香港城市大學
City University of Hong Kong

Analysis-Synthesis Based Speech Enhancement with Improved Spectrum Envelope Estimation by Tracking Speech Dynamics

Ruofei Chen and Cheung-Fat Chan

Department of Electronic Engineering

City University of Hong Kong

ICASSP 2011, Prague, Czech Republic

27 May 2011

Outline

Backgrounds

Model-based Speech Enhancement

Speech Dynamics Tracking

Performance Evaluation

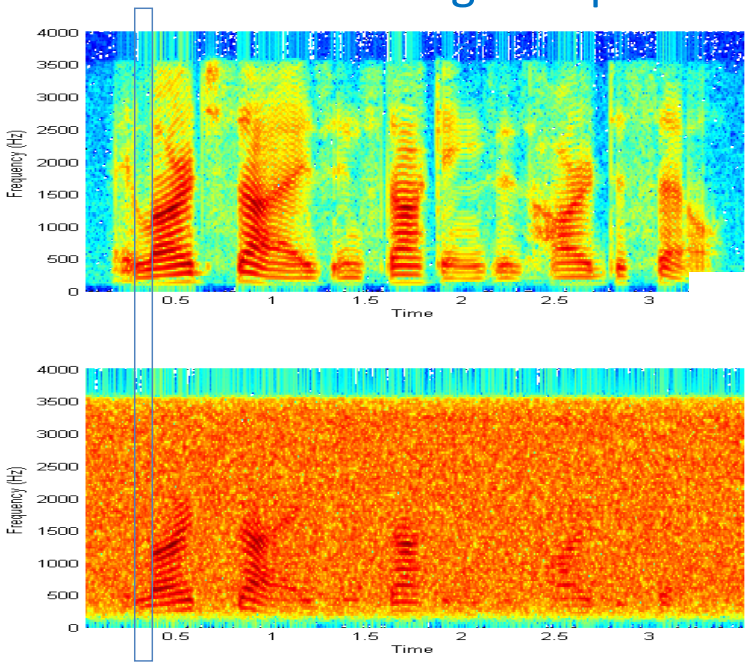
Conclusion

Backgrounds

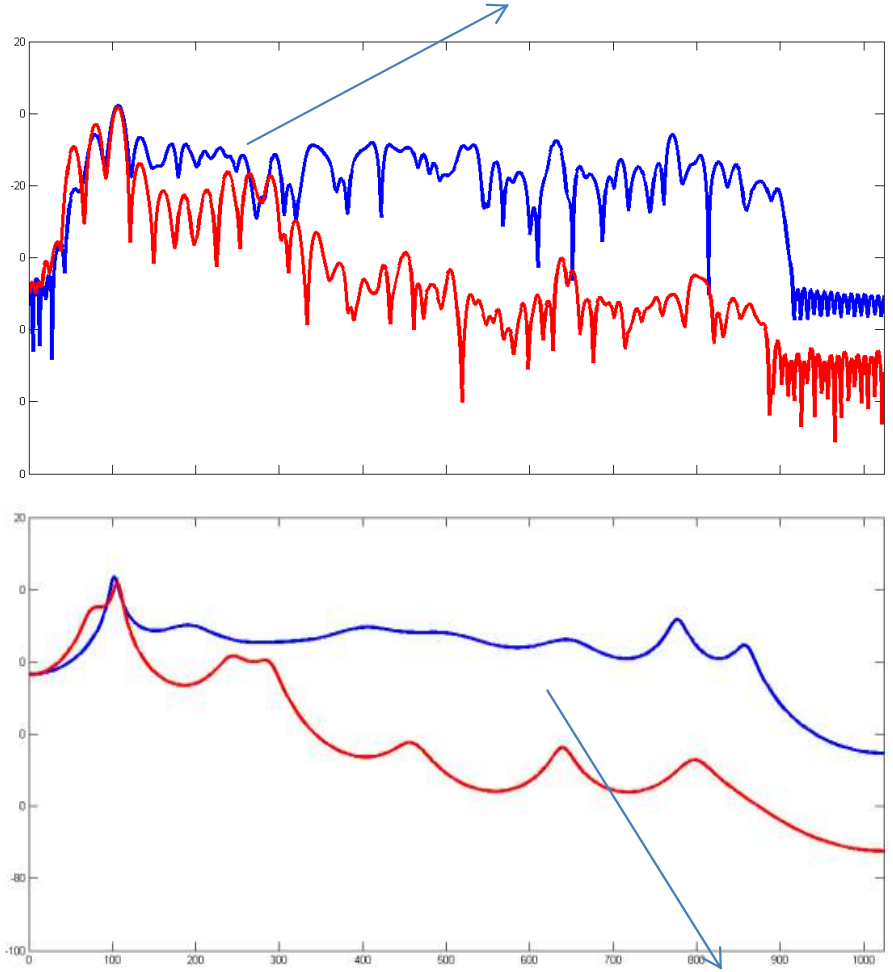
Effect of noise corruption (spectral view)

Damaged harmonic structure

Original Speech



White Noise
SNR=0dB



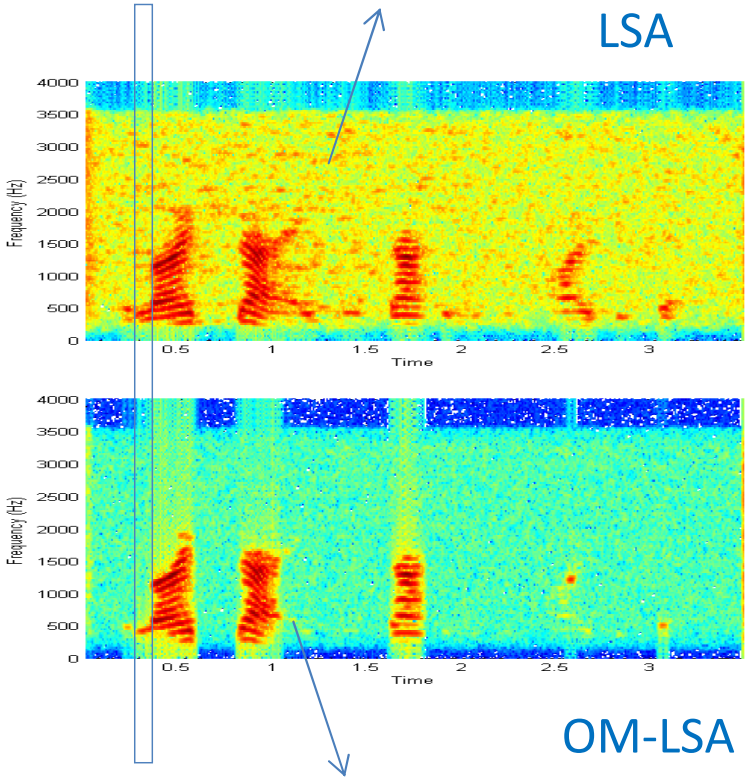
Large spectral distortion
(smeared spectrum envelope)

Backgrounds

Effect of conventional speech enhancement

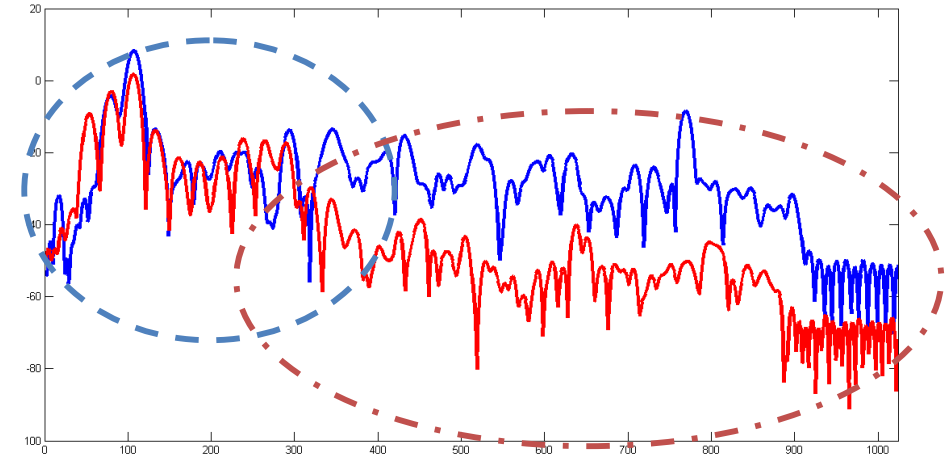
Musical Tones and
Residual Noise

LSA

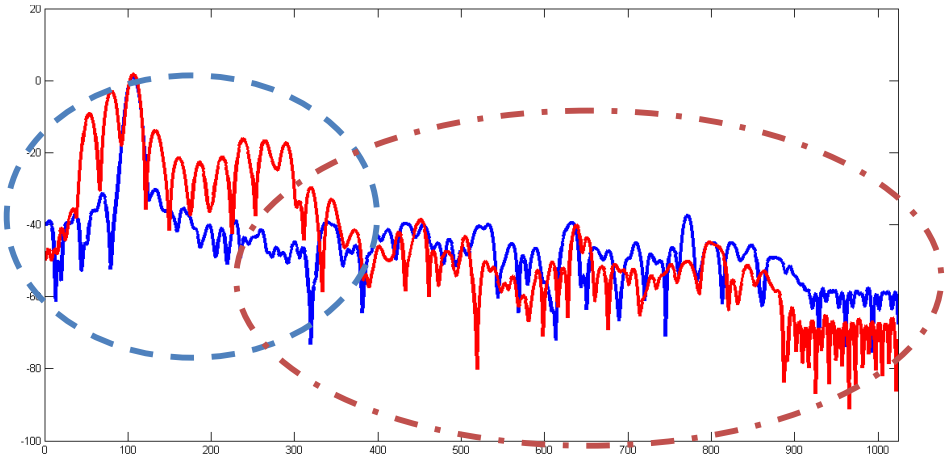


OM-LSA

Less Voice Fidelity



Higher
PESQ score

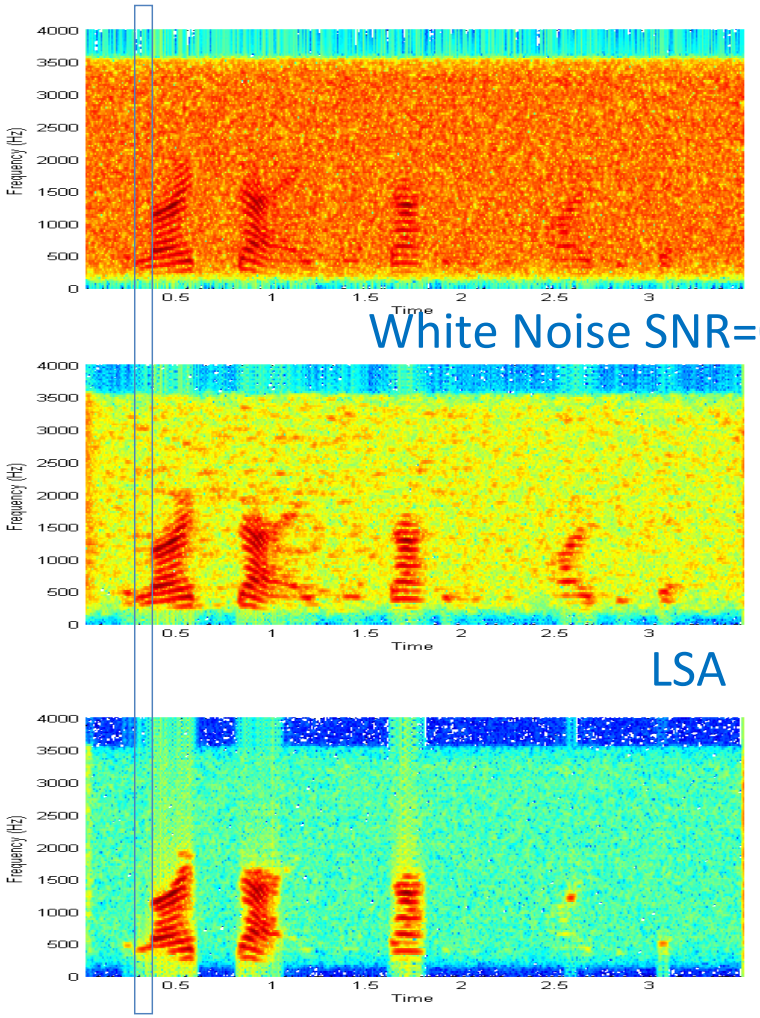


Higher
Segmental SNR

Trade-off : Noise Suppression VS Harmonic Distortion

Backgrounds

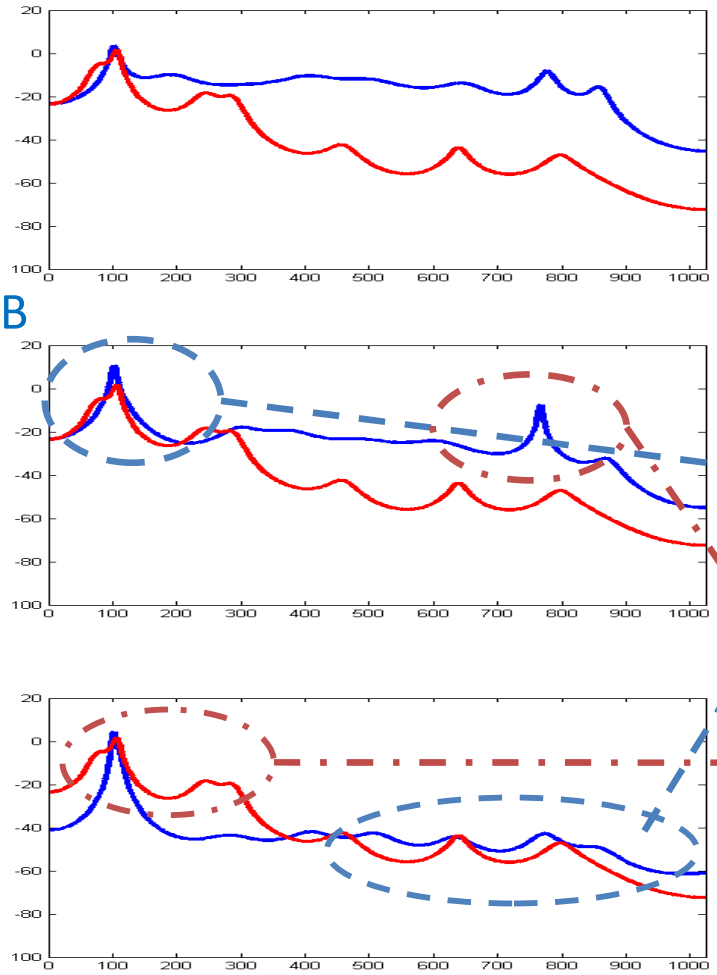
Effect of conventional speech enhancement (cont.)



White Noise SNR=0dB

LSA

OM-LSA



Noise distort the spectrum envelope

Conventional methods:

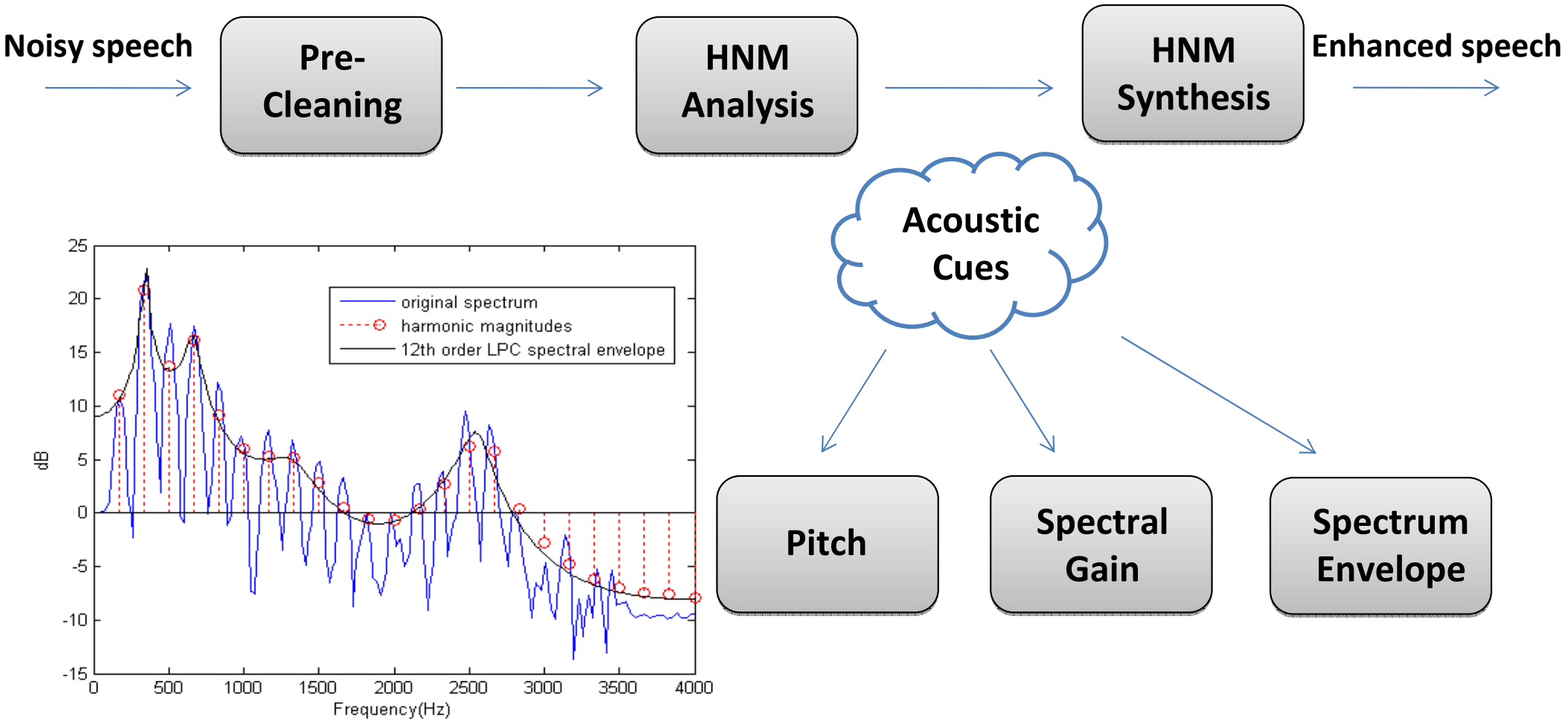
Partially restore the spectrum envelope

Partially further distort the spectrum envelope

Potentially account for common deficiencies: musical tones, low intelligibility

Model-based Speech Enhancement

An analysis-synthesis approach based on harmonic noise model (HNM)



Model-based Speech Enhancement

Why this approach?

- **Noise Suppression**: HNM generates clean harmonics and hence background noise is automatically removed

- **Natural Speech Restoration**: HNM is able to retrieve the damaged harmonic structure (no isolated spectral peaks and hence no “musical tone” problem)

- **Flexible decomposition**: HNM allows independent adjustment of different model parameters (e.g. Modify the spectrum envelope)

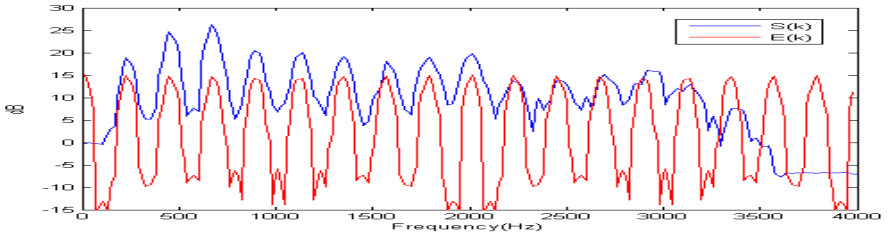
~~**Trade-off : Noise Suppression VS Harmonic Distortion**~~

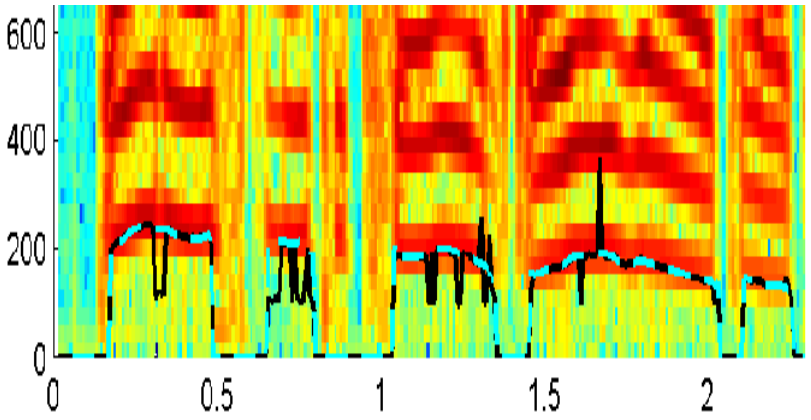
Model-based Speech Enhancement

Model parameter estimation

- **Pitch Estimation**

$$\alpha(\tau) = \frac{\sum_{m=1}^{M(\tau)} \sum_{k=a_m(\tau)}^{b_m(\tau)} [|S(k)| - A_m(\tau)|E(\tau,k)|]^2}{(1-\tau B) \sum_{m=1}^{M(\tau)} \sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)|^2}$$





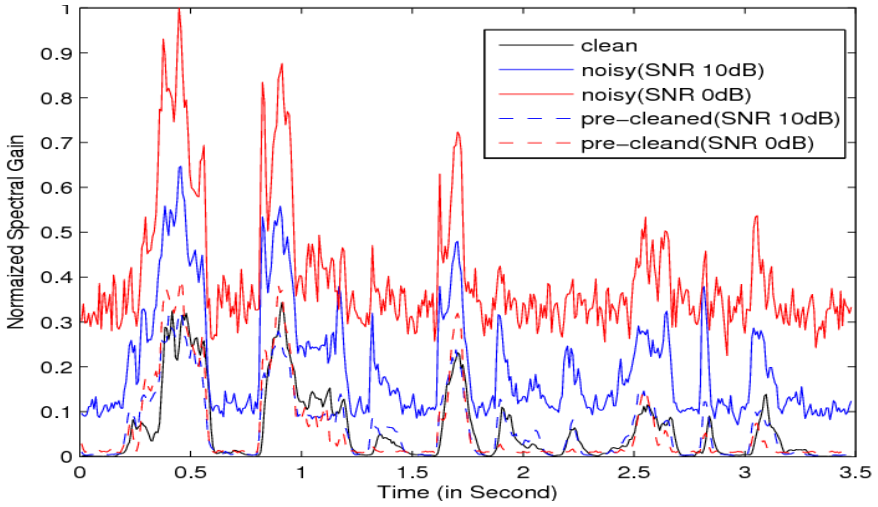
- **Spectral Gain Estimation**

Input magnitude spectrum

Input energy-normalized envelope spectrum

$$\sum_{k=1}^{K} [|S_\ell(k)| - g_\ell |\bar{S}_\ell(k)|]^2$$

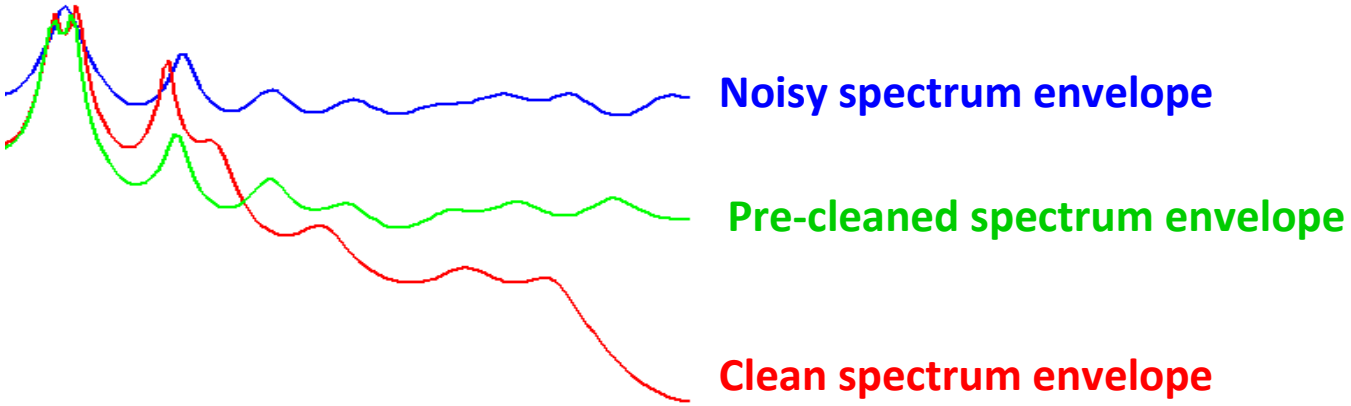
$$g_\ell = \frac{\sum_{k=1}^{K} |S_\ell(k)||\bar{S}_\ell(k)|}{\sum_{k=1}^{K} |\bar{S}_\ell(k)|^2}$$



Model-based Speech Enhancement

Model parameter estimation (cont.)

- **Spectrum Envelope Estimation**

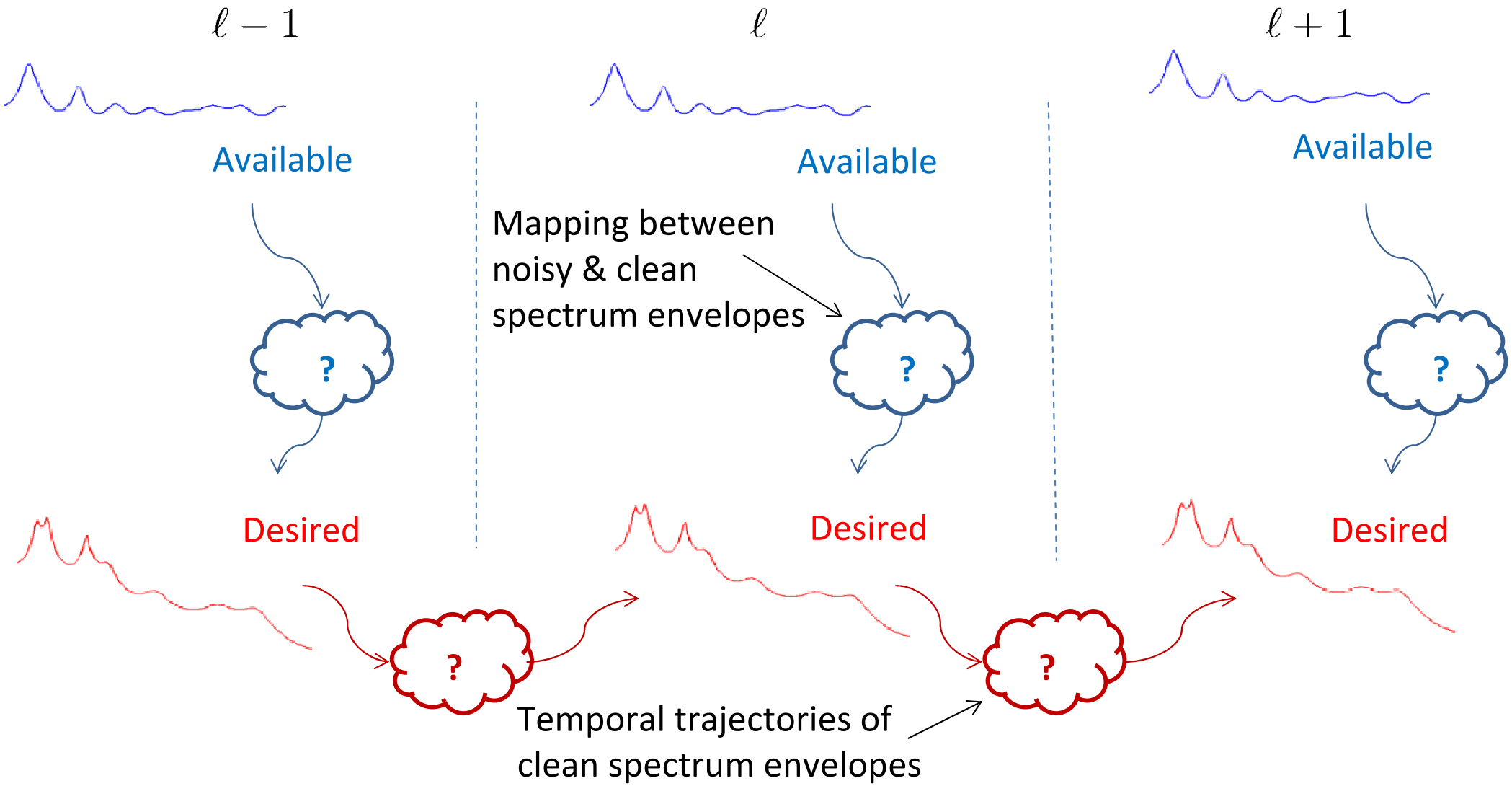


- **Preliminary Results**

SNR = 0dB, White Noise	PESQ (1 ~ 4.5)
Without enhancement	1.55
Best conventional approach (LSA_SPU)	1.87
Analysis-synthesis approach (pre-cleaned envelope)	2.05
Analysis-synthesis approach (clean envelope)	3.17

Speech Dynamics Tracking

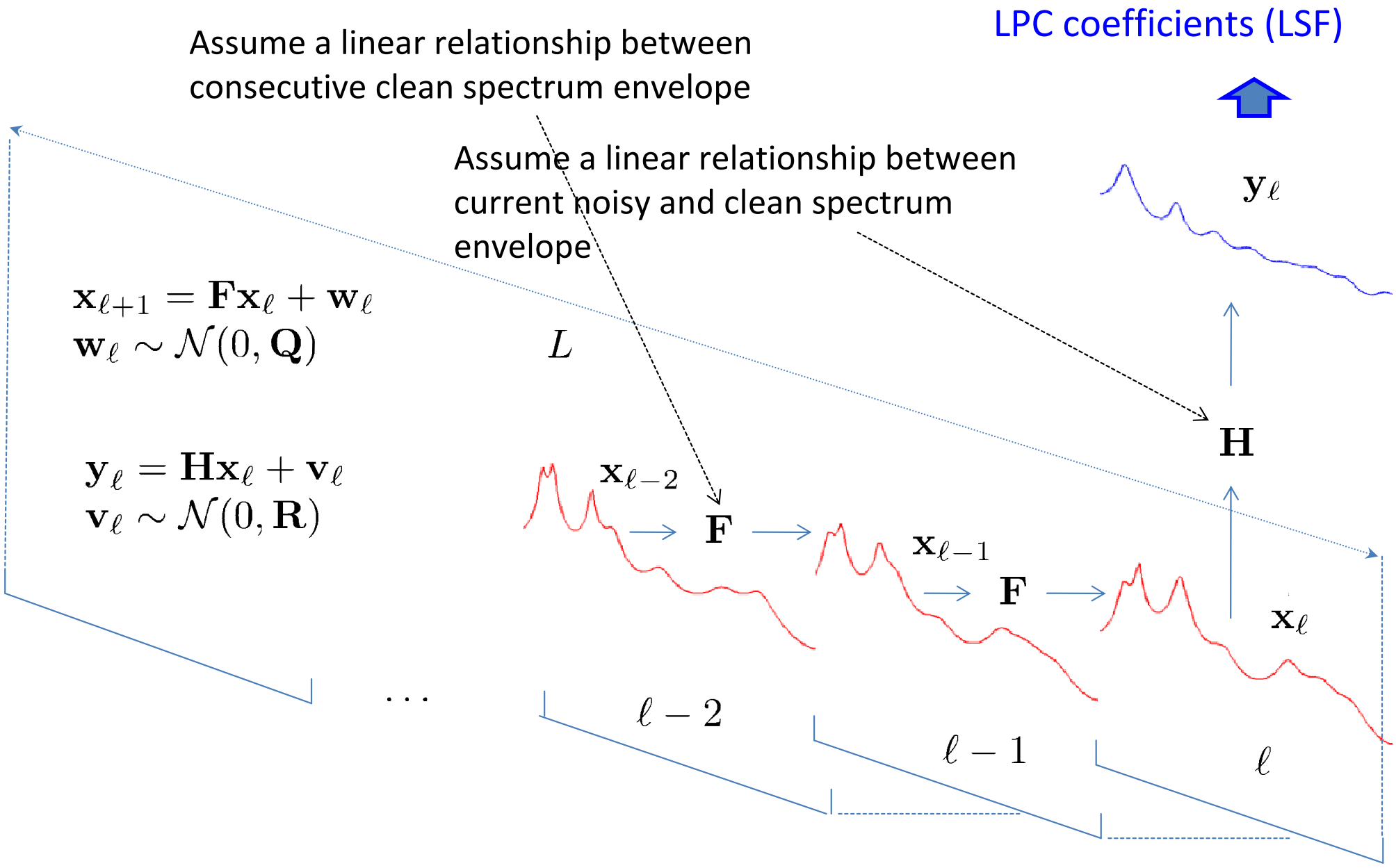
Problem Statement



Estimate clean spectrum envelope by looking for a long term speech evolution

Speech Dynamics Tracking

Linear Dynamical System (LDS)



Speech Dynamics Tracking

Kalman Filter

For each analysis block, we have the observation, noisy LPC coefficients

$$\mathbf{Y} = \{\mathbf{y}_\ell,\ \ell = 1, \ldots, K\}$$

Given Kalman system parameters

$$\Theta = \{\mathbf{F}, \mathbf{H}, \mathbf{Q}, \mathbf{R}, \hat{\mathbf{x}}_1, \boldsymbol{\Sigma}_1\}$$

$$\begin{aligned}
\hat{\mathbf{x}}_{\ell|\ell-1} &= \mathbf{F}\hat{\mathbf{x}}_{\ell-1|\ell-1} \\
\boldsymbol{\Sigma}_{\ell|\ell-1} &= \mathbf{F}\boldsymbol{\Sigma}_{\ell-1|\ell-1}\mathbf{F}^T + \mathbf{Q} \\
\mathbf{e}_\ell &= \mathbf{y}_\ell - \mathbf{H}\hat{\mathbf{x}}_{\ell|\ell-1} \\
\boldsymbol{\Sigma}_{e_\ell} &= \mathbf{H}\boldsymbol{\Sigma}_{\ell|\ell-1}\mathbf{H}^T + \mathbf{R} \\
\mathbf{K}_\ell &= \boldsymbol{\Sigma}_{\ell|\ell-1}\mathbf{H}^T\boldsymbol{\Sigma}_{e_\ell}^{-1} \\
\hat{\mathbf{x}}_{\ell|\ell} &= \hat{\mathbf{x}}_{\ell|\ell-1} + \mathbf{K}_\ell\mathbf{e}_\ell \\
\boldsymbol{\Sigma}_{\ell|\ell} &= \boldsymbol{\Sigma}_{\ell|\ell-1} - \mathbf{K}_\ell\boldsymbol{\Sigma}_{e_\ell}\mathbf{K}_\ell^T
\end{aligned}$$

Kalman filter

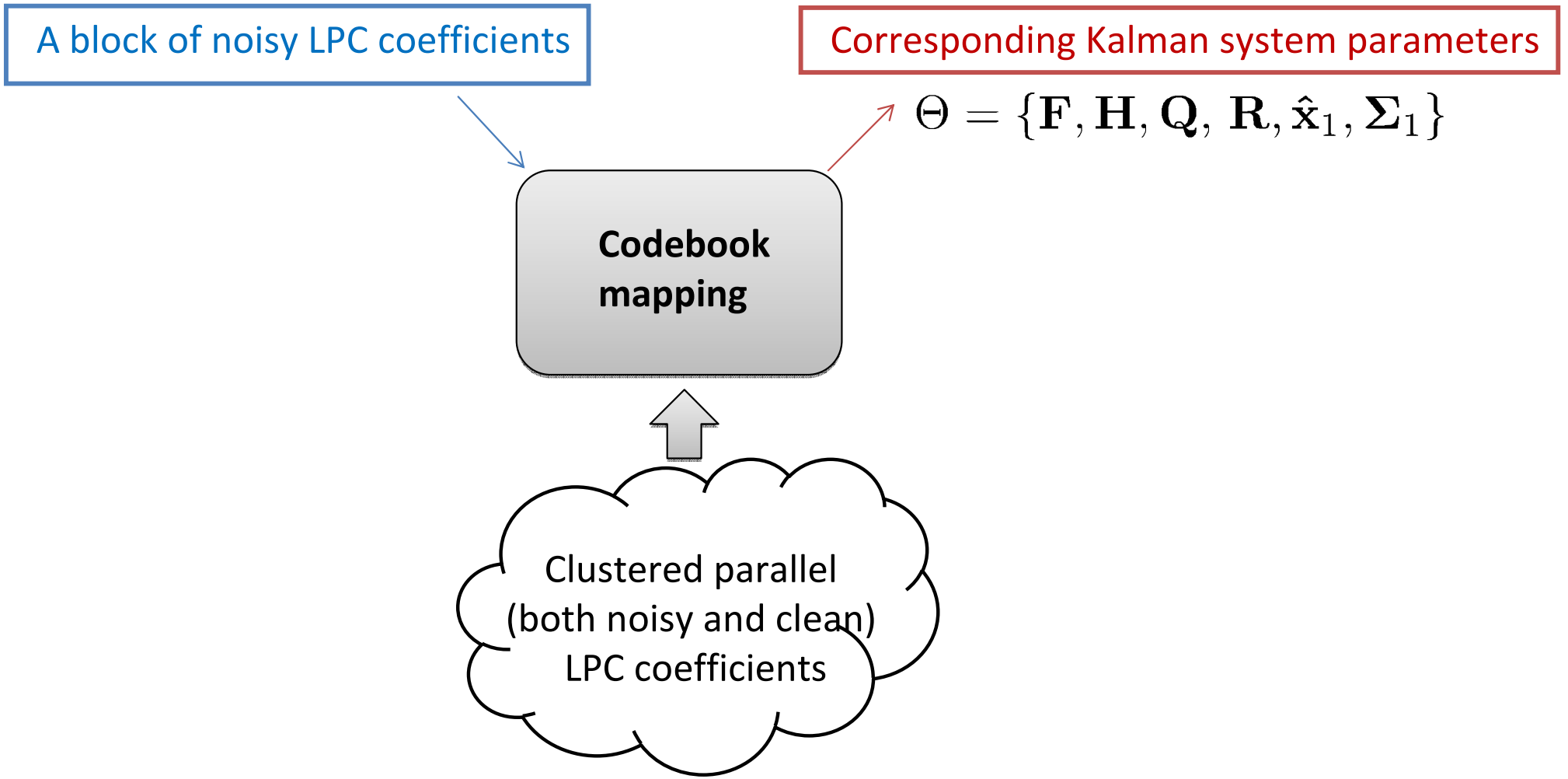
$$\mathbf{X} = \{\mathbf{x}_\ell,\ \ell = 1, \ldots, K\}$$

Desired clean LPC coefficients

Speech Dynamics Tracking

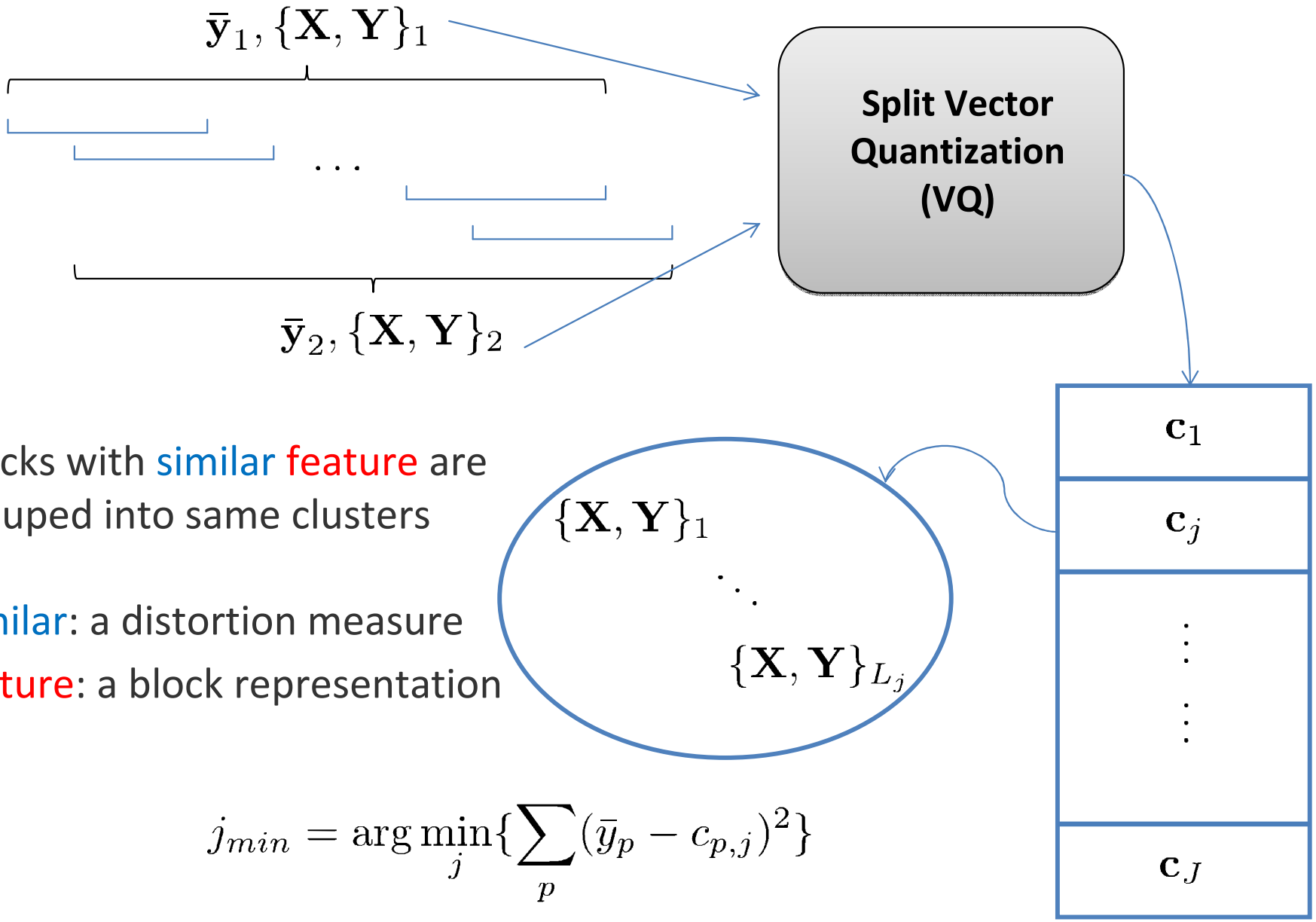
Kalman System Identification

Codebook mapping



Speech Dynamics Tracking

Offline Training



Blocks with similar feature are grouped into same clusters

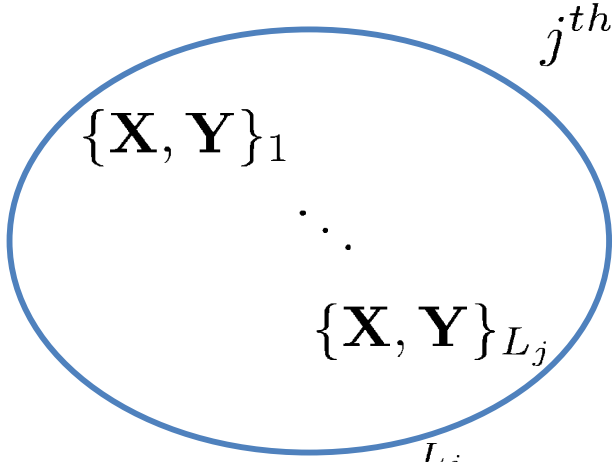
Similar: a distortion measure

feature: a block representation

$$j_{min} = \arg\min_j \{\sum_p (\bar{y}_p - c_{p,j})^2\}$$

Speech Dynamics Tracking

Offline Training (cont.)



Obtain the set of model parameters by minimizing a total negative log-likelihood function of this cluster (Maximum Likelihood)

$$\mathcal{J}_j(\mathbf{X}, \mathbf{Y}, \Theta_j) = -\sum_{i=1}^{L_j} \mathcal{L}(\mathbf{X}_i, \mathbf{Y}_i, \Theta_j)$$

$$= \sum_{i=1}^{L_j}\sum_{l=2}^{N}(\mathbf{x}_\ell^{(i)} - \mathbf{F}\mathbf{x}_{\ell-1}^{(i)})^T\mathbf{Q}^{-1}(\mathbf{x}_\ell^{(i)} - \mathbf{F}\mathbf{x}_{\ell-1}^{(i)})$$

$$+ \sum_{i=1}^{L_j}\sum_{l=1}^{N}(\mathbf{y}_\ell^{(i)} - \mathbf{H}\mathbf{x}_\ell^{(i)})^T\mathbf{R}^{-1}(\mathbf{y}_\ell^{(i)} - \mathbf{H}\mathbf{x}_\ell^{(i)})$$

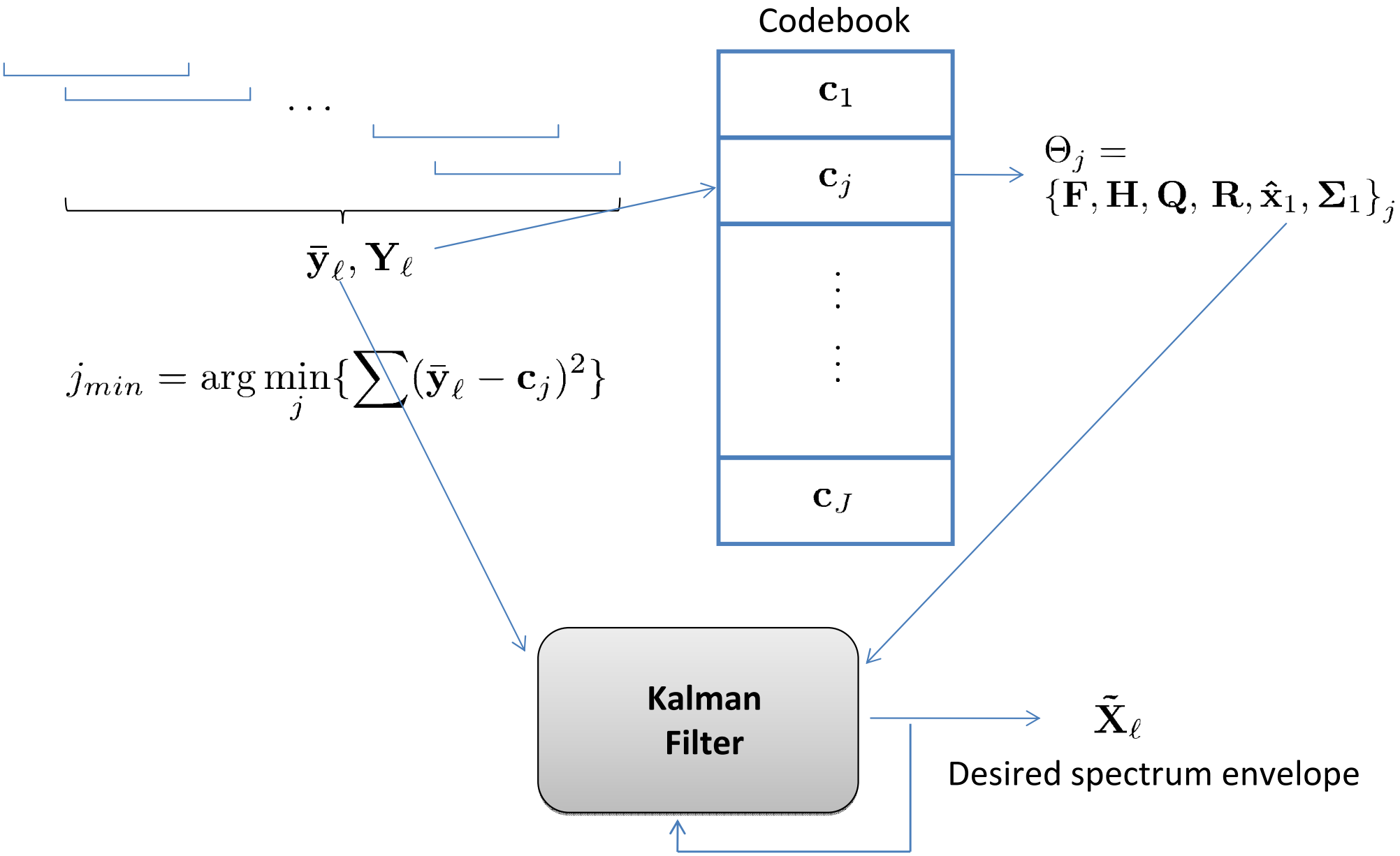
$$+ \sum_{i=1}^{L_j}(\mathbf{x}_1^{(i)} - \hat{\mathbf{x}}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}_1^{(i)} - \hat{\mathbf{x}}_1) + L_j \ln|\mathbf{\Sigma}_1|$$

$$+ L_j(N-1)\ln|\mathbf{Q}| + L_j N \ln|\mathbf{R}| + \text{constant}$$

$$\Theta_j = \{\mathbf{F}, \mathbf{H}, \mathbf{Q}, \mathbf{R}, \hat{\mathbf{x}}_1, \mathbf{\Sigma}_1\}_j$$

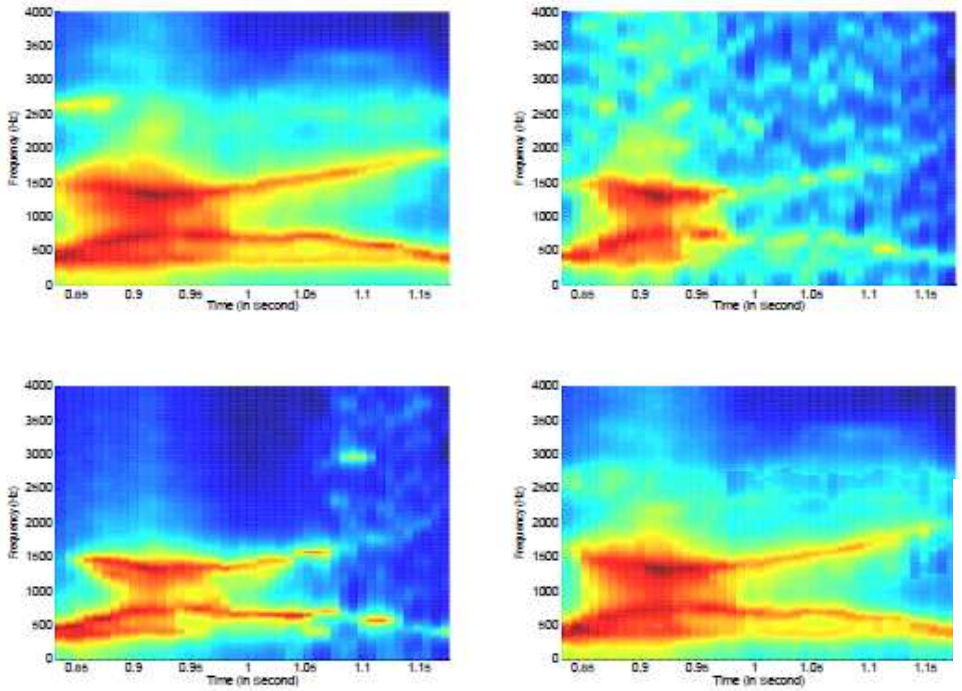
Speech Dynamics Tracking

Online Adaptation

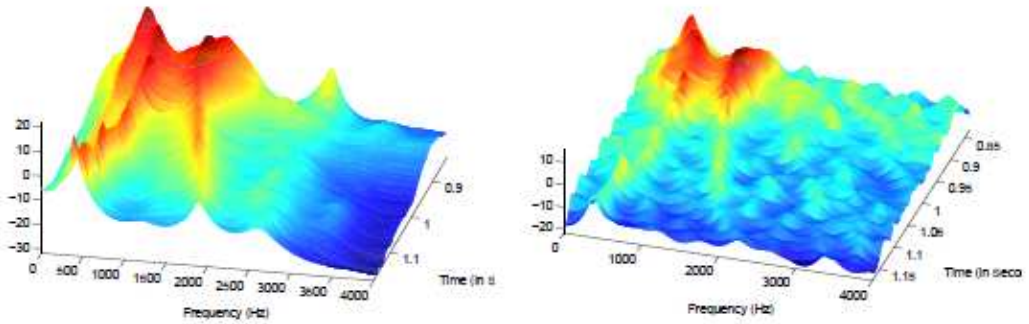


Performance Evaluation

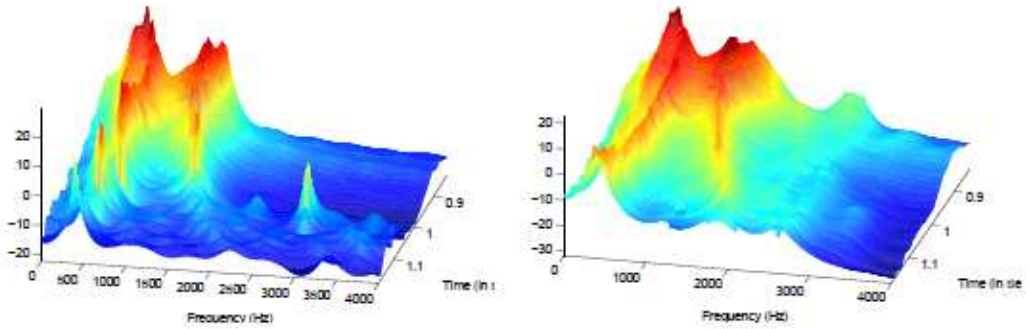
Spectrum Envelope Improvement



- Smooth envelope trajectories without musical tones
- Extended speech evolution for transition period

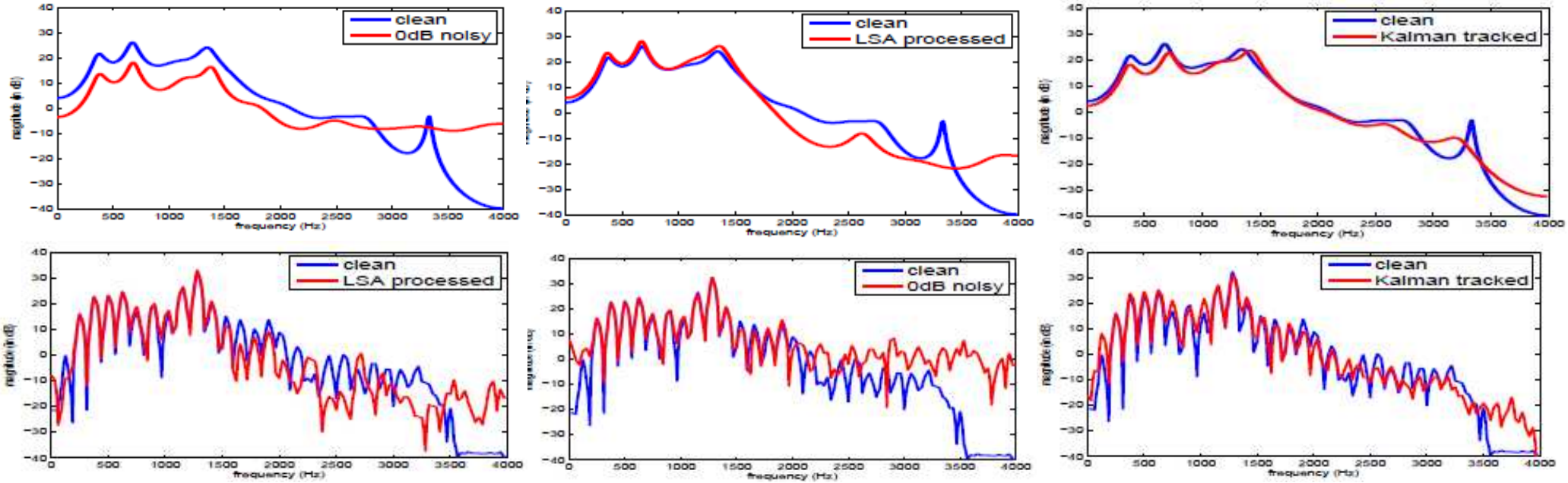


- Expanded formant bandwidth as compared to conventional method
- Close to original temporal envelope trajectories

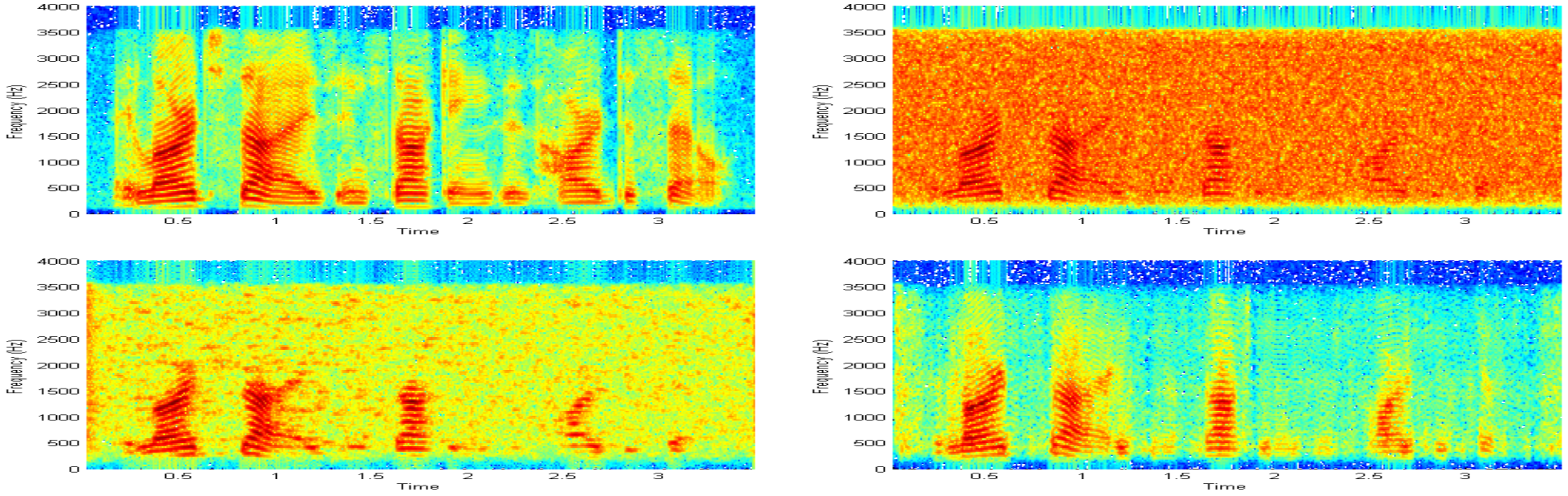


Performance Evaluation

Short-time spectra



Spectrograms



Performance Evaluation

Objective Evaluation

PESQ (1 ~ 4.5)

SNR = 0dB, White Noise

Without enhancement	1.55
Best conventional approach (LSA_SPU)	1.87
Analysis-synthesis approach (pre-cleaned envelope)	2.05
Analysis-synthesis approach (tracked envelope)	**2.41**
Analysis-synthesis approach (clean envelope)	3.17

		Speaker-dependent testing						Speaker-independent testing					
		LSD improvement (in dB)			PESQ improvement (out of 4.5)			LSD improvement (in dB)			PESQ improvement (out of 4.5)		
Noise		Input SNR			Input SNR			Input SNR			Input SNR		
Type	Method	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Gaussian White Noise	STSA	3.81	3.41	3.02	0.31	0.47	0.60	3.79	3.45	3.05	0.32	0.48	0.61
	LSA	3.75	3.36	2.83	0.35	0.54	0.64	3.75	3.39	2.84	0.33	0.49	0.65
	LSA_SPU	3.78	3.39	2.83	0.26	0.37	0.50	3.74	3.42	2.88	0.24	0.35	0.52
	HNM	4.32	4.12	3.89	0.42	0.48	0.49	4.31	4.16	3.77	0.44	0.49	0.48
	KF_GC_HNM	**8.57**	**7.49**	**6.95**	**0.58**	**0.73**	**0.68**	**5.41**	**5.16**	**4.86**	**0.48**	**0.56**	**0.54**
Car Interior Noise	STSA	4.12	3.62	3.25	0.39	0.52	0.58	4.11	3.67	3.24	0.37	0.51	0.56
	LSA	4.09	3.51	3.02	0.42	0.56	0.62	4.05	3.56	3.02	0.42	0.54	0.57
	LSA_SPU	4.13	3.52	2.99	0.37	0.42	0.54	4.12	3.54	2.99	0.36	0.41	0.53
	HNM	4.41	4.22	3.96	0.41	0.46	0.47	4.39	4.19	3.95	0.41	0.48	0.43
	KF_GC_HNM	**8.44**	**7.52**	**6.97**	**0.52**	**0.69**	**0.68**	**5.48**	**5.18**	**4.90**	**0.45**	**0.55**	**0.58**
F16 Cockpit Noise	STSA	3.96	3.41	3.02	0.38	0.30	0.45	3.99	3.43	3.11	0.36	0.31	0.43
	LSA	3.91	3.36	2.83	0.49	0.48	0.56	3.88	3.38	2.94	**0.45**	0.47	0.55
	LSA_SPU	3.99	3.39	2.89	0.36	0.39	0.62	4.02	3.41	2.95	0.35	0.36	0.59
	HNM	4.62	4.12	3.89	0.43	0.47	0.45	4.59	4.10	3.95	0.41	0.45	0.44
	KF_GC_HNM	**8.50**	**7.50**	**6.93**	**0.55**	**0.71**	**0.69**	**5.57**	**5.20**	**4.92**	0.43	**0.52**	**0.57**

Conclusion

- Effect of noise corruption and conventional speech enhancement are discussed
- A novel analysis-synthesis approach is presented
- A speech dynamic tracking approach that incorporates VQ training and Kalman filtering is proposed
- Improved spectrum envelope estimation is illustrated
- Objective results in terms of LSD and PESQ score are shown

Thank You !

Q & A