Improved Cepstra Minimum-Mean-Square-Error Noise Reduction Algorithm for Robust Speech Recognition

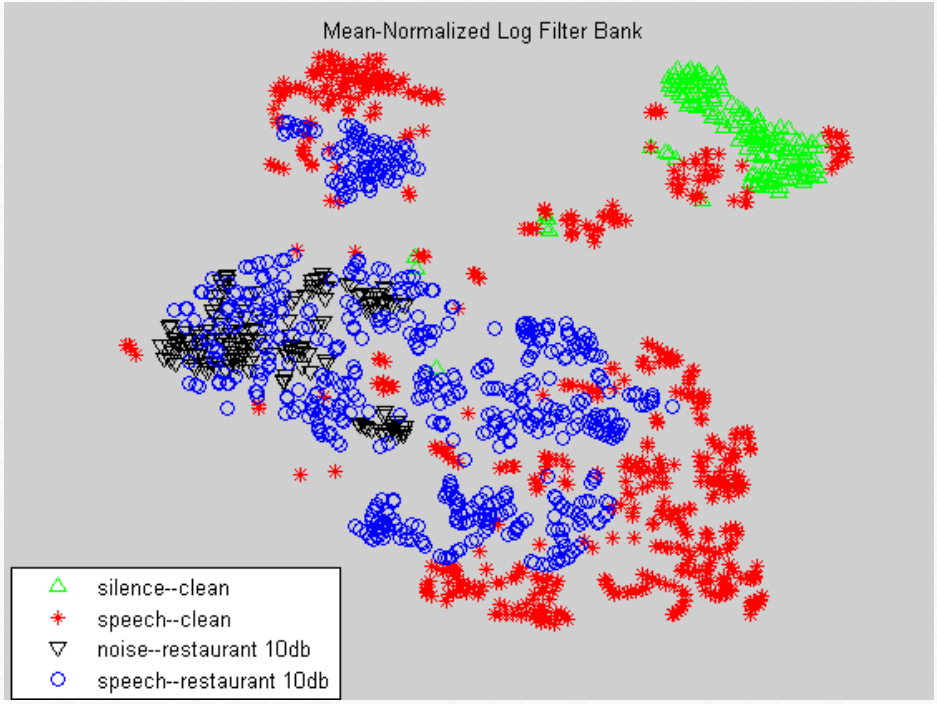
Jinyu Li, Yan Huang, Yifan Gong

Microsoft

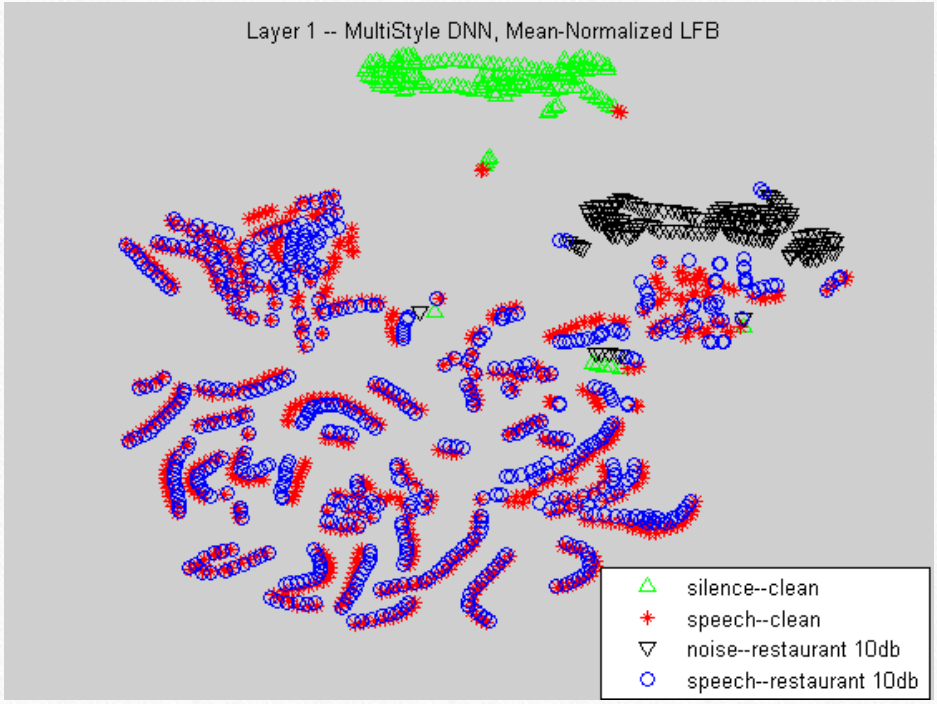
Robust Front-End

- Conventional noise-robust front-ends work very well with the Gaussian mixture models (GMMs).
- Single-channel robust front-ends were reported not helpful to multi-style deep neural network (DNN) training.
  - DNN's layer-by-layer structure provides a feature extraction strategy that automatically derives powerful noise-resistant features

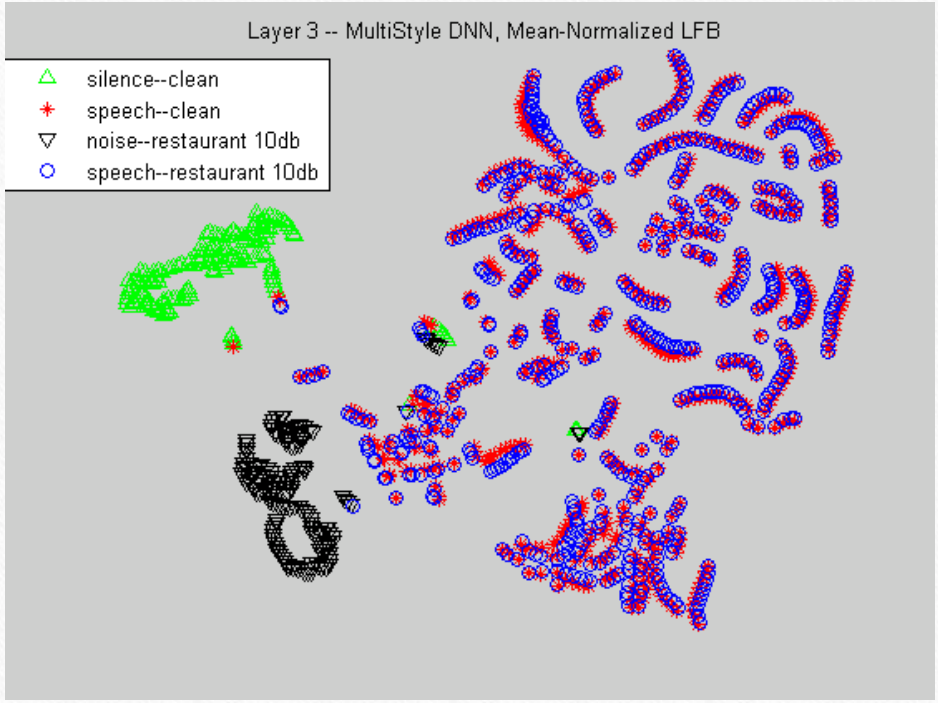
t-SNE Plot for Paired Clean and Noisy Utterances in Training Set (LFB)



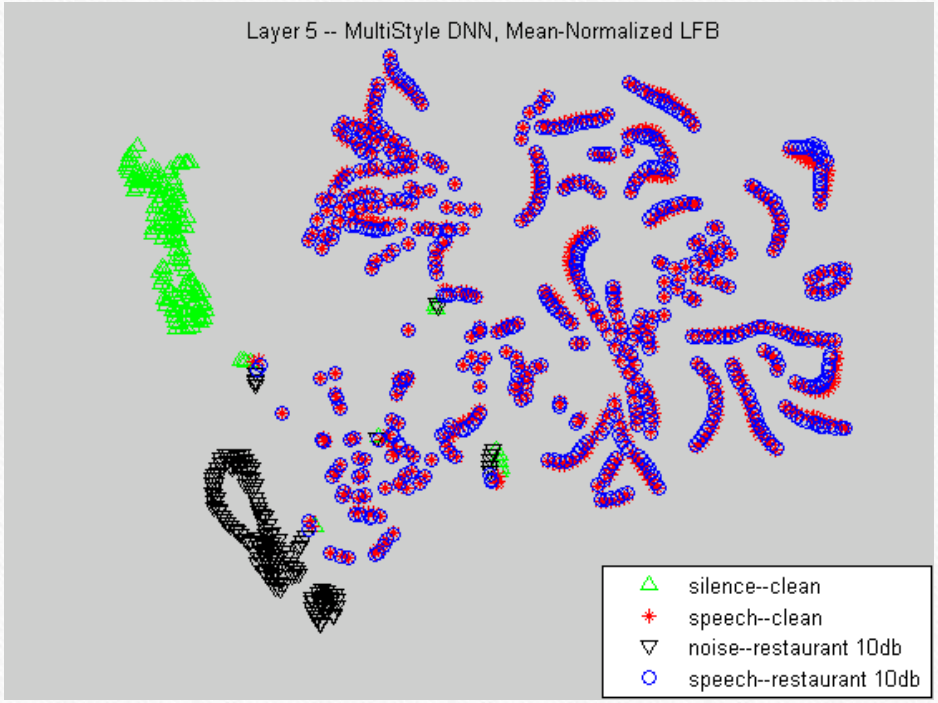
t-SNE Plot for Paired Clean and Noisy Utterances in Training Set (Layer 1)



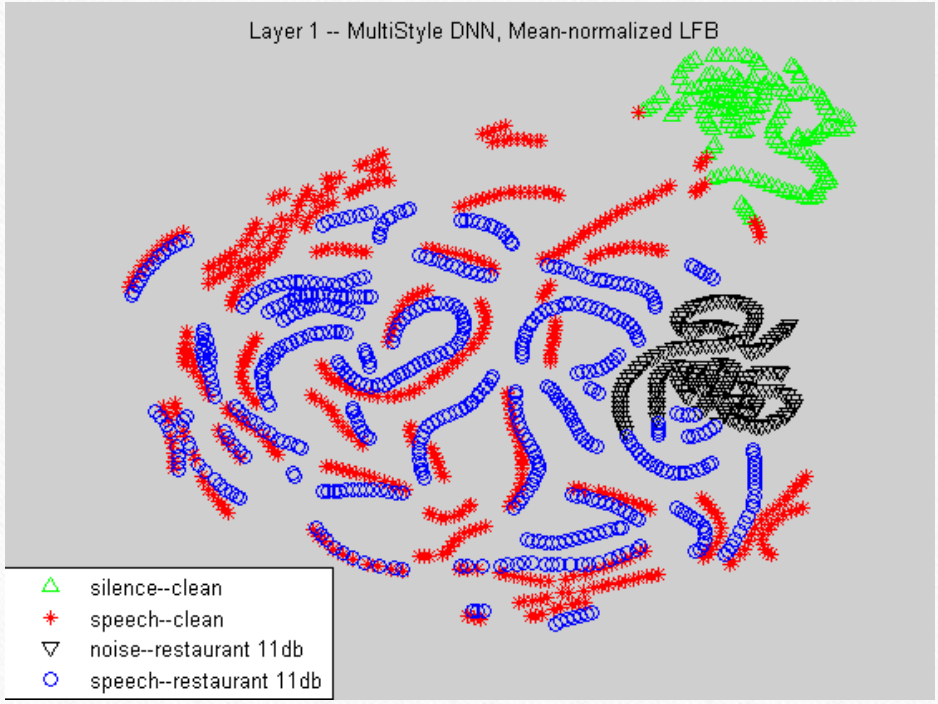
t-SNE Plot for Paired Clean and Noisy Utterances in Training Set (Layer 3)



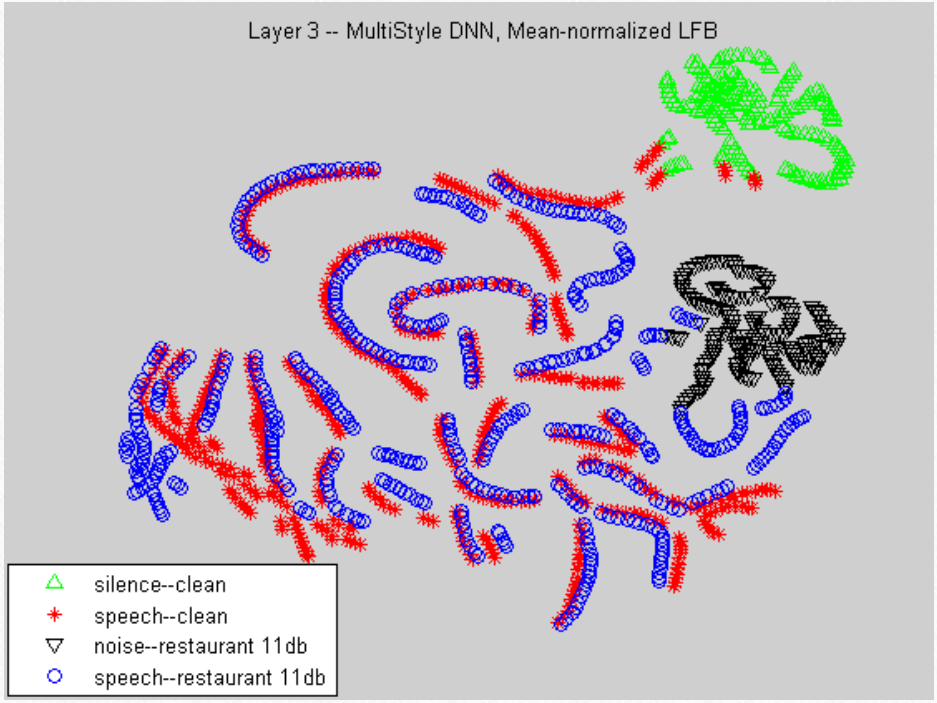
t-SNE Plot for Paired Clean and Noisy Utterances in Training Set (Layer 5)



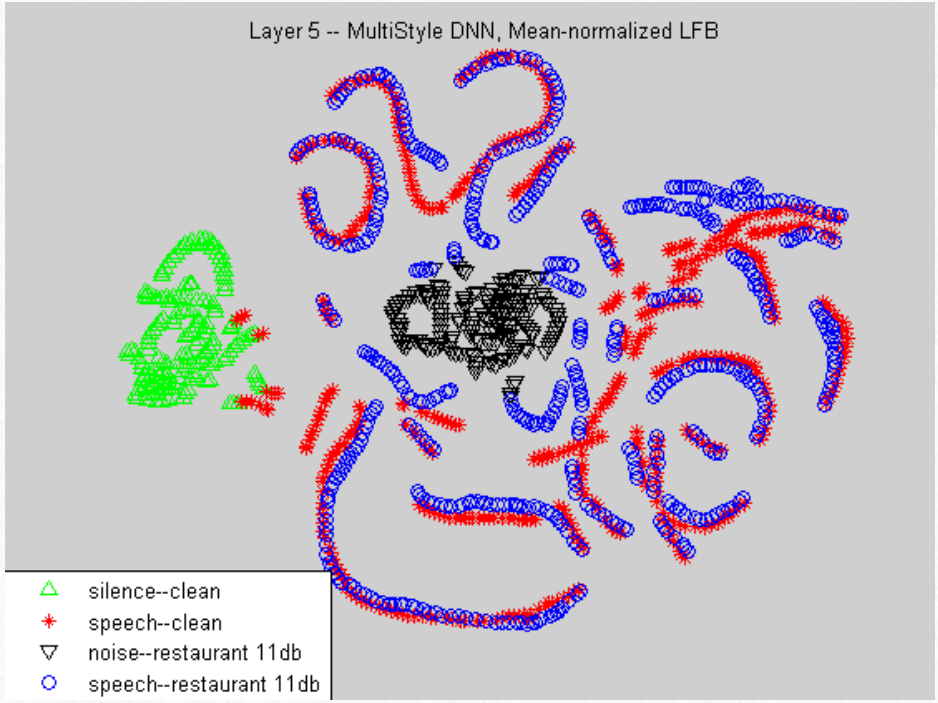
t-SNE Plot for Paired Clean and Noisy Utterances in Testing Set (Layer 1)



t-SNE Plot for Paired Clean and Noisy Utterances in Testing Set (Layer 3)



t-SNE Plot for Paired Clean and Noisy Utterances in Testing Set (Layer 5)



Robust Front-End

- Conventional noise-robust front-ends work very well with the Gaussian mixture models (GMMs).
- Single-channel robust front-ends were reported not helpful to multi-style deep neural network (DNN) training although multi-channel signal processing still helps.
- In this study, we show that the single-channel robust front-end is still beneficial to deep learning models as long as it is well designed.

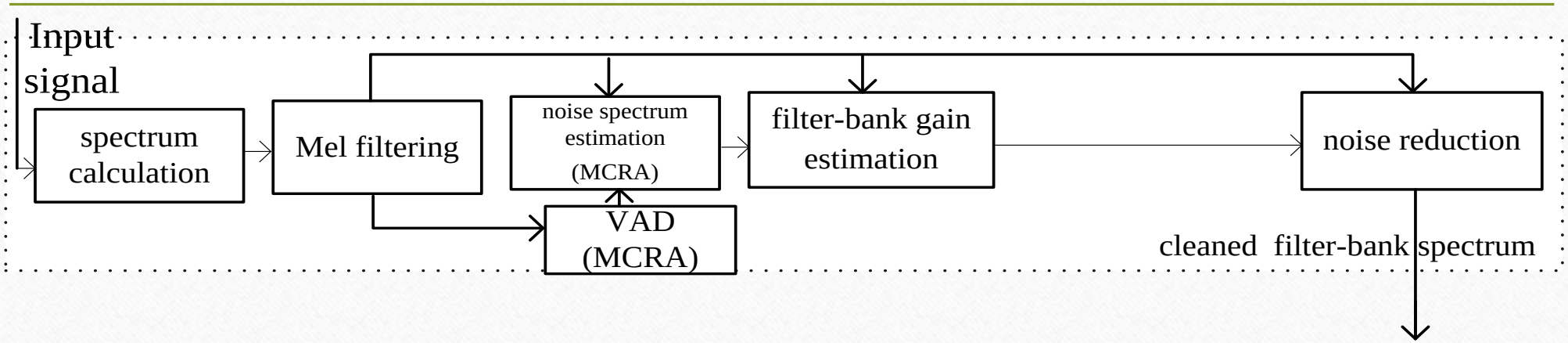
Cepstra Minimum Mean Square Error (CMMSE)

- Reported very effective in dealing with noise when used in the GMM-based acoustic models
- The solution to CMMSE for each element of the dimension-wise MFCC: $\hat{c}_x(t,k) = E\{c_x(t,k)|\boldsymbol{m}_y(t)\} = \sum_b a_{k,b}\, E\{\log m_x(t,b)\,|\boldsymbol{m}_y(t)\}$
- The problem is reduced to finding the log-MMSE estimator of the Mel filter-bank's output: $\hat{m}_x(t,b) \approx \exp\big(E\{\log m_x(t,b)\,|m_y(t,b)\}\big)$ given a weak independent assumption between Mel-filterbanks.

4-Step Processing of CMMSE

- Voice activity detection (VAD): detects the speech probability at every time-filterbank bin;
- Noise spectrum estimation: uses the estimated speech probability to update the estimation of noise spectrum;
- Gain estimation: uses the noisy speech spectrum and the estimated noise spectrum to calculate the gain of every time-filterbank bin;
- Noise reduction: applies the estimated gain to the noisy speech spectrum to generate the clean spectrum.

CMMSE


Input signal
spectrum calculation
Mel filtering
noise spectrum estimation (MCRA)
filter-bank gain estimation
noise reduction
VAD (MCRA)
cleaned filter-bank spectrum

VAD

- CMMSE uses a minimum controlled recursive moving average (MCRA) noise tracker (Cohen and Berdugo, 2002) to detect the speech probability $p(t, b)$ in each filterbank bin *b* and time *t*.
  - I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, *9*(1), pp.12-15, 2002.

Noise Spectrum Estimation

- The noise power spectrum $m_n(t,b)$ is estimated using MCRA(Cohen and Berdugo, 2002) as

$$m_n(t,b) = \alpha * m_n(t-1,b) + (1.0-\alpha) * m_y(t,b)$$

with

$$\alpha = \alpha_D + (1.0-\alpha_D) * p(t,b)$$

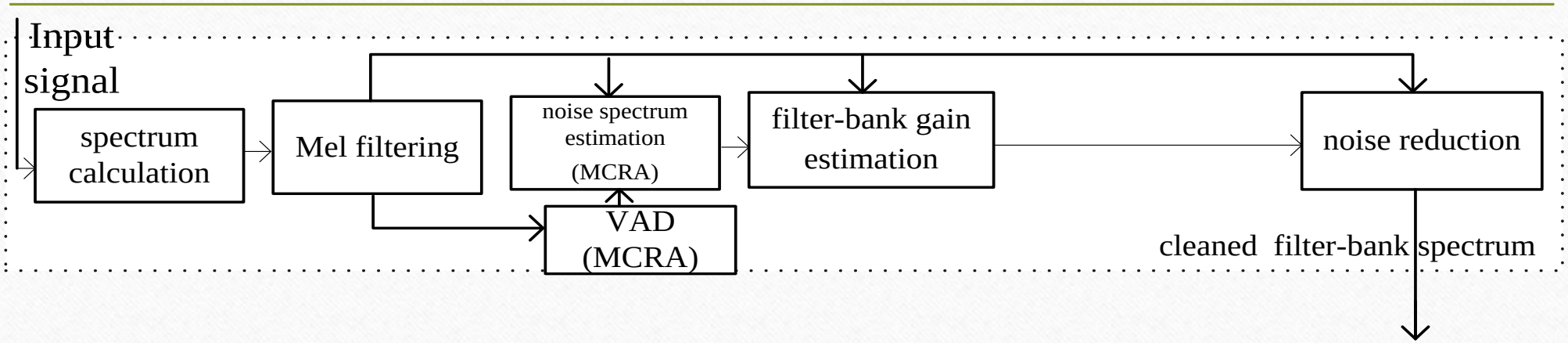
Gain Estimation

- The gain of time-filterbank bin $G(t,b) = \frac{\hat{\xi}(t,b)}{1+\hat{\xi}(t,b)} \exp\left\{\frac{1}{2}\int_{v(t,b)}^{\infty} \frac{e^{-\tau}}{\tau} d\tau\right\}$.
  - posterior SNR : $\gamma(t,b) = \frac{m_y(t,b)}{m_n(t,b)}$
  - prior SNR is calculated using a decision-directed approach (DDA) $\hat{\xi}(t,b) = \beta * G(t-1,b) * \gamma(t-1,b) + (1.0-\beta) * \xi(t,b)$
  - $\xi(t,b) = \max(\gamma(t,b) - 1, 0.0)$
  - $v(t,b) = \frac{\hat{\xi}(t,b)}{1+\hat{\xi}(t,b)}\ \gamma(t,b)$

Noise Reduction

- $\widehat{m}_x(t,b) \approx G(t,b)m_y(t,b)$

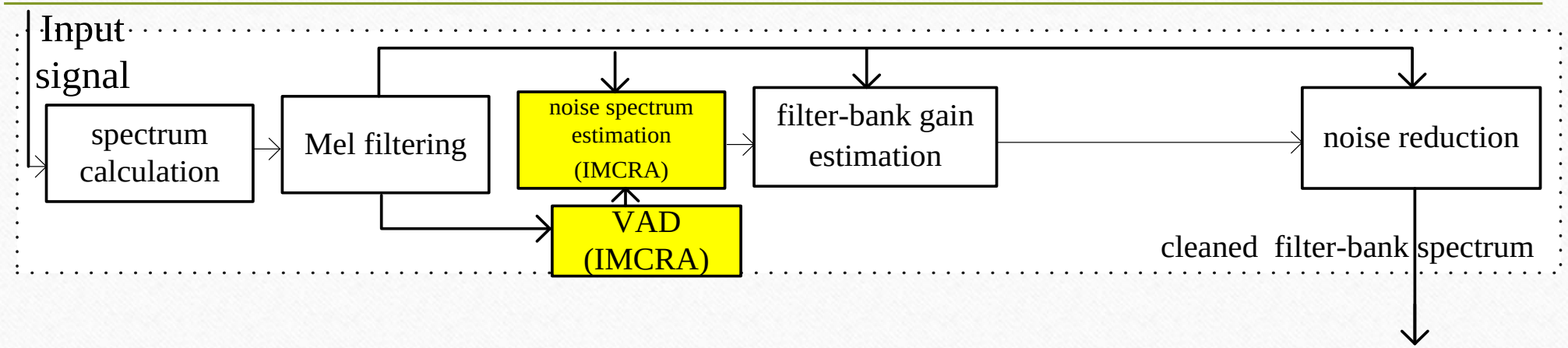
CMMSE


Input signal
spectrum calculation
Mel filtering
noise spectrum estimation (MCRA)
filter-bank gain estimation
noise reduction
VAD (MCRA)
cleaned filter-bank spectrum

Speech Probability Estimation

- Reliable speech probability estimation is critical to the estimation of noise spectrum: $m_n(t,b) = \alpha * m_n(t-1,b) + (1.0-\alpha) * m_x(t,b)$.
- We use IMCRA (Cohen 2003) to estimate the speech probability $p(t,b)$ in each time-filterbank bin.
  - I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, Vol. 11, No. 5, pp. 466-475, 2003.

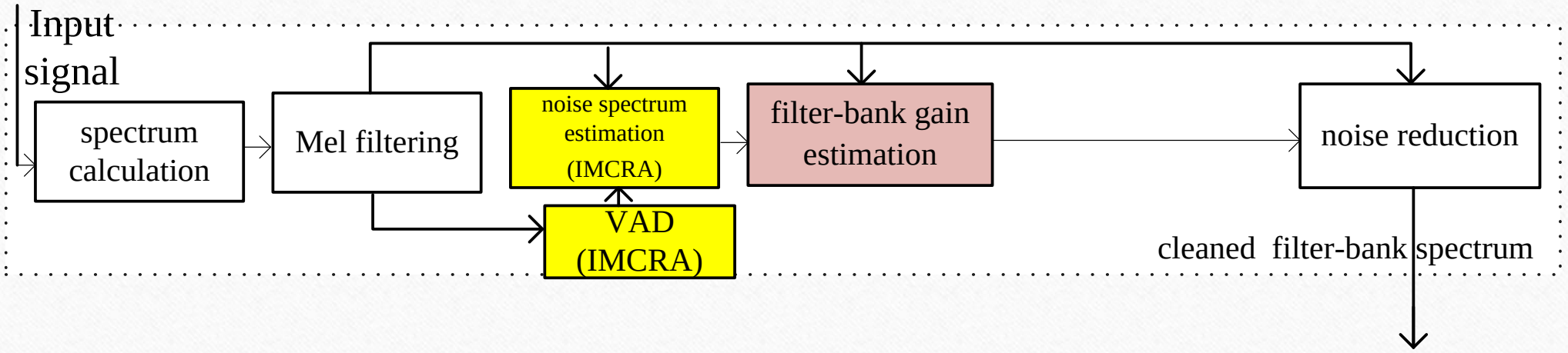
Improving CMMSE



Refined Prior SNR Estimation

- We do further gain estimation to get a converged gain
  - $\dot{\xi}(t,b) = G(t,b) * \gamma(t,b)$
  - $\dot{v}(t,b) = \frac{\dot{\xi}(t,b)}{1+\dot{\xi}(t,b)}\,\gamma(t,b)$
  - $\dot{G}(t,b) = \frac{\dot{\xi}(t,b)}{1+\dot{\xi}(t,b)} \exp\left\{\frac{1}{2}\int_{\dot{v}(t,b)}^{\infty} \frac{e^{-\tau}}{\tau} d\tau\right\}$

Improving CMMSE



OMLSA

- Optimally-modified log-spectral amplitude (OMLSA) speech estimator (Cohen and Berdugo, 2001) is used to modify time-filterbank gain:

$$\hat{G}(t,b) = \dot{G}(t,b)^{p(t,b)} G_0^{1-p(t,b)}$$

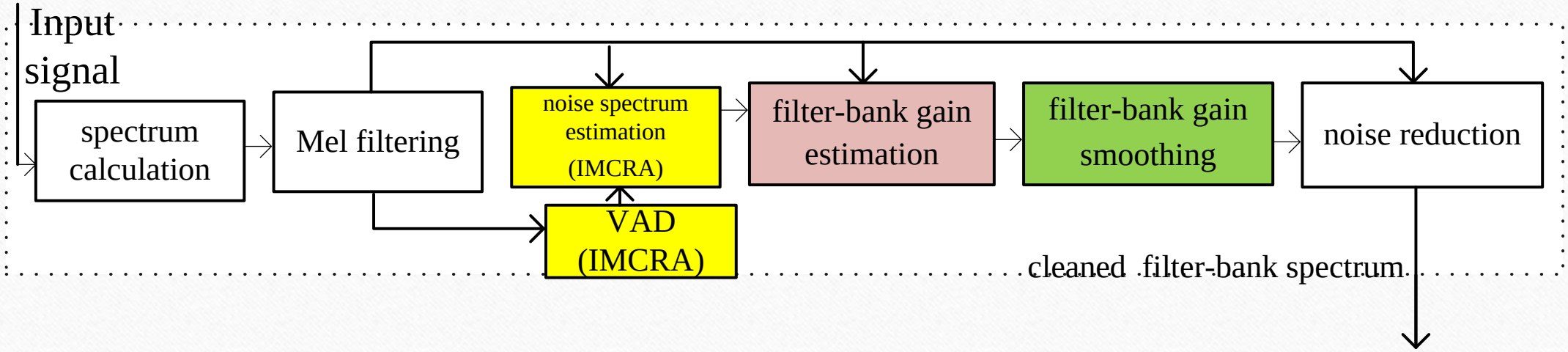
  - I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, 2001.

Gain Smoothing

- It is better to use OMLSA when the speech probability estimation is reliable.
- We do cross filterbank gain smoothing in this stage to partially address the weak independent assumption of Mel-filterbanks.

$$\hat{G}(t,b) = (\dot{G}(t,b-1) + \dot{G}(t,b) + \dot{G}(t,b+1))/3$$

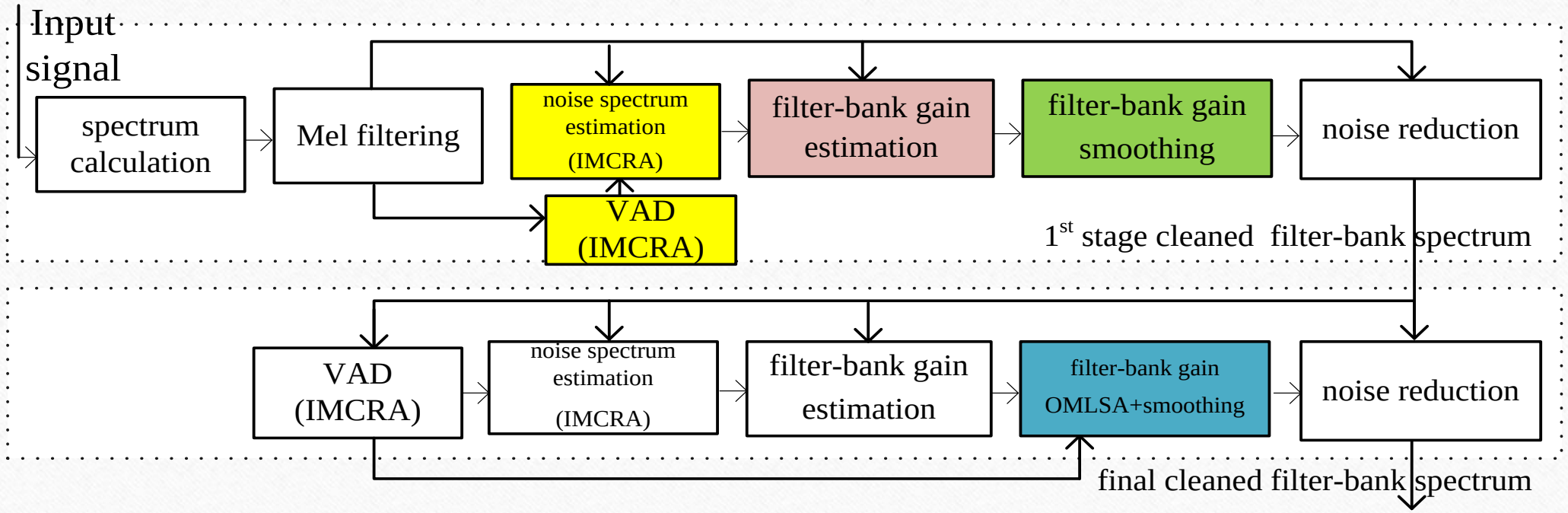
Improving CMMSE



Two-stage Processing

- The noise reduction process is not perfect due to the factors such as imperfect noise estimation.
- A second stage noise reduction can be used to further reduce the noise.
- We use OMLSA together with gain smoothing for the gain modification in the second stage because the residual noise has less impact to speech probability estimation after the first stage noise reduction

Improved Cepstra Minimum Mean Square Error (ICMMSE)



Experiments

Task	Training Data	Acoustic Model
Aurora 2	8440 utterances	GMM

Experiments

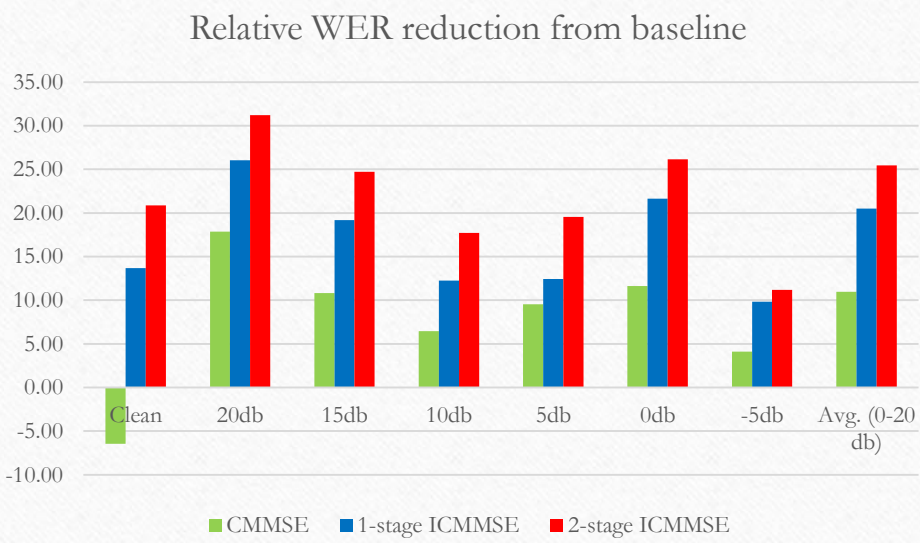
Task	Training Data	Acoustic Model
Aurora 2	8440 utterances	GMM
Chime 3	1600 real utterances + 7138 simulated utterances	feed-forward DNN

Experiments

Task	Training Data	Acoustic Model
Aurora 2	8440 utterances	GMM
Chime 3	1600 real utterances + 7138 simulated utterances	feed-forward DNN
Cortana	3400hr Live data	LSTM-RNN

Aurora 2

	Baseline	CMMSE	1-stage ICMMSE	2-stage ICMMSE
Clean	1.39	1.48	1.20	1.10
20db	2.69	2.21	1.99	1.85
15db	3.6	3.21	2.91	2.71
10db	6.04	5.65	5.30	4.97
5db	14.38	13.01	12.59	11.57
0db	43.41	38.36	34.02	32.06
-5db	75.93	72.8	68.47	67.45
Avg. (0-20 db)	13.67	12.17	10.87	10.19



Improvement Breakdown

Method	Avg. WER
Baseline	13.67
CMMSE	12.17

Improvement Breakdown

Method	Avg. WER
Baseline	13.67
CMMSE	12.17
+ IMCRA	11.66

Improvement Breakdown

Method	Avg. WER
Baseline	13.67
CMMSE	12.17
+ IMCRA	11.66
+OMLSA	10.99

Improvement Breakdown

Method	Avg. WER
Baseline	13.67
CMMSE	12.17
+ IMCRA	11.66
+OMLSA	10.99
+refined prior SNR	10.87

Improvement Breakdown

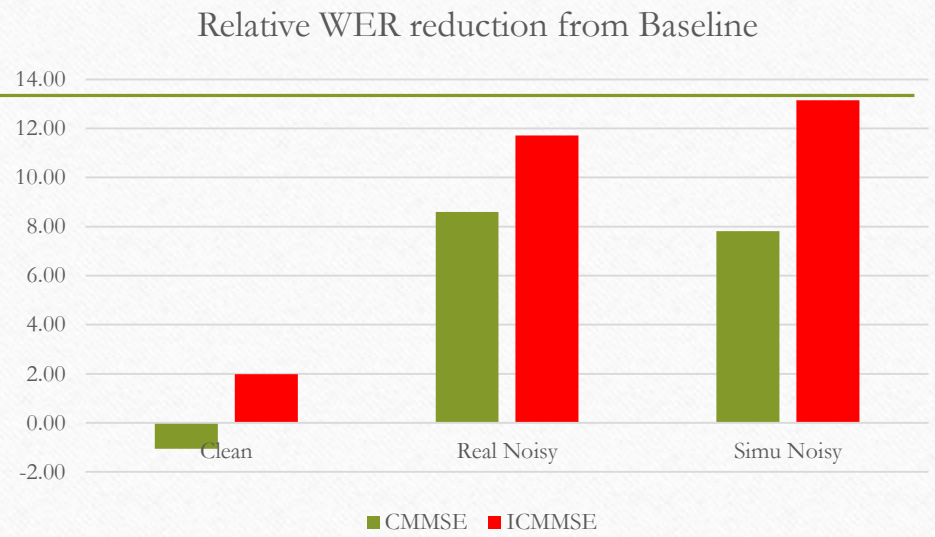
Method	Avg. WER
Baseline	13.67
CMMSE	12.17
+ IMCRA	11.66
+OMLSA	10.99
+refined prior SNR	10.87
-OMLSA +gain smoothing (1-stage ICMMSE)	10.74

Improvement Breakdown

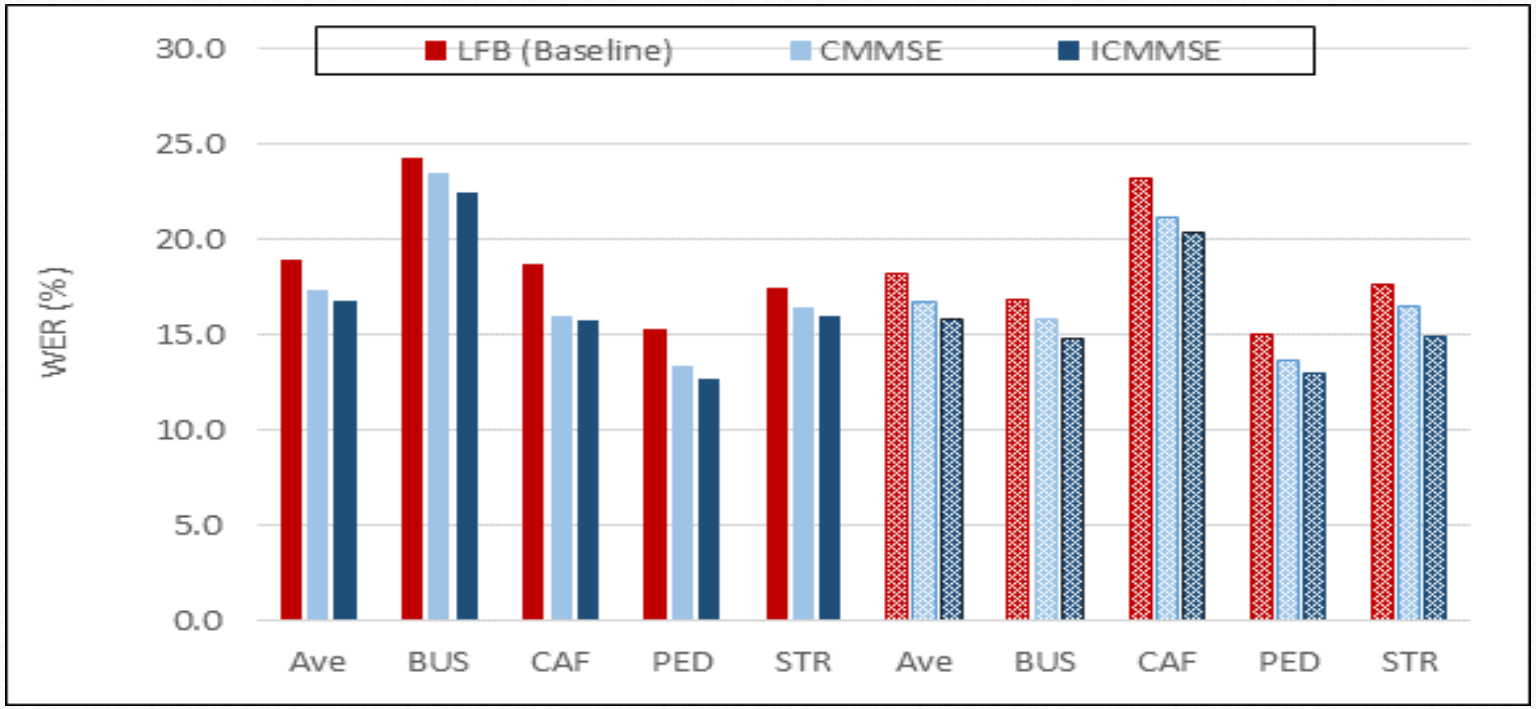
Method	Avg. WER
Baseline	13.67
CMMSE	12.17
+ IMCRA	11.66
+OMLSA	10.99
+refined prior SNR	10.87
-OMLSA +gain smoothing (1-stage ICMMSE)	10.74
+2nd stage processing (2-stage ICMMSE)	10.19

Chime 3

Model FE	Test	Real	Simulate
Baseline	Clean	7.56	N/A
CMMSE	Clean	7.64	N/A
ICMMSE	Clean	7.41	N/A
Baseline	Noisy	18.95	18.18
CMMSE	Noisy	17.32	16.76
ICMMSE	Noisy	16.73	15.79

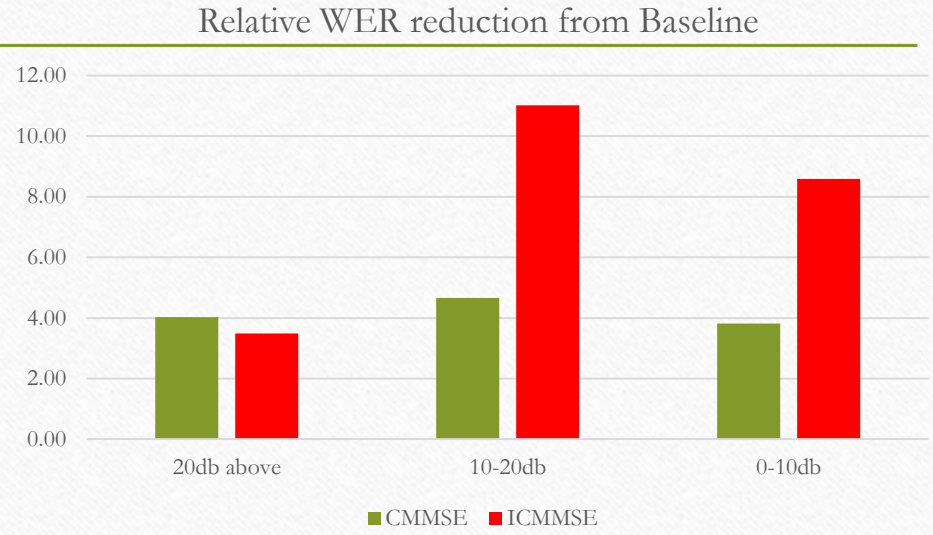


Chime 3 WER Breakdown with Real and Simu. Noisy Test Set



Cortana

WER	Baseline	CMMSE	ICMMSE
20db above	13.17	12.64	12.71
10-20db	20.8	19.83	18.51
0-10db	27.03	26	24.71



Conclusion

- A new robust front-end called ICMMSE is proposed to improve the previous CMMSE front-end with several advanced components
  - The IMCRA algorithm helps to generate more accurate speech probability.
  - The refined prior SNR estimation helps to get a converged gain.
  - Either cross filterbank gain smoothing or OMLSA is helpful to further modify the gain function.
  - The two-stage processing helps to reduce the residual noise after the first-stage processing.
- ICMMSE is superior regardless of the underlying models and evaluation tasks

Result Summary

Task	Training Data	Acoustic Model	Relative WER reduction
Aurora 2	8440 utterances	GMM	25.46%
Chime 3	1600 real utterances + 7138 simulated utterances	feed-forward DNN	11.98%
Cortana	3400hr Live data	LSTM-RNN	11.01%

Thank You