify "oracle channels"
- **Labels:** Oracle results → multi-label, multi-class problem
- **Features:** Signal-based and/or decoder-based features correlating with oracle results

Signal-based features:

- Signal energy:

$$x_m^u[n] = \frac{1}{N_e - N_s + 1} \sum_{n=N_s}^{N_e} |s_m^u[n]|^2$$

- Peak of GCC-PHAT:

$$\hat{R}_{i,ref}(d) = \mathcal{F}^{-1}\left(\frac{X_i(f)X_{ref}^*(f)}{|X_i(f)X_{ref}^*(f)|}\right)$$

- Envelope variance:

$$C^* = \operatorname*{argmax}_m \sum_k w_m[k] \frac{V_m[k]}{\max_m(V_m[k])}$$

- Mel-filterbank

Decoder-based features:

- Average posterior entropy:

$$H_t^m = -\sum_S \boldsymbol{p}_t^m \cdot log_2\left(\boldsymbol{p}_t^m\right)$$

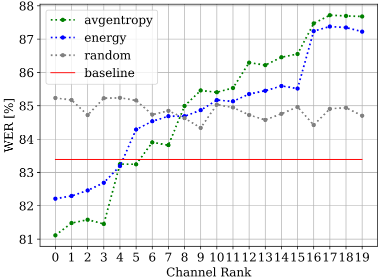
$$H_{avg}^m = \frac{1}{T}\sum_{t=0}^{T} H_t^m$$

- Average posterior moments: mean, variance, skewness, kurtosis

Channel Selection – Results

Feature direct classification:

Channels	Feature	Dev S02	Dev S09	Dev Overall
U+Bflt (5)	Energy	81.2	81.6	81.3
	GCC-PHAT	81.1	81.7	81.4
U (20)	Energy	82.2	82.0	82.1
	Avg. Entropy	81.1	81.8	81.4



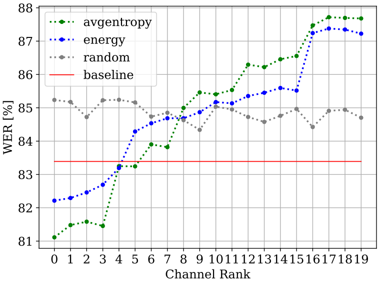
DNN classification:

Channels	Feature	Dev S02	Dev S09	Dev Overall
U (20)	Energy	82.2	82.7	82.8
	EV	83.7	82.6	82.7
	Fbank	83.8	83.5	83.7
	Avg. Entropy	81.7	82.8	82.1
	Avg. Moments	82.8	81.3	82.3
	Stacked	82.3	82.3	82.3
U+Bflt (5)	Avg. Entropy	80.8	80.1	80.5
	Avg. Moments	81.1	80.7	81.0

Channel Selection – Results

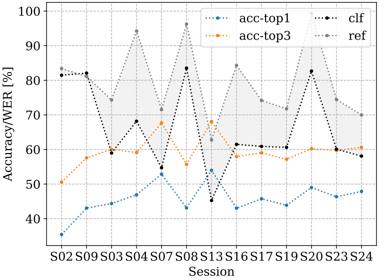
Feature direct classification:

Channels	Feature	Dev S02	Dev S09	Dev Overall
U+Bflt (5)	Energy	81.2	81.6	81.3
	GCC-PHAT	81.1	81.7	81.4
U (20)	Energy	82.2	82.0	82.1
	Avg. Entropy	81.1	81.8	81.4



DNN classification:

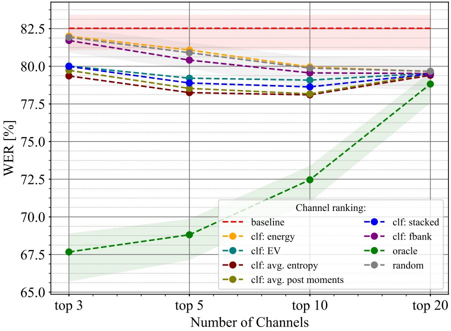
Channels	Feature	Dev S02	Dev S09	Dev Overall
U (20)	Energy	82.2	82.7	82.8
	EV	83.7	82.6	82.7
	Fbank	83.8	83.5	83.7
	Avg. Entropy	81.7	82.8	82.1
	Avg. Moments	82.8	81.3	82.3
	Stacked	82.3	82.3	82.3
U+Bflt (5)	Avg. Entropy	80.8	80.1	80.5
	Avg. Moments	81.1	80.7	81.0



Hypothesis fusion:

- ROVER combination of the {3, 5, 10, 20}-best hypothesis as determined from the DNN-classifier
- Combination for all features
- **Upper baseline:** combine hypothesis from oracle ranking
- **Lower baseline:** random combination of N hypothesis

# Channels	3	5	10	20
Energy	82.00	81.08	79.96	79.65
EV	80.02	79.21	79.08	79.54
Avg. Entropy	79.36	78.25	**78.10**	79.40
Avg. Moments	79.73	78.53	78.17	79.51
Stacked	79.99	78.89	78.63	79.49
Fbank	81.71	80.41	79.56	79.52
Oracle	67.67	68.81	72.46	78.82
Random	81.92	80.90	79.88	79.67



Conclusion and Future Work

Summary:

- The oracle results show a high possible theoretical performance gain from a on utterance-level based channel selection.
- Channel selection does not deliver notable improvements in WER → Informative value of the extracted features, difficulty of the dataset, bad network generalisation.

Ideas:

- Investigation on a curated dataset to trace back the problem to the channel selection stage rather conflicting with a difficult dataset.
- Application of other/more informative features, having a stronger correlation with the oracle labels.

Thank you!