Research Internship Report

Daniel Gerber

supervised by Christian Hümmer

Winter Term 2015

Experiments

This research internship focuses on the investigation of the coherence-based postfilter in an automatic speech recognition (ASR) system. The postfilter is a Wiener Filter, which performs spectral substraction in the short-time Fourier transform (STFT) domain. It's weights are computed based on the estimation of the coherence between two microphones. We evaluated the recognition accuracy of a deep neural network (DNN)-based ASR system using the Kaldi Toolkit and the REVERB Challenge dataset. The dataset consists of real (RealData) and synthesized (SimData) data. Only the latter ones are touched by the experiments and the others are used for validation.

Exp.1 - Influence of different RIRs and enhancement by postfilter

The first experiment deals with different room impulse responses (RIRs). Three cases are distinguished, namely full, early and late. The full case describes the full RIR, whereas the early case has zero-valued entries in the area of the late reverberation, which begins 50 ms after the direct path. The late case is constructed accordingly vice versa. Also two cases are distinguished with enhancement through postfilter (Enh) and without (NoEnh). This gives 6 configurations, for which the word error rate (WER) scores are shown in Tab.1.

Exp.2 - Missmatch between training and classification regarding postfilter

The second experiment introduces a mismatch between training and classification stage. Meaning the training of the DNN is performed without the front-end enhancement at all. Later the classification stage is evaluated with the appliance of the postfilter. Again this is done in a full, early and late RIR manner. Tab.2 shows the corresponding results.

Exp.3 - Incorporate variance in feature extraction

The third experiment also used the missmatch setup and the variance information given in the estimation of the postfilter coefficients is omitted, whereas in the above cases it is also passed through the nonlinear function of the log-mel filter bank and incorporated in the resulting features. The results are given in Tab.3 and covering again full, early and late case.

Conclusion

The foregoing experiments showed that the late reverberation has a greater influence on the recognition rates than the early reflections. Also the postfilter revealed to be quite beneficial in the context of ASR.

Results

Data	SimData							RealData		
Distance	Far			Near			**Avg**	Far	Near	**Avg**
Room	1	2	3	1	2	3		1	1	
Full-NoEnh	6.23	10.75	12.28	5.93	6.74	7.28	8.20	21.27	21.02	21.15
Full-Enh	6.96	10.01	12.30	6.10	6.90	7.16	8.24	21.94	22.96	22.45
Early-NoEnh	5.59	6.99	6.79	5.08	5.32	5.82	5.93	51.18	50.43	50.81
Early-Enh	6.54	7.70	7.95	5.76	5.84	6.43	6.70	31.87	32.67	32.27
Late-NoEnh	5.32	24.27	11.65	5.27	8.92	6.50	10.32	29.44	28.62	29.03
Late-Enh	5.62	27.26	11.38	5.45	8.91	6.89	10.92	24.92	26.51	25.72

Table 1: ASR word error rates in percentage of REVERB challenge evaluation set with different room impulse responses (Full, Early, Late) and enhancement or no enhancement via postfilter (Enh, NoEnh)

Data	SimData							RealData		
Distance	Far			Near			**Avg**	Far	Near	**Avg**
Room	1	2	3	1	2	3		1	1	
Full	7.50	10.64	12.79	6.15	6.82	7.40	8.55	21.20	21.72	21.46
Early	11.06	13.57	15.39	6.74	8.47	10.01	10.87	50.54	51.68	51.11
Late	5.64	28.53	12.50	5.35	10.53	7.34	11.65	28.76	31.04	29.90

Table 2: ASR word error rates in percentage of REVERB challenge evaluation set with different room impulse responses (Full, Early, Late) and missmatch of training (without postfilter) and classification (with postfilter)

Data	SimData							RealData		
Distance	Far			Near			**Avg**	Far	Near	**Avg**
Room	1	2	3	1	2	3		1	1	
Full	7.49	10.11	12.38	6.57	6.93	7.51	8.50	21.71	22.61	22.16
Early	6.74	7.39	8.41	6.0	6.25	6.52	6.89	37.58	38.23	37.91
Late	6.18	30.17	12.83	5.88	10.5	8.03	12.27	26.74	27.72	27.23

Table 3: ASR word error rates in percentage of REVERB challenge evaluation set with different room impulse responses (Full, Early, Late) and incorporated variance information during feature extraction