Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones

Mingzi Li
Department of Electrical Engineering
Supervised by: Prof. Israel Cohen
November 2013

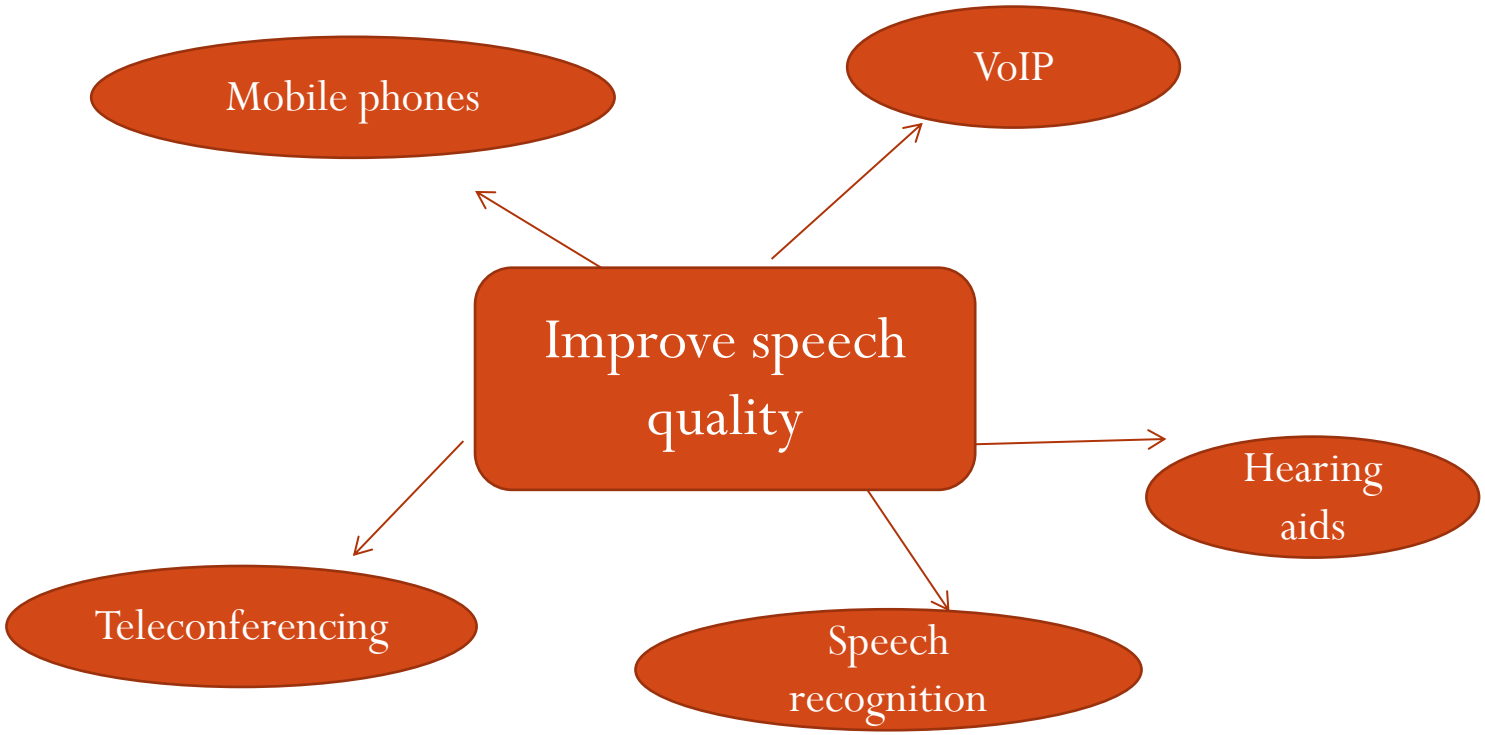
Outline

- Introduction
- Review of existing methods
- Probabilistic approach
- Geometric extension approach
- Summary

Introduction

- Speech enhancement
- Multi-sensory speech enhancement
- Bone-conducted microphone
- Research objectives

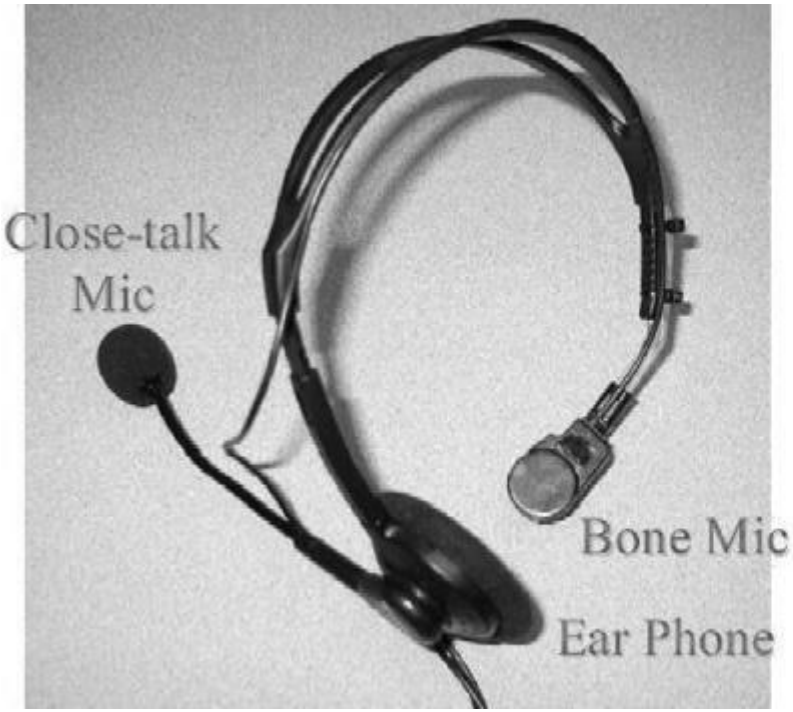
Speech enhancement



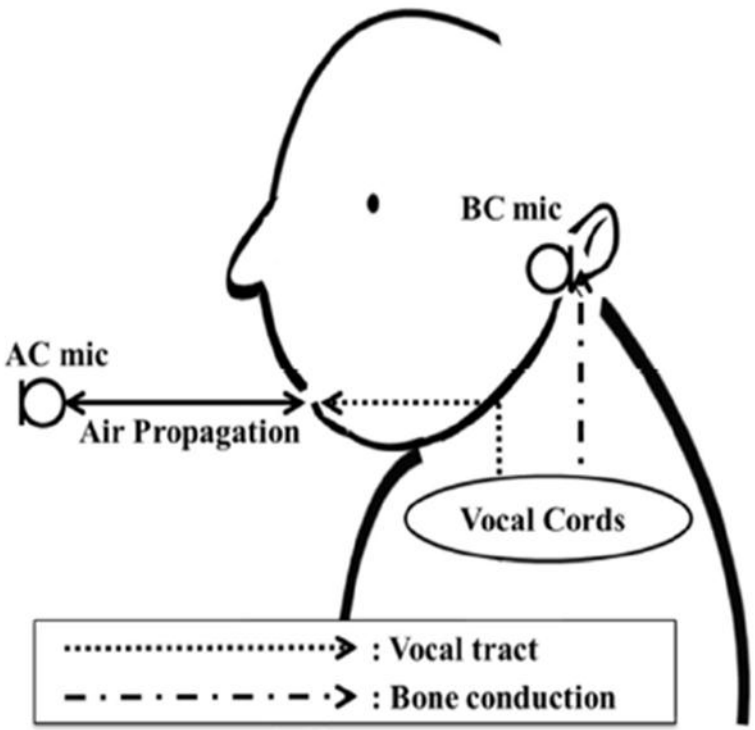
Multi-sensory speech enhancement

- Audio-visual speech processing (G. Potamianos et al,2004)
- Air and throat microphones (M. Graciarena et al,2003)
- Ear plug (O. M. Strand et al, 2003)
- Stethoscope device (P. Heracleous et al, 2003)
- Aliph's Jawbone headsets
- Electromagnetic motion sensor (GEMS) (G. C. Burnett,1999)
- Physiological microphone (P-Mic) (M.V. Scanlon,1998)
- Electroglottograph (EGG) (M. Rothenberg,1992)
- Bone-Conducted Microphone (T.Yanagisawa et al, 1975)

Bone-conducted microphone



Conducting path (K. Kondo et al,2006)



Conducting path of AC and BC microphones

Model

$$Y_t = X_t + V_t + U_t \quad \rightarrow AC$$

$$B_t = H_t X_t + G_t V_t + W_t \quad \rightarrow BC$$

$$X_t \quad \rightarrow Clean$$

$$H_t, G_t \quad \rightarrow Transfer\ function$$

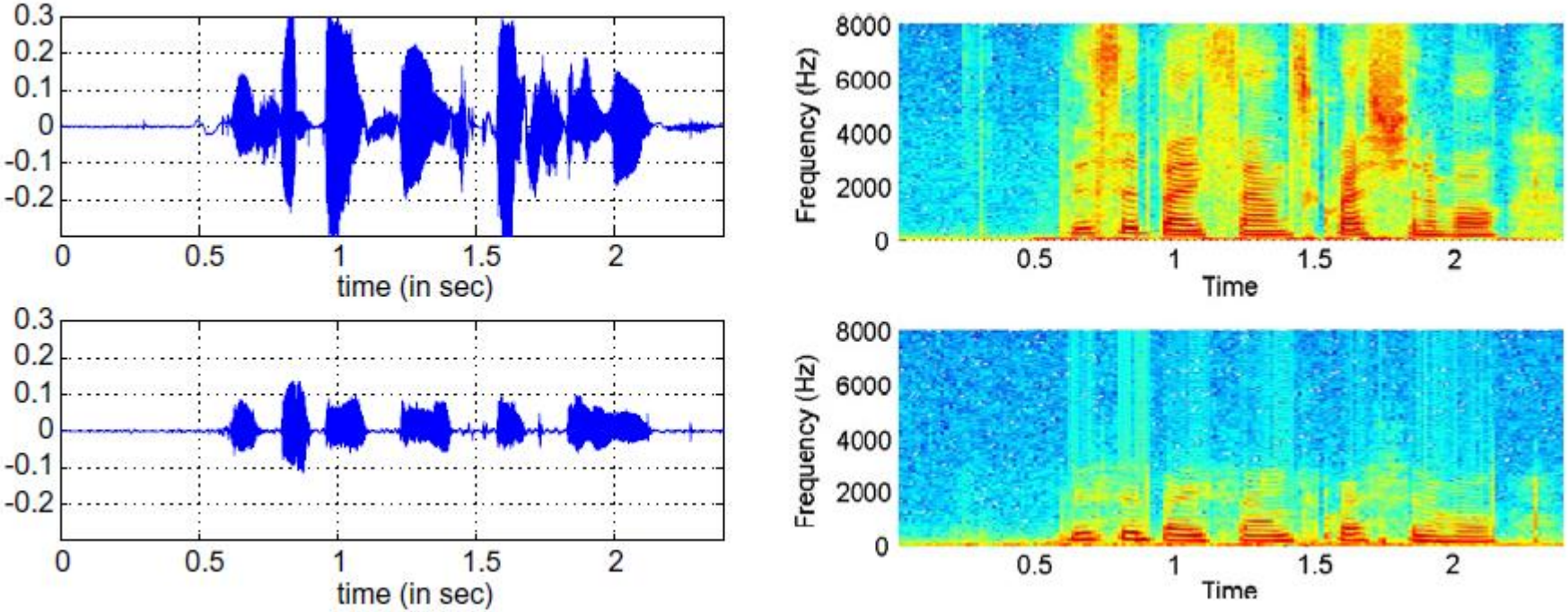
$$V_t \quad \rightarrow Noise$$

$$U_t \quad \rightarrow AC\ \ sensor\ Noise$$

$$W_t \quad \rightarrow BC\ \ sensor\ Noise$$

- Bone conducted:
- Less noise & low frequency
- Air conducted:
- More noise & complete frequency

Signal and spectrogram (A. Subramanya et al,2008)



Waveforms and spectrograms of the signals captured by the ABC microphone. The first row shows the signal captured by the air microphone and the second row shows the signal captured by the bone microphone.

Research objectives

- BC microphone as a dominant sensor:
  - Geometric harmonics method
  - Laplacian pyramid method
- Compare the proposed methods with an existing method

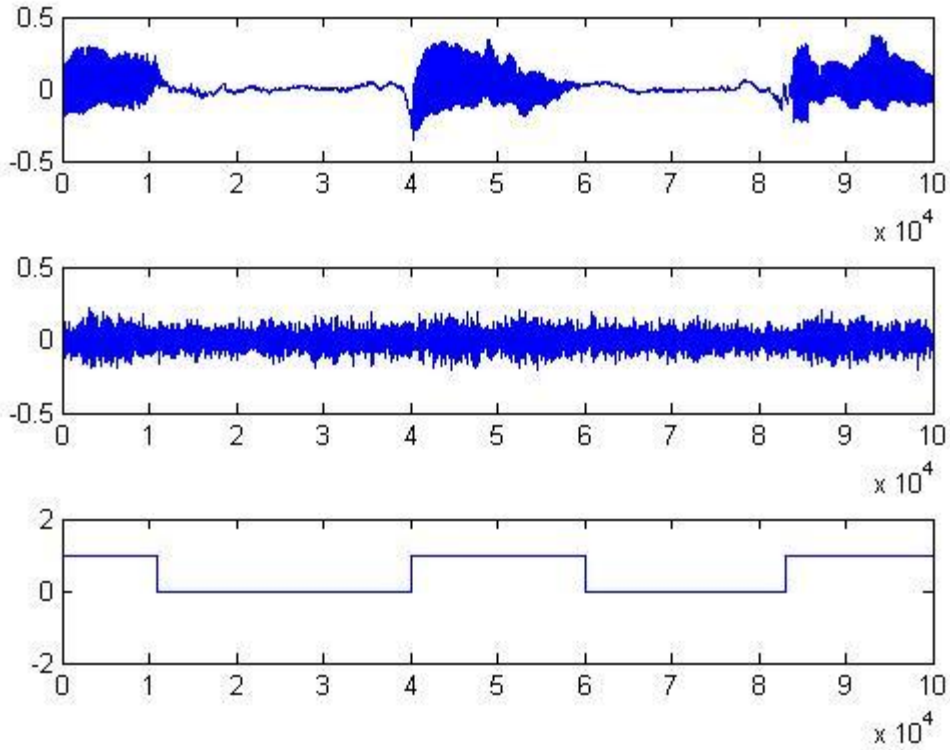
Review of existing methods

- BC microphone as a supplementary sensor
- BC microphone as a dominant sensor

Methods

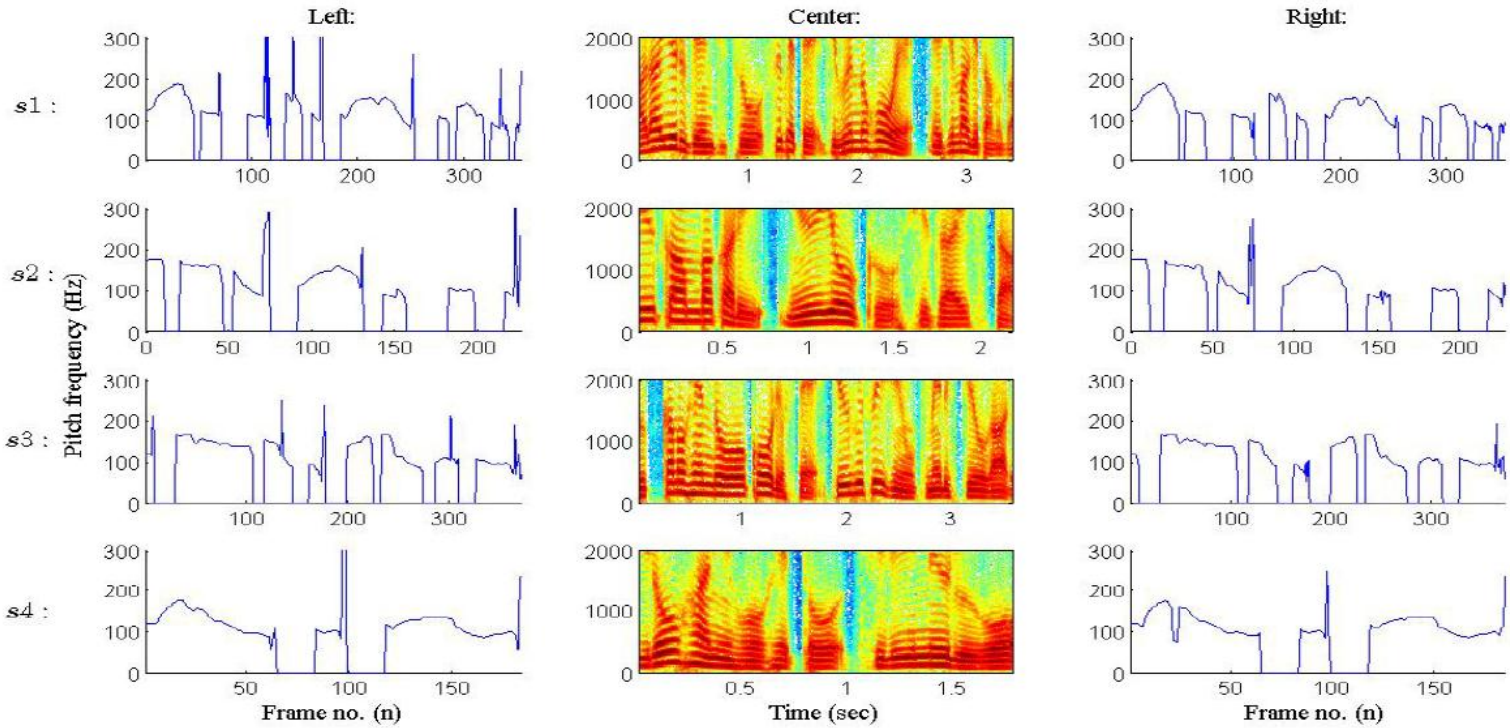
BC m as a supplementary sensor	BC m as a dominant sensor
BC m for voice activity detection	Equalization: IDFT; DFT; LMS
BC m for pitch detection	Analysis and synthesize: LP; LSF (Neural Network)
BC m for low frequency enhancement	Probabilistic: ML; MMSE (DBN)

Voice activity detection (M.Zhu et al,2007)



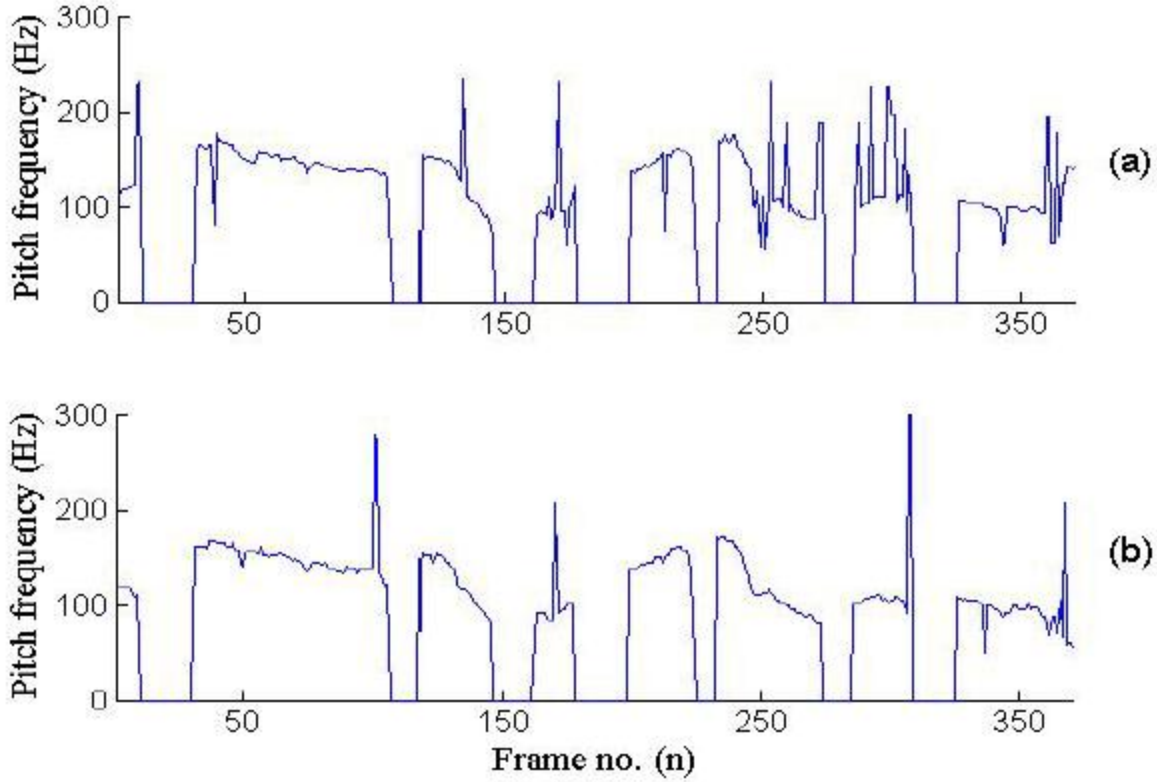
Speech detection using bone sensors: The top figure illustrates the speech signal captured by the bone sensor when two people are talking at the same time. The middle figure shows the signal captured by the regular microphone and bottom figure presents the detection result.

Pitch detection (M. S. Rahman et al, 2010)



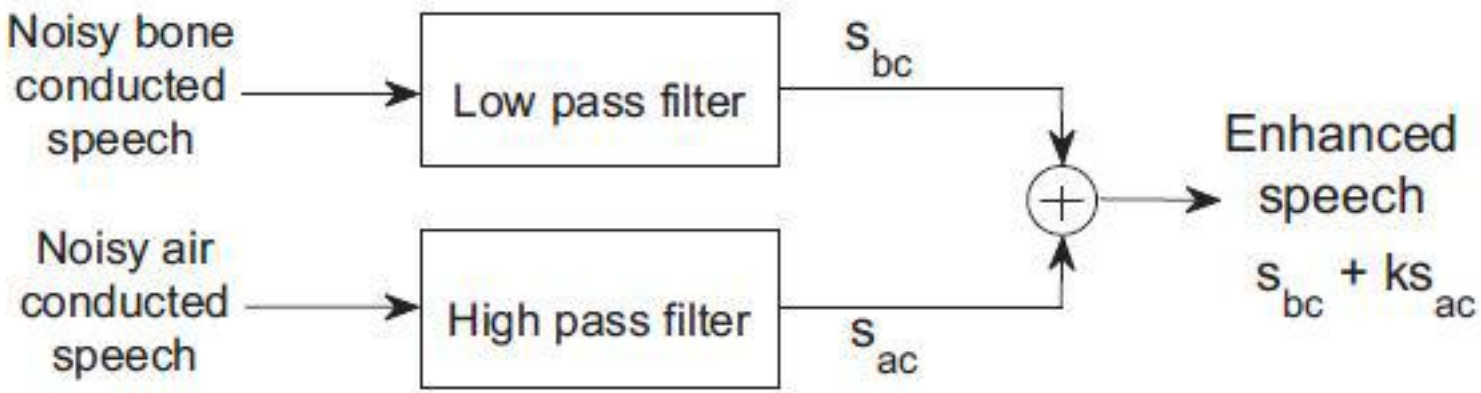
Left: Pitch tracking of air-conducted speech in noiseless condition. Center: Speech spectrogram. Right: Pitch tracking of bone-conducted speech in noiseless condition. The experiments have been conducted on four speeches.

Pitch detection (Cont.)



Pitch contours estimated from speech when corrupted by noise. a) pitch contours estimated from air-conducted speech, b) pitch contours estimated from bone-conducted speech.

BC m for low frequency enhancement (M. S. Rahman,2011)



Block diagram when BC Speech is used for low frequency enhancement.

Equalization

- IDFT (T. Shimamura et al,2005)

$$\hat{h}^{INV} = IDFT\left[\left|X_t^{AC}\right| / \left|X_t^{BC}\right|\right]$$

- DFT (K. Kondo,2006)

$$\hat{H}^{INV} = E\left\{\left|X_t^{AC}\right| / \left|X_t^{BC}\right|\right\}$$

- Least Mean Square (LMS) filter (T. Shimamura,2006)

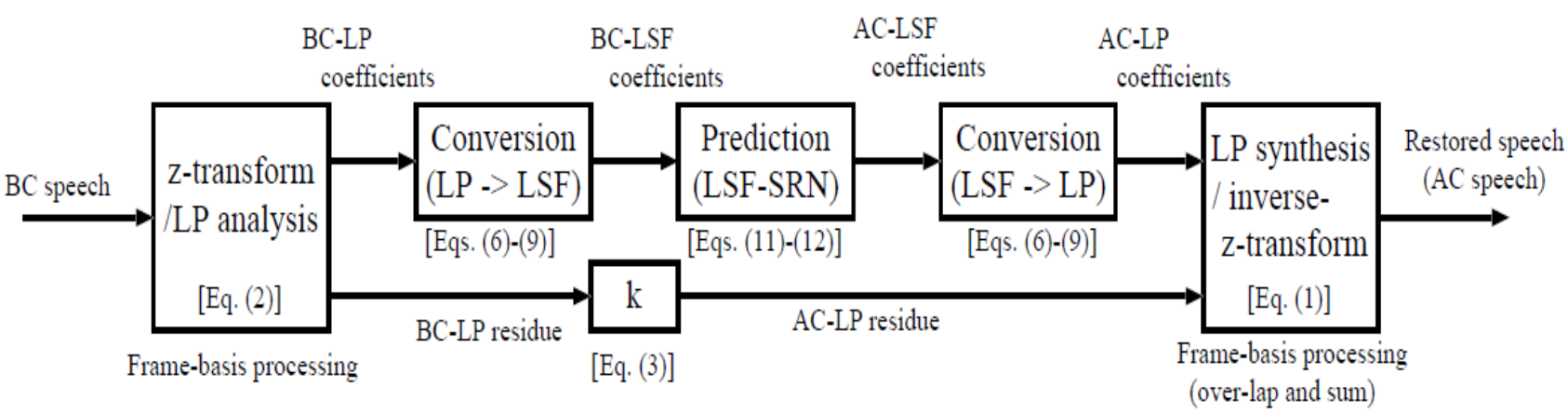
$$\hat{h}^{INV}(n) = \hat{h}^{INV}(n-1) + 2\mu e(n-1)b(n-1)$$

Analysis and synthesize

- Linear prediction (LP) filter (T.T.Vu,2006)

$$\hat{H}^{INV} = \frac{Z\left[e^{AC}(n)\right]\sum_{i=0}^{Q} a^{AC}(i)z^{-i}}{Z\left[e^{BC}(n)\right]\sum_{i=0}^{P} a^{BC}(i)z^{-i}}$$

- Line spectral frequency (LSF) filter (T.T.Vu,2008)



Probabilistic

- Maximum likelihood estimation (MLE) (Z.Liu et al, 2004)

$$R=\sum_{n=l}^{N}\left\{\begin{array}{l}\left|Y(l,k)-X(l,k)\right|/2\sigma_V^2\\+\left|B(l,k)-H(l,k)X(l,k)\right|/2\sigma_W^2\end{array}\right\}$$

- MMSE estimator (A. Subramanya et al, 2005)

$$\hat{X}(l,k)=\sum_{t=0}^{1}\left\{\begin{array}{l}P\left(T(l)=t\left|Y(l,k),B(l,k)\right.\right)\\\cdot E\left\{X(l,k)\left|Y(l,k),B(l,k),T(l)=t\right.\right\}\end{array}\right\}$$

- Dynamic Bayesian Network (DBN) (A. Subramanya et al,2008)

$$p(X_t|Y_t,B_t)=\sum_{s,m}p(X_t,S_t=s,M_t=m|Y_t,B_t)$$

$$=\sum_{s,m}p(X_t|Y_t,B_t,S_t=s,M_t=m)p(M_t=m|Y_t,B_t,S_t=s)\,p(S_t=s|Y_t,B_t)$$

Probabilistic approach

- Model
- Network description
- Transfer function & leakage factor
- MMSE estimator
- Result

Model

- Air conducted(AC): $Y_t = X_t + V_t + U_t$
- Bone conducted(BC): $B_t = H_t X_t + G_t V_t + W_t$
- Background noise: $V_t \sim N(0, \sigma_v^2)$
- AC Sensor noise: $U_t \sim N(0, \sigma_u^2)$
- BC Sensor noise: $W_t \sim N(0, \sigma_w^2)$
- Optimal linear mapping: H_t
- Leakage of noise: G_t

Assumption

- Feature: Magnitude-normalized complex spectra

$$\tilde{X}_t = \frac{X_t}{\|X_t\|}$$

- Training: Speech model(k-means)

$$d\left(\tilde{X}_i, \tilde{X}_j\right) = \left\| \left[d\left(\tilde{X}_i^1, \tilde{X}_j^1\right), \ldots, d\left(\tilde{X}_i^f, \tilde{X}_j^f\right) \ldots, d\left(\tilde{X}_i^N, \tilde{X}_j^N\right) \right]^T \right\|, 1 \le i, j \le T$$

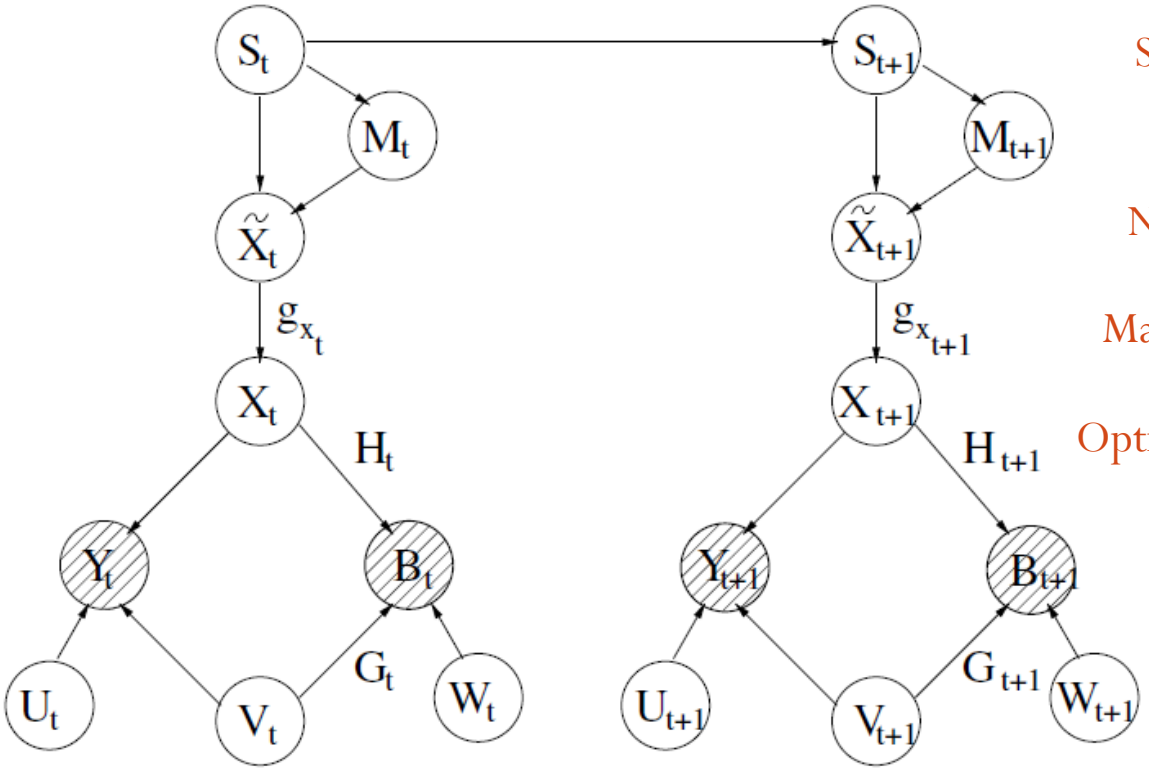
$$d\left(\tilde{X}_i^f, \tilde{X}_j^f\right) = \log\left|\tilde{X}_i^f\right| - \log\left|\tilde{X}_j^f\right|, \ 1 \le f \le N$$

- Method: dynamic Bayesian network (DBN)

$$\begin{aligned} & p(Y_t, B_t, X_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) \\ & = p(Y_t | X_t, V_t, U_t) p(B_t | X_t, V_t, W_t) p(X_t | \tilde{X}_t) p(\tilde{X}_t | M_t, S_t) \\ & \times p(M_t |, S_t) p(S_t) p(V_t) p(U_t) p(W_t) \end{aligned}$$

Network description

- Dynamic Bayesian network



Speech/non-speech

Mixture index

Normalized speech

Match the clean speech

Optimal linear mapping

Leakage noise

Transfer function & leakage factor

- Transfer function (non-speech)

$$G_t = \frac{\sum_{t\in N_v}(\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2) + \sqrt{\left(\sum_{t\in N_v}(\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2)\right)^2 + 4\sigma_v^2\sigma_w^2\left|\sum_{t\in N_v} B_t^* Y_t\right|}}{2\sigma_v^2\sum_{t\in N_v} B_t^* Y_t}$$

- Leakage factor (speech)

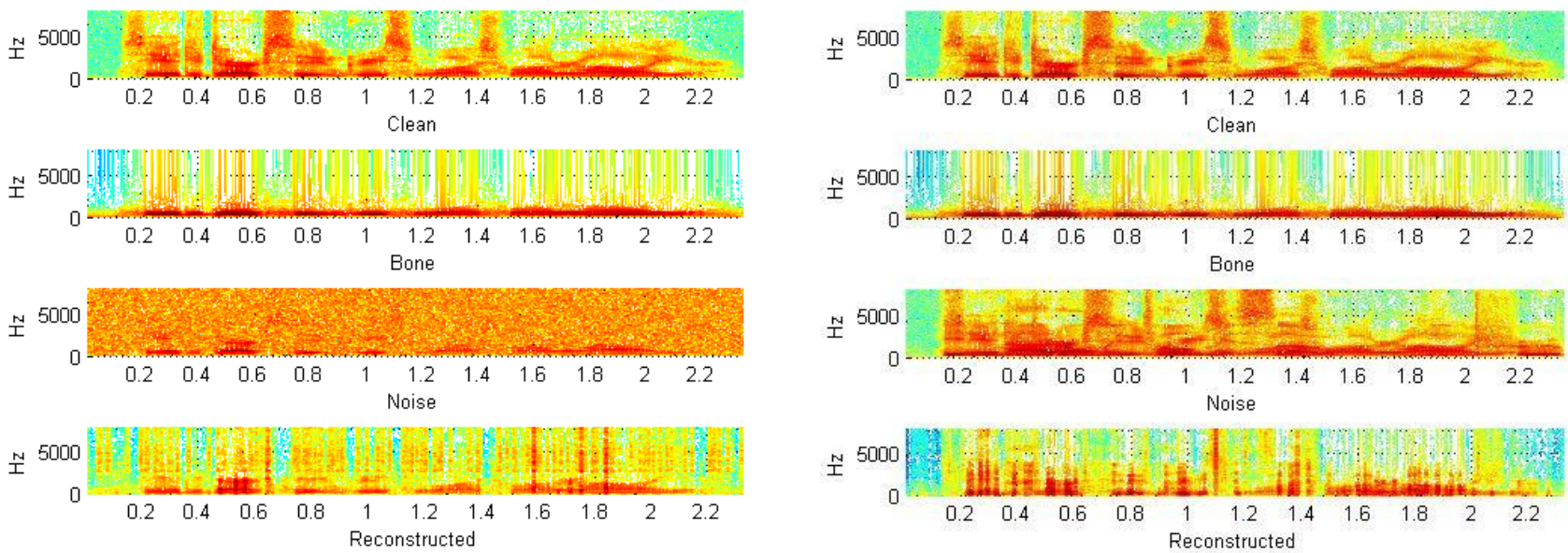
$$H_t = G_t + \frac{\sum_{t\in N_s}(\sigma_v^2|B'_t|^2 - \sigma_w^2|Y_t|^2) + \sqrt{\left(\sum_{t\in N_s}(\sigma_v^2|B'_t|^2 - \sigma_w^2|Y_t|^2)\right)^2 + 4\sigma_v^2\sigma_w^2\left|\sum_{t\in N_s}(B'_t)^* Y_t\right|}}{2\sigma_v^2\sum_{t\in N_s}(B'_t)^* Y_t}$$

MMSE estimator & result

- Estimator

$$\hat{\tilde{X}}_t = E(\tilde{X}_t | Y_t, B_t) = p(S_t = 0 | Y_t, B_t) E(\tilde{X}_t | Y_t, B_t, S_t = 0, M_t = 0)$$
$$+ p(S_t = 0 | Y_t, B_t) \sum_m p(M_t = m | Y_t, B_t, S_t = 1) E(\tilde{X}_t | Y_t, B_t, S_t = 1, M_t = m)$$

- Result



Spectrogram of clean, BC, noisy AC and reconstructed speech: Left: Gaussian noise, Right: interfering speaker

Geometric extension approach

- Model
- Nyström extension method
- Geometric harmonics
- Laplacian pyramid estimation
- Result

Model

- Train: (Mapping from concatenation of noisy AC and BC speech to clean speech.)

$$YB_t = \begin{bmatrix} Y_t \\ B_t \end{bmatrix} \xrightarrow{f:R^{512} \to R^{256}} X_t$$

- Test: (Extension of the mapping from concatenation of noisy AC and BC speech to clean speech.)

$$YB^*_t = \begin{bmatrix} Y^*_t \\ B^*_t \end{bmatrix} \xrightarrow{f:R^{512} \to R^{256}} X^*_t$$

Nyström extension method (C.T.H. Baker, 1977)

- Goal: extend relevant "information" about a large dataset in a high dimensional space.
- Method: find a low-rank approximation to a symmetric, positive semi-definite kernel.
- In essence: only use partial information about the kernel to solve a simpler eigenvalue problem, and then to extend the solution using complete knowledge of the kernel.

Nyström extension method (C.T.H. Baker, 1977)

- Eigen function approximation

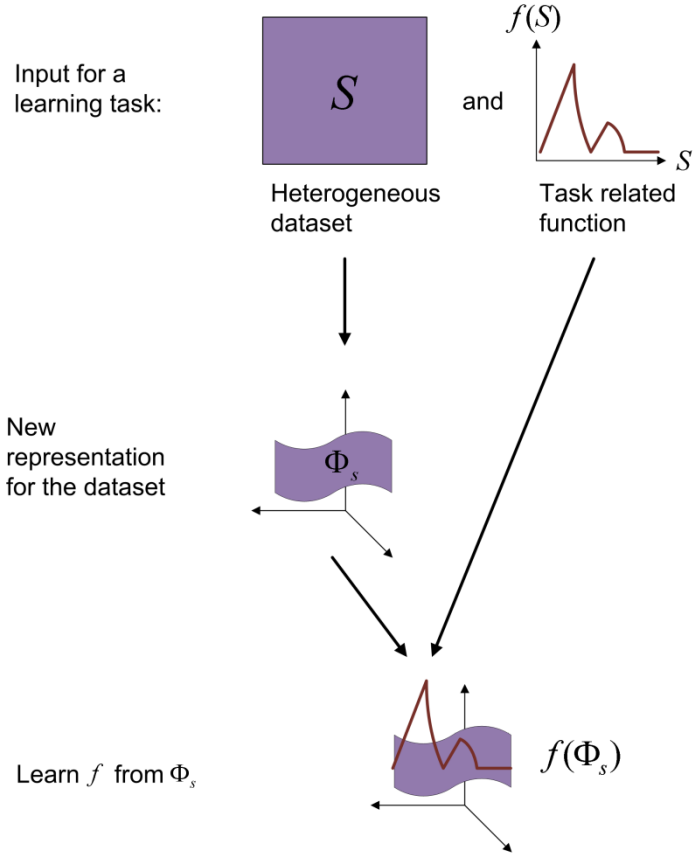
$$\int_a^b G(x,y)\phi(y)dy = \lambda\phi(x)$$

$$\frac{b-a}{n}\sum_{j=1}^{n} G(x_i,x_j)\phi(x_j) = \lambda\phi(x_i)$$

- Nyström extension

$$\hat{\phi}(x_*) \triangleq \frac{b-a}{n\lambda}\sum_{j=1}^{n} G(x_*,x_j)\phi(x_j)$$

$$f_* \triangleq \sum_{i=1}^{n} \left(f^T \cdot \phi_i\right) \hat{\phi}_i\left(x_*\right)$$



Scheme of learning functions (N. Rabin, 2012)

Geometric harmonics (GH) (R.R. Coifman, 2006)

- Definition

$$\Psi_j(\bar{x}) = \frac{1}{\lambda_j} \int_X k(\bar{x}, y)\psi_j(y)\,\mathrm{d}\mu(y).$$

- Example
  - Gaussian extension $e^{-B^2\|x-y\|^2}$
  - Harmonic extension $k(x,y) = \begin{cases} -\log(\|x-y\|) & \text{if } n=2, \\ \frac{1}{\|x-y\|^{n-2}} & \text{if } n \geqslant 3. \end{cases}$
  - Wavelet extension $k(x,y) = \sum 2^{-nj}\Phi(2^j x - m)\Phi(2^j y - m)$

Geometric harmonics (GH) (R.R. Coifman, 2006)

- Eigenvector approximation

$$\lambda_l \varphi_l(x_i) = \sum_{x_j \in \Gamma} e^{\frac{-\|x_i - x_j\|^2}{2\epsilon}} \varphi_l(x_j), \ x_i \in \Gamma.$$

- Extension

$$\varphi_l(y) = \frac{1}{\lambda_l} \sum_{x_j \in \Gamma} e^{\frac{-\|y - x_j\|^2}{2\epsilon}} \varphi_l(x_j), \ y \in \mathbb{R}^m.$$

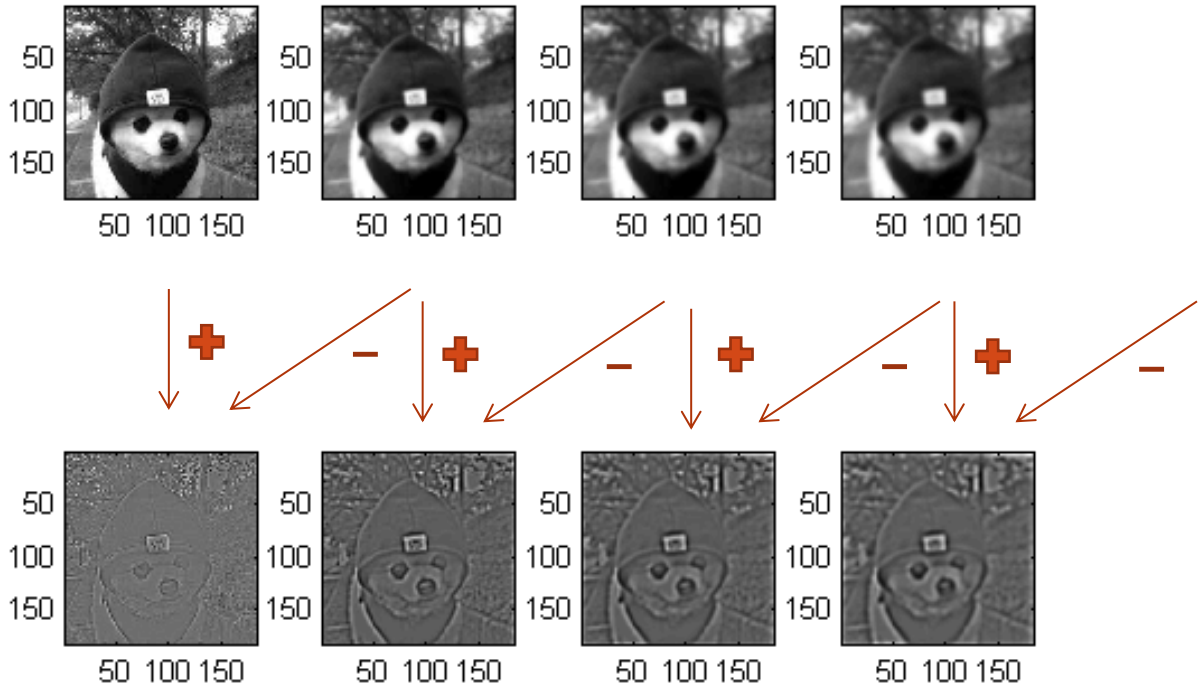
$$f(y) = \sum_l \langle \varphi_l, f \rangle \varphi_l(y), \ y \in \mathbb{R}^m.$$

$$\tilde{f}(y) = \sum_{\lambda_l \geq \delta \lambda_0} \langle \varphi_l, f \rangle \bar{\varphi}_l(y), \ y \in \mathbb{R}^m.$$

Comments of GH

- Need to tune the parameters ε, l.
- Extension of the function is not the original function but the projection of the function.
- The extension range has relation to the complexity of the function.

Laplacian pyramid (LP) (Burt and Adelson,1983)



Laplacian pyramid (LP) (N. Rabin, 2012)

- Algorithm:
  - Kernel: $W_0 = \exp(-|x_i - x_j|^2 / \sigma_0)$ $\quad K_0 = q_0^{-1}(x_i) w_0(x_i, x_j)$
  - Iteration:

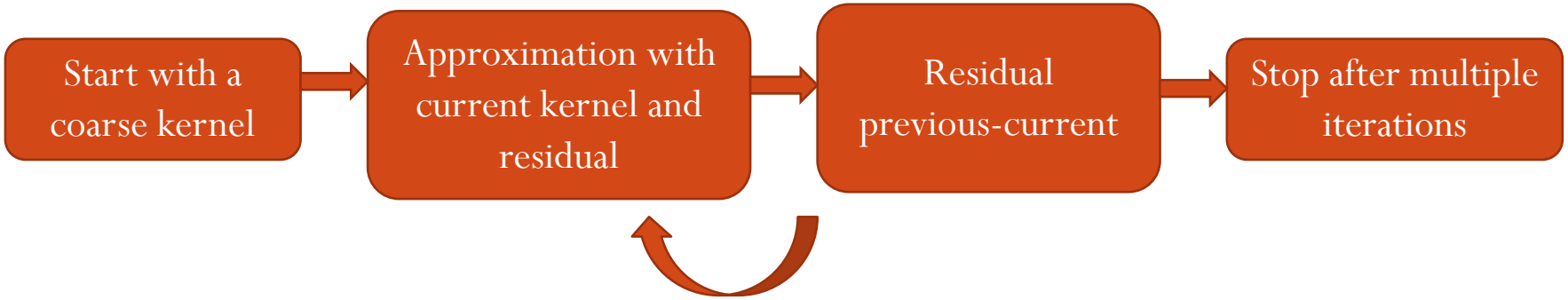
$$s_0(x_k) = \sum_{i=1}^{n} k_0(x_i, x_k) f(x_i)$$

$$d_1(x_k) = f(x_k) - s_0(x_k) \quad d_l(x_k) = f(x_k) - \sum_{i=1}^{l-1} s_i(x_k)$$

$$W_l = \exp(-|x_i - x_j|^2 / \frac{\sigma_0}{2^l}) \quad K_l = q_l^{-1}(x_i) w_l(x_i, x_j)$$

$$s_l(x_k) = \sum_{i=1}^{n} k_l(x_i, x_k) d_l(x_i)$$

  - Estimation: $f(x_k) = \sum_l s_l(x_k)$



Comments of LP

- Kernel method

$$\hat{f}(x_k) = \sum_{i=1}^{n} k_0(x_i, x_k) f(x_i)$$

- Improve(Iterate)
  - Diffusion $\hat{f}_{l+1}(x_k) = \hat{f}_l(x_k) + (K_0 - I)\,\hat{f}_l(x_k)$
  - Residual $\hat{f}_{l+1}(x_k) = \hat{f}_l(x_k) + K_0\left(f(x_k) - \hat{f}_l(x_k)\right)$
  - LP $\hat{f}_{l+1}(x_k) = \hat{f}_l(x_k) + K_l\left(f(x_k) - \hat{f}_l(x_k)\right)$

- Statistic analysis

$$\text{Model}: y(x_k) = f(x_k) + n_k$$

$$E\left[s_l(x_k)\right] = E\left[\sum_{i=1}^{n} k_l(x_i, x_k)\left(f(x_i) + n_i - \sum_{i=1}^{l-1} s_i(x_i)\right)\right] = s_l(x_k)$$

$$Bias = E\left[\hat{f}(x_k)\right] - f(x_k) = E\left[\sum s_l(x_k)\right] - f(x_k) = \sum s_l(x_k) - f(x_k) = e_s$$

$$MSE = E\left[\hat{f}(x_k) - f(x_k)\right]^2$$

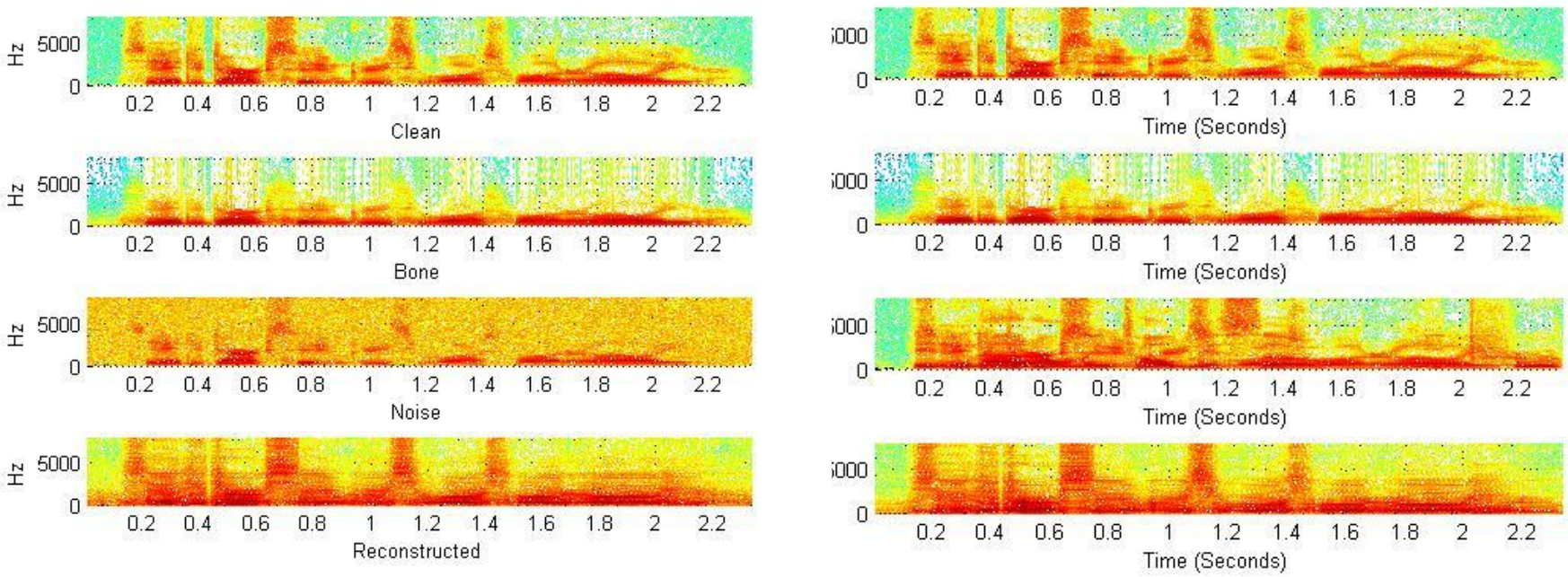
$$= E\left[\sum_{i=1}^{n} k_0(x_i, x_k)\left(f(x_i) + n_i\right) + \ldots + \sum_{i=1}^{n} k_l(x_i, x_k)\left(f(x_i) + n_i - \sum_{i=1}^{l-1} s_i(x_i)\right) - \sum_{i=1}^{l-1} s_i(x_k) - e_s\right]^2$$

$$= E\left[\sum_{i=1}^{n} k_0(x_i, x_k) n_i + \ldots + \sum_{i=1}^{n} k_l(x_i, x_k) n_i - e_s\right]^2$$

$$= E\left[\sum_{i=1}^{n} k_0^2(x_i, x_k) n_i^2 + \ldots + \sum_{i=1}^{n} k_l^2(x_i, x_k) n_i^2 + e_s^2\right]$$

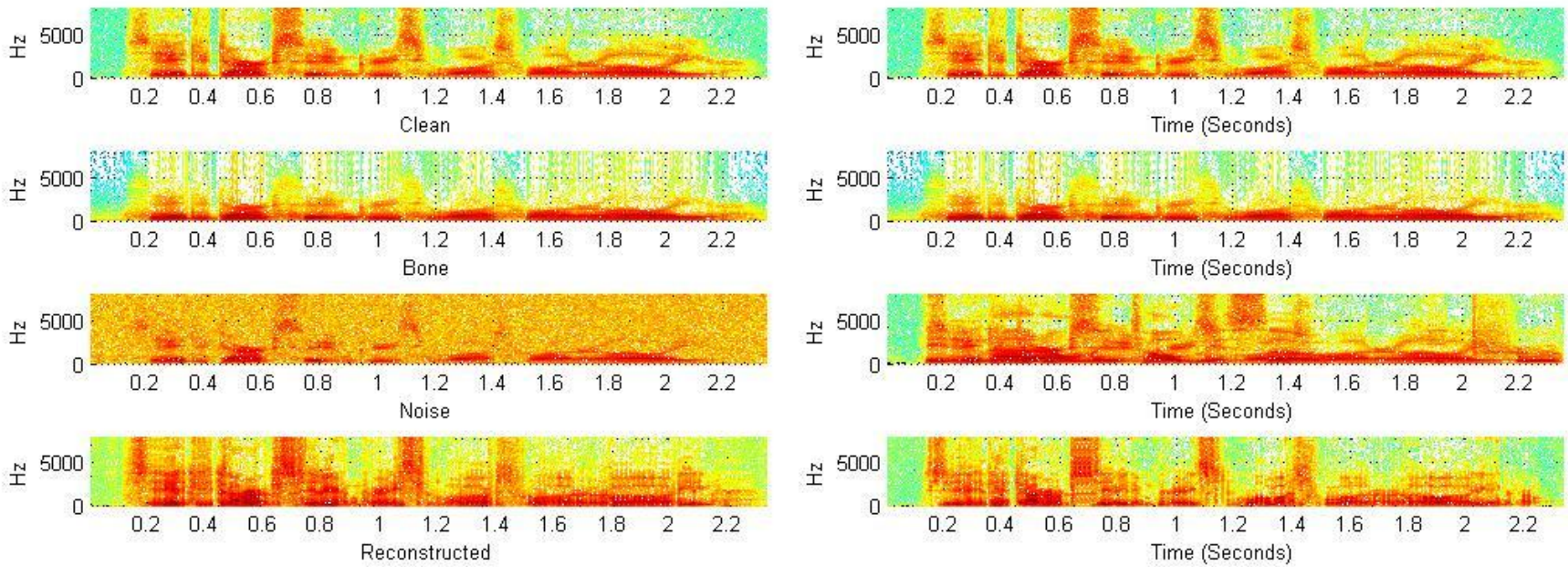
$$= E\left[\sum_{i=1}^{n}\sum_{l=1}^{L} k_l^2(x_i, x_k) n_i^2 + e_s^2\right] = -\sum_{i=1}^{n} \frac{Ei(-2^l \ln(k_l^2(x_i, x_k)))}{\ln 2}\sigma_i^2 + e_s^2$$

Result (GH)



Spectrogram of clean, BC, noisy AC and reconstructed speech: Left: Gaussian noise, Right: interfering speaker

Result (LP)



Spectrogram of clean, BC, noisy AC and reconstructed speech: Left: Gaussian noise, Right: interfering speaker

Comparison of Log Spectral Distortion

LSD	GH	LP	OM-LSA	PA
SNR =0	1.5726	1.1003	2.0901	1.9613
SNR = 10	1.5564	1.1028	1.4979	2.1592
SNR = 20	1.5755	1.1021	1.1085	2.2531
Interfering speech	1.5660	1.1768	1.3604	2.2672

Conclusion

- Probabilistic approach scheme improves the quality of reconstructed speech.
- Geometric harmonics can not describe the map very well.
- Laplacian pyramid method enable further noise reduction, but at the cost of distortion for the reconstructed speech.

Future research

- Geometric harmonics in a multi-scale manner.
- Find the relation between iteration number and noise level for Laplacian pyramids.
- Further processing needs to reduce distortions of reconstructed speech in geometric methods.