One microphone blind dereverberation based on quasi-periodicity of speech signals

Tomohiro Nakatani, Masato Miyoshi
and Keisuke Kinoshita

Speech Open Lab.
NTT Communication Science Labs.
NTT Corporation, Japan

Background & Purpose

- Reverberation degrades automatic speech recognition
  - Dereverberation is required as pre-processing
- Existing blind dereverberation methods such as ICA
  - Assumption: source signal is i.i.d., but speech is not i.i.d.

➡ Destroy formant structure and periodicity of speech

- Our method – single channel blind dereverberation
  - Assumption: speech signal is quasi-periodic
  - Estimates inverse filter based on this assumption

What is blind dereverberation ?

Only observed signal $x(n)$ is given, where $x(n) = h(n) * s(n)$

$h(n)$ room impulse response (unknown) $s(n)$ source signal (unknown speech)

Purpose: Estimate inverse filter $w(n)$ for $-N < n < N$

to obtain dereverberated signal $y(n) = w(n) * x(n)$

Desired condition:

$$w(n) * h(n) = d(n) = \begin{cases} c & \text{for } n = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $d(n)$ is a direct signal component of $h(n) = d(n) + r(n)$

$q(n) = w(n) * h(n)$: dereverberated impulse response

Features of quasi-periodic (QP) signals

- A QP-signal $s(n)$ has the following features:
  - $s(n)$ is approximately periodic in each local time region
  - The period gradually changes with time, and $s(n)$ has different periods in different time regions
- $x(n)$ becomes non-periodic if long reverberation is added

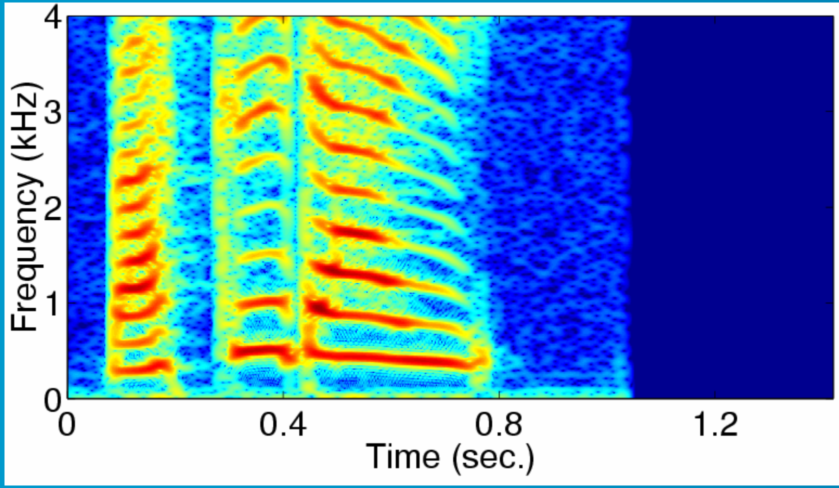
$$x(n)=\sum_m h(m)s(n-m)=h(0)\underline{s(n)}+\ldots+h(m)\underline{s(n-m)}+\ldots$$

Signals that have different periods are added.

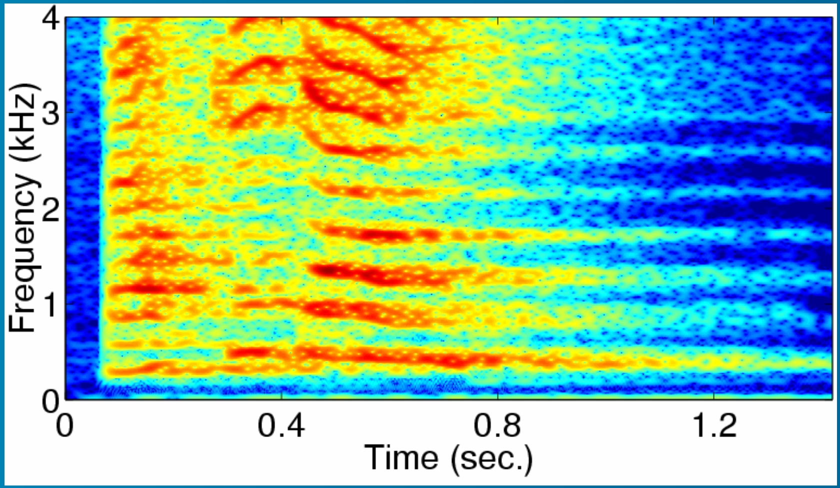
degrades the periodicity of $x(n)$.

Quasi-periodicity (QP) of speech signal

Clean speech



Reverberant speech



$$S(\omega) = S_h(\omega) + S_n(\omega)$$

$S_h(\omega)$ QP-signal (harmonic parts)

$S_n(\omega)$ non-periodic signal (deviation from harmonics)

$$X(\omega) = H(\omega)S(\omega)$$

$$= \underbrace{D(\omega)S(\omega)}_{\text{Direct signal}} + \underbrace{R(\omega)S(\omega)}_{\substack{\text{Reverberation} \\ \text{(should be removed)}}}$$

where $H(\omega) = D(\omega) + R(\omega)$

New dereverberation principle

Goal: estimate $w(n)$ that makes $y(n)$ a QP-signal

- Once such a filter is obtained, $q(m)$ must be zero in a long time region when assuming $q(0) \neq 0$.

$$y(n) = \sum_k w(m)x(n-m) = \sum_k q(m)s(n-m) = \underbrace{q(0)s(n)}_{\neq 0} + \ldots + \underbrace{q(m)s(n-m)}_{=0} + \ldots$$

- Therefore, long reverberation is eliminated by $w(n)$.

An inverse filter that eliminates the long reverberation of a quasi-periodic signal can be obtained by enhancing the quasi-periodicity of the observed signal

Two dereverberation methods

HERB: Harmonicity-based dEReverBeration methods

ATF-HERB

- Calculates an Average Transfer Function (ATF) that transforms reverberant signals into approximated direct QP-signals.

MMSE-HERB

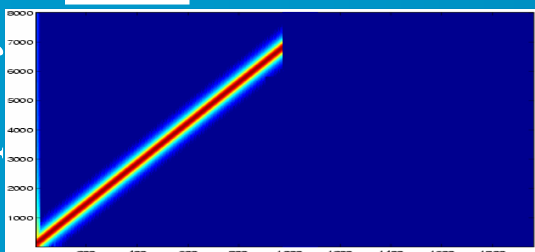
- Evaluates the QP of target signals in terms of a Minimum Mean Squared Error (MMSE) criterion, and minimizes the criterion.

ATF-HERB: basic idea

Source signal: QP-signal

$S(\omega)$

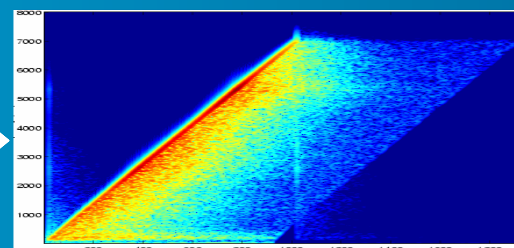
Frequency



Time

Reverberant observed signal

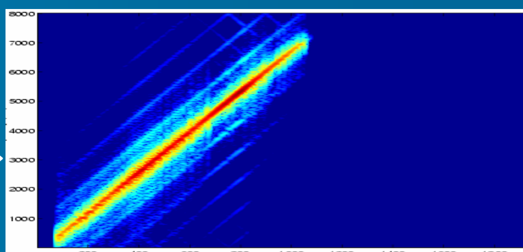
$X(\omega) = H(\omega)S(\omega)$



Room transfer function : H

Approx. direct QP-signal

$\hat{X}(\omega) \approx D(\omega)S(\omega)$

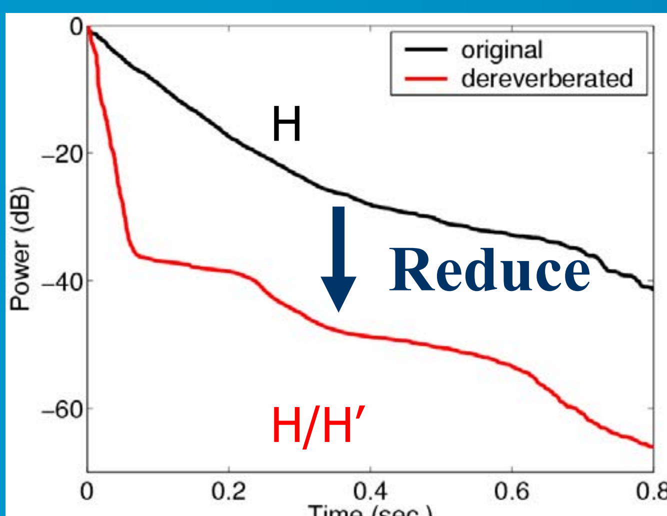


Extract a dominant sinusoid

Transfer function

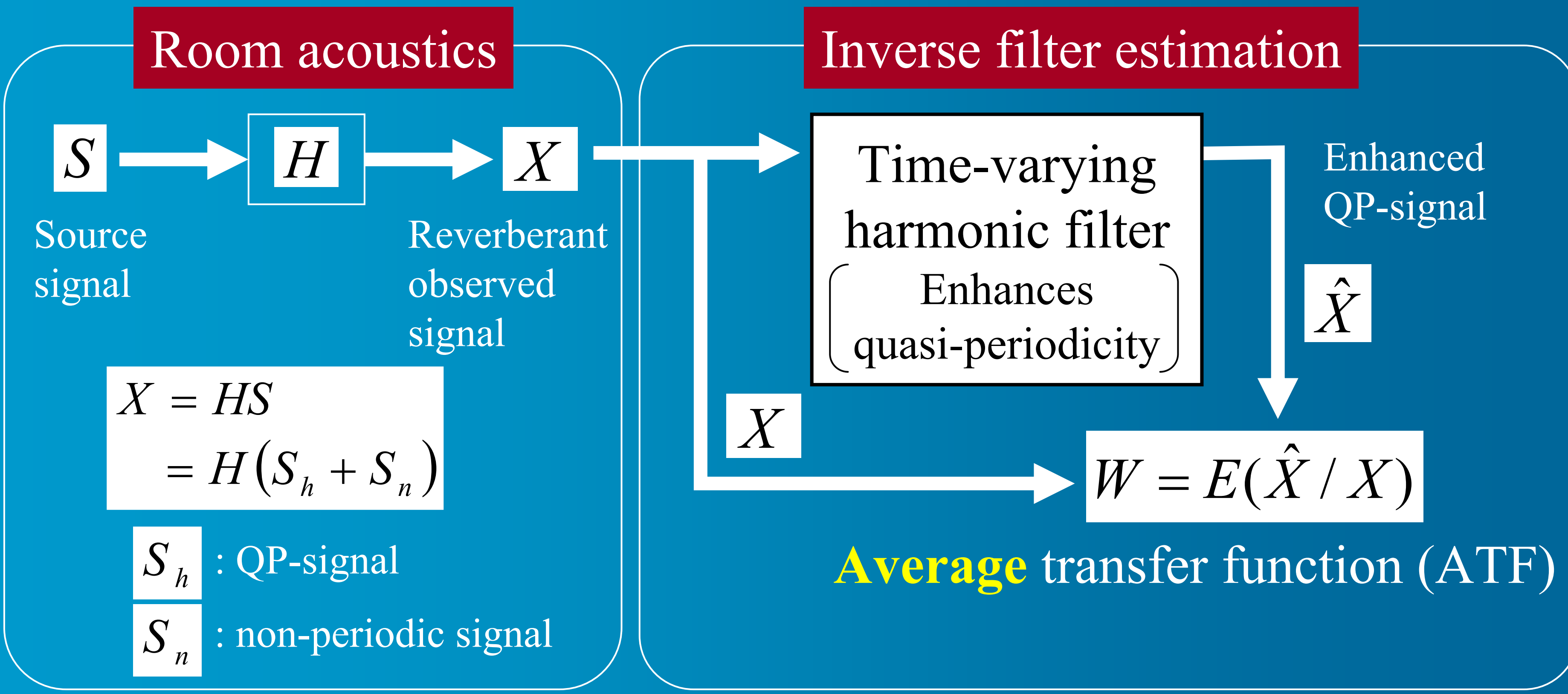
$$W = \hat{X}(\omega) / X(\omega)$$
$$\approx D(\omega) / H(\omega)$$

Utilize as inverse filter



Reverberation curve

ATF-HERB: definition of ATF



ATF is a time-invariant filter that is expected to enhance the quasi-periodicity of reverberant observed signals

Time-varying harmonic filter

1. Estimates fundamental freq. $f_0(n_0)$ at each frame n_0
2. Applies an adaptive comb filter that preserves the phase and amplitude of each harmonic component

Models of harmonic filtering

Reverberant speech signal

$$X = H(S_h + S_n)$$

$$H = \underbrace{D}_{\text{Direct signal component}} + \underbrace{R}_{\text{Reverberation component}}$$

$$= \underbrace{DS_h}_{\text{Direct QP-signal}} + \underbrace{(RS_h + HS_n)}_{\text{Non-periodic signal}}$$

Harmonic filtering

Approximated direct QP-signal

$$\hat{X} = DS_h + \underbrace{(\hat{R}S_h + \hat{N})}_{\text{Reduced non-periodic signal}}$$

ATF-HERB: property of ATF

- ATF is mathematically shown to be:

$$E\left(\hat{X}'(\omega)/X(\omega)\right)=\frac{\left(D(\omega)+\hat{R}(\omega)\right)}{H(\omega)}P\left(\left|S_h(\omega)\right|^2>\left|S_n(\omega)\right|^2\right)$$

  - $\left(D+\hat{R}\right)/H$: dereverberation operator
    - reduces reverberation by multiplying X by the operator

$$\left(\left(D+\hat{R}\right)/H\right)X=DS+\underline{\hat{R}S}$$

Reduced reverberation

    - This operator can dereverberate both periodic and non-periodic parts of signals because an inverse filter is independent of signal characteristics.
  - $P\left(\left|S_h\right|^2>\left|S_n\right|^2\right)$: probability of $\left|S_h\right|^2>\left|S_n\right|^2$ $\left(0.0<P(\cdot)<1.0\right)$
    - only affects filter gain without degrading dereverberation effects

ATF-HERB: derivation of property

Derivation:

$$E\left(\frac{\hat{X}}{X}\right)=E\left(\frac{DS_h+\left(\hat{R}S_h+\hat{N}\right)}{X}\right),$$

$$=E\left(\frac{D+\hat{R}}{H}\frac{S_h}{S_h+S_n}\right)+E\left(\frac{\hat{N}}{X}\right),$$

$$=\frac{D+\hat{R}}{H}E\left(\frac{1}{1+S_n/S_h}\right)+E\left(\frac{1}{1+\left(X-\hat{N}\right)/\hat{N}}\right).$$

Assume $\left|X-\hat{N}\right|>\left|\hat{N}\right|$, then the following is shown based on the lemma-1:

$$E\left(\frac{1}{1+S_n/S_h}\right)=P\left(\left|S_h\right|^2>\left|S_n\right|^2\right),$$

$$E\left(\frac{1}{1+\left(X-\hat{N}\right)/\hat{N}}\right)=0.$$

Then, the next equation is derived:

$$E\left(\frac{\hat{X}}{X}\right)=\frac{D+\hat{R}}{H}P\left(\left|S_h\right|^2-\left|S_n\right|^2\right)$$

Lemma-1:

Let $f(z)=\dfrac{1}{1+z}=\dfrac{1}{1+re^{j\theta}}$, where $z=re^{j\theta}$, and

assume θ is uniformly distributed,

θ and r are independent, and

$r\neq 1$.

Then, $E(f(z))=P(|z|<1)$.

Proof:

$$E(f(z))=P(|z|<1)E(f(z))_{|z|<1}+P(|z|>1)E(f(z))_{|z|>1},$$

$$E(f(z))_{|z|<1}=E\left(\frac{1}{1+re^{j\theta}}\right)_{r<1},$$

Taylor expansion

$$=1+\sum_{n=1}^{\infty}E\left((-r)^n e^{jn\theta}\right)_{r<1},$$

$$=1.$$

$$E(f(z))_{|z|>1}=E\left(\frac{1}{1+re^{j\theta}}\right)_{r>1},$$

$r'=1/r$

$$=E\left(\frac{r'e^{-j\theta}}{1+r'e^{-j\theta}}\right)_{r'<1},$$

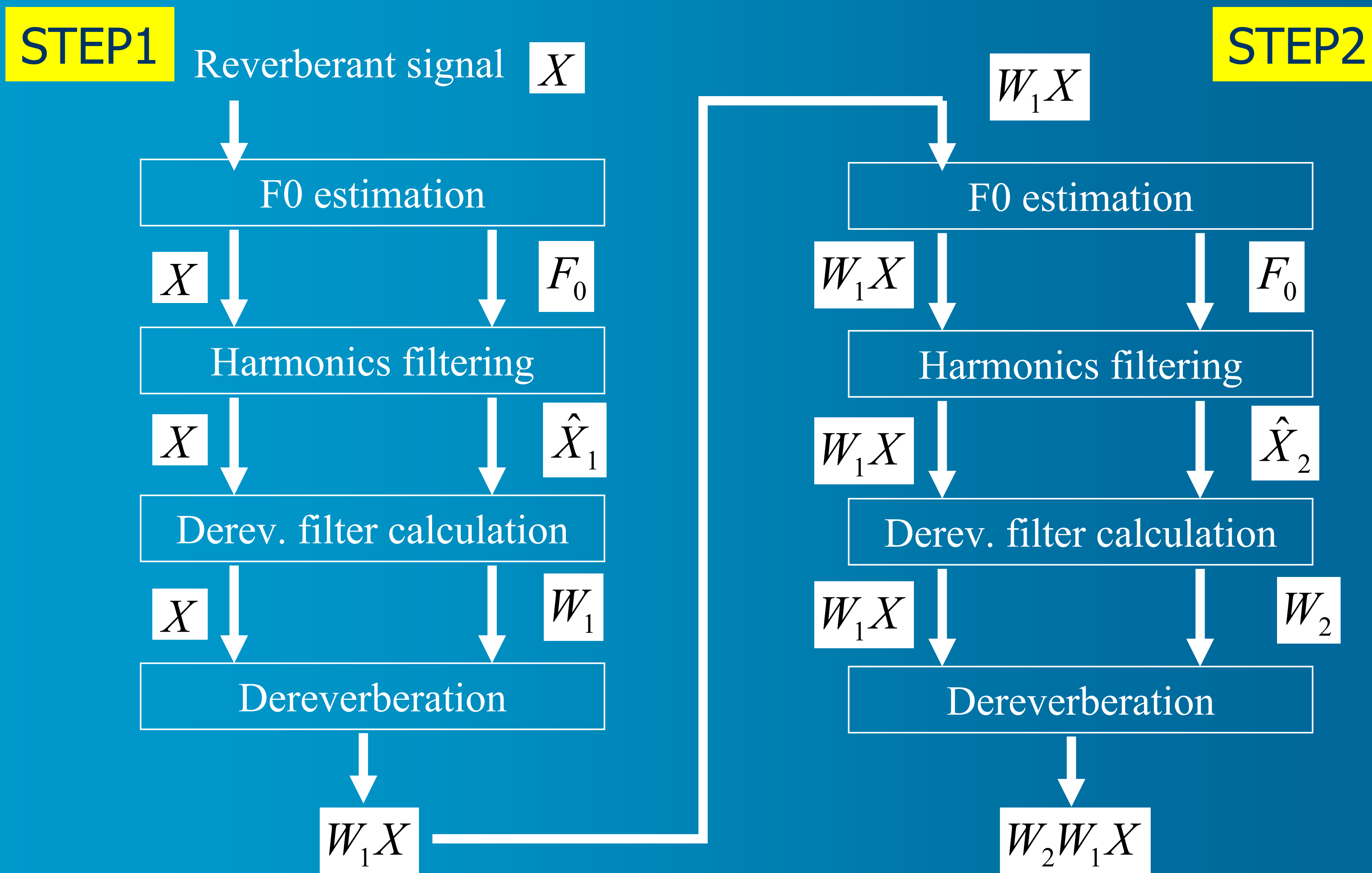
Taylor expansion

$$=-\sum_{n=1}^{\infty}E\left((-r')^n e^{-jn\theta}\right)_{r'<1},$$

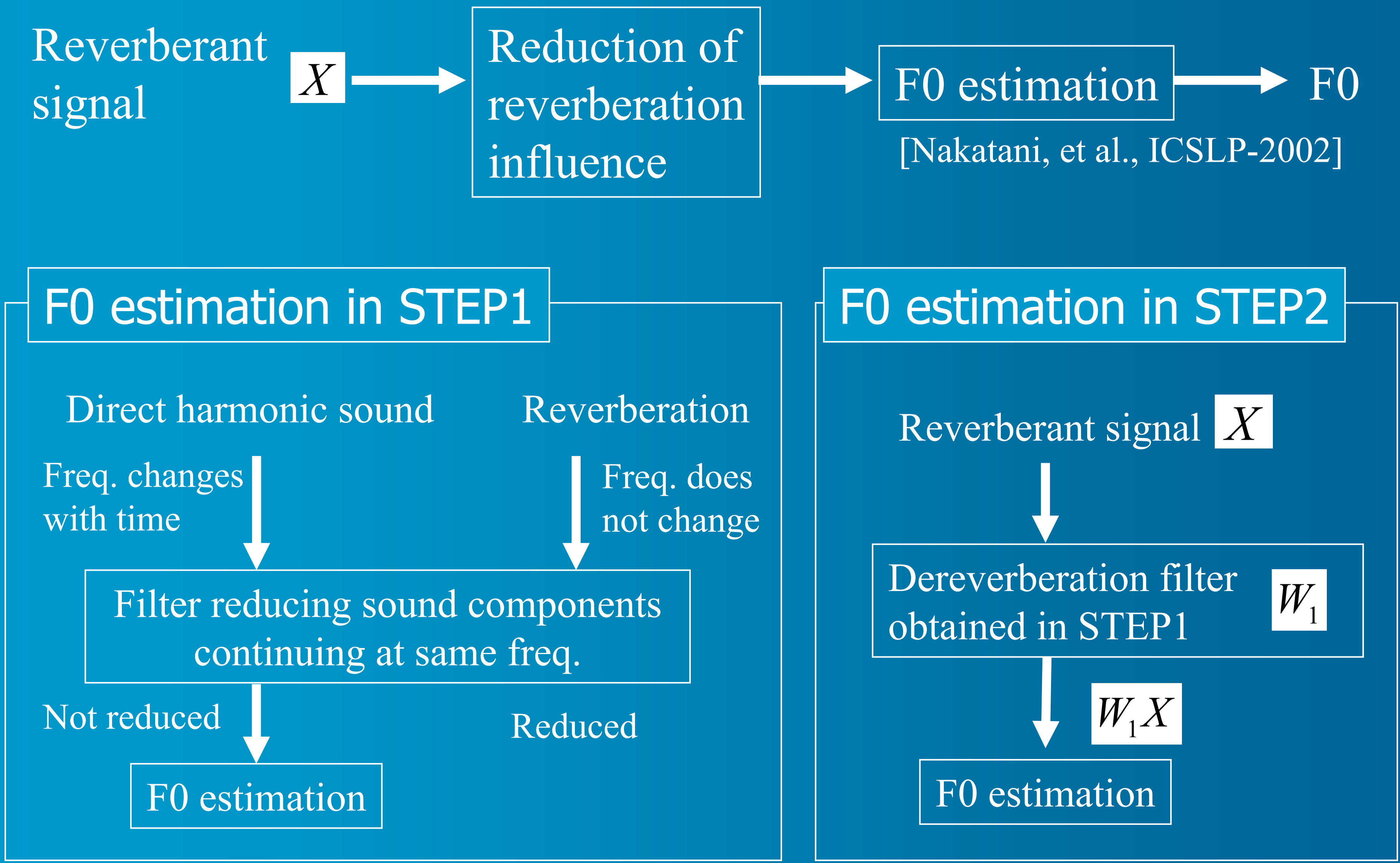
$$=0.$$

$$\therefore E(f(z))=P(|z|<1).$$

Processing flow



Robust F0 estimation with reverberation



MMSE-HERB: basic idea & problems

- MMSE criterion to evaluate QP of target signal

$$C(w) = \sum_n \left(y(n) - \hat{y}(n)\right)^2$$

where $\hat{y}(n)$ is harmonic filter output for $y(n)$

  – $C(w) \approx 0$ when $y(n)$ is a QP-signal

$\Rightarrow$ $y(n)$ is expected to be a QP-signal by minimizing $C(w)$

- Problems
  – $w(n)$ in short time region is not specifically determined because of features of QP-signal.
  – Self-evident solution $w(n) = 0$ for all n .
  – Computing cost for the optimization is high.

MMSE-HERB: simple solution

- Simplified MMSE criterion

$$C(W(\omega)) = \sum_n \left(Y(\omega) - \hat{X}(\omega)\right)^2$$

  - Desired signal is specifically given as $\hat{X}(\omega)$
  - $W(\omega) = 0$ is no longer optimum
  - The solution is given analytically:

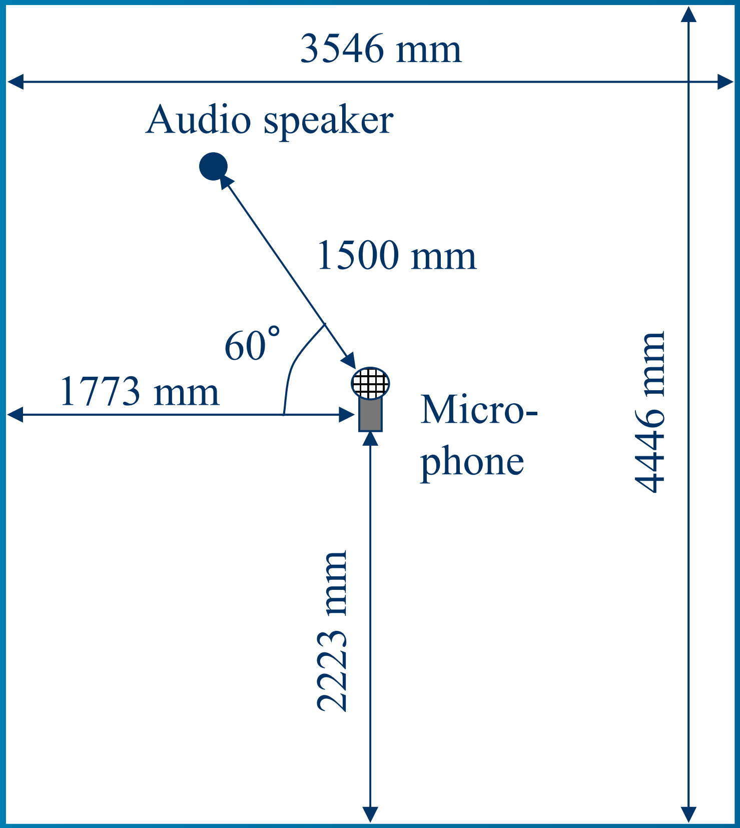
$$W(\omega) = \frac{E\left(\hat{X}(\omega) X^*(\omega)\right)}{E\left(X(\omega) X^*(\omega)\right)}$$

- The solution again approaches the dereverberation operator

$$W(\omega) \approx \mathrm{O}\left(\hat{R}(\omega), \omega\right) \frac{E\left(\left|S_h(\omega)\right|^2\right)}{E\left(\left|S_h(\omega)\right|^2\right) + E\left(\left|S_n(\omega)\right|^2\right)}$$

Experimental conditions

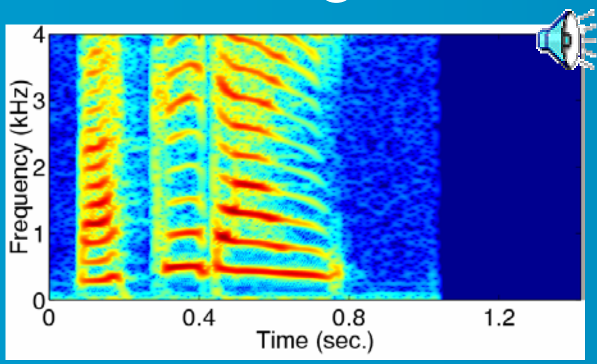
- Source signal $s(n)$
  - ATR word DB (12 kHz,16 bits)
  - Female: FKM（5240 words）
  - Male: MAU（5240 words）
- Room impulse response $h(n)$
  - Measured with reverberation time of 0.1, 0.2, 0.5, and 1.0 sec.
- Reverberant signal $h(n)*s(n)$
  - Synthesized by convolving source signal with impulse responses
- Dereverberation filter $w(n)$
  - Delayed inverse filter
  - 131,072 taps (10.9 sec DFT window)

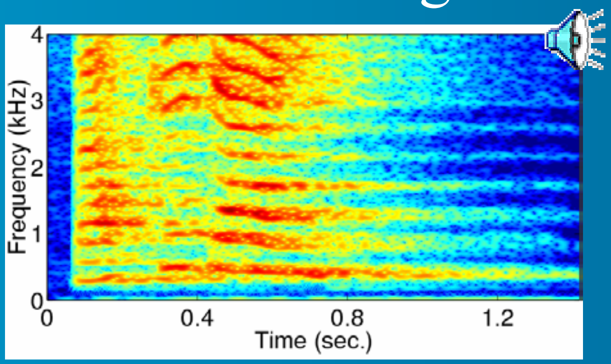


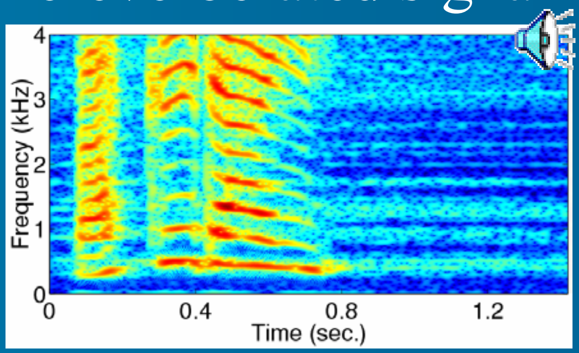
Impulse response measurement

Demonstration –ATF-HERB

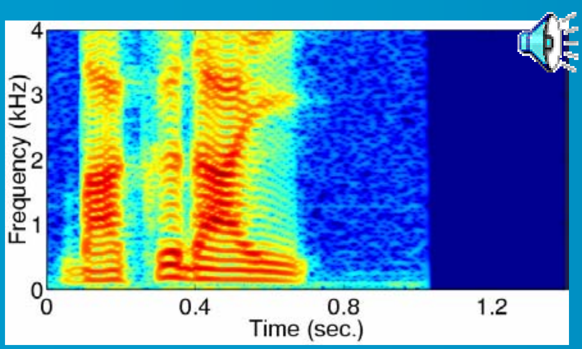
Source signal | Reverberant signal | Dereverberated signal

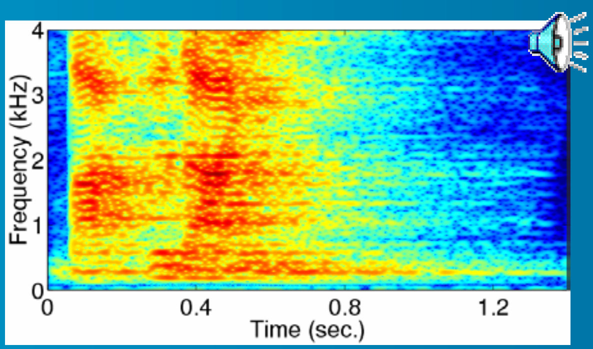


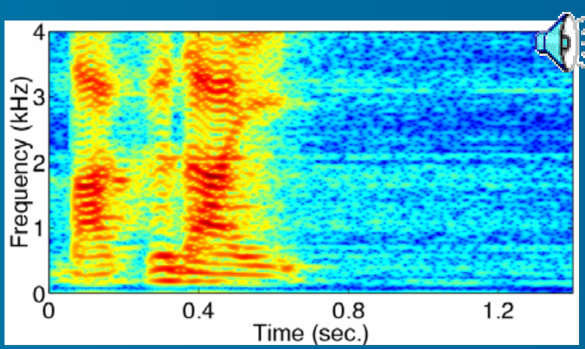




(1) Dereverberation of female voice (reverberation time: 1.0 sec)



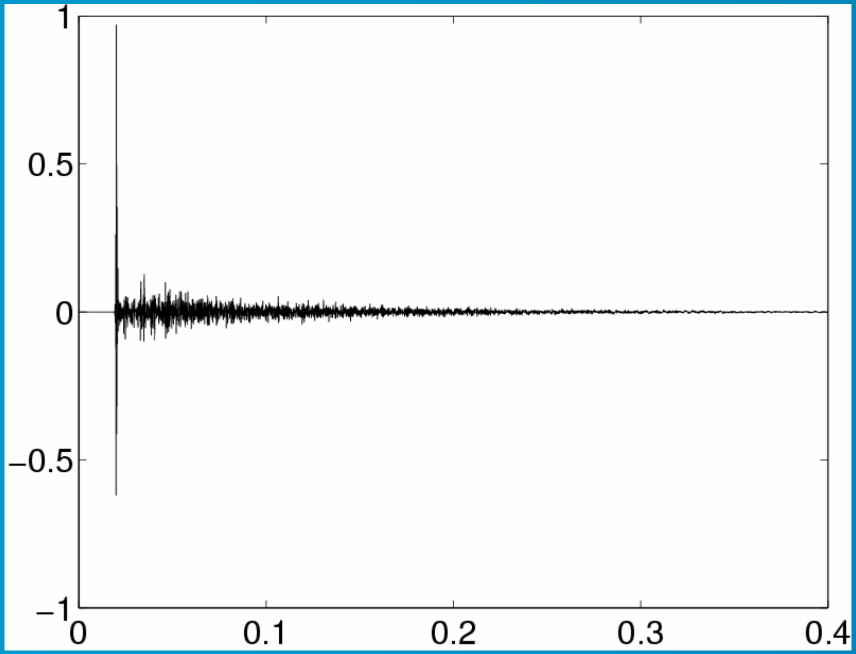




(2) Dereverberation of male voice (reverberation time: 1.0 sec)

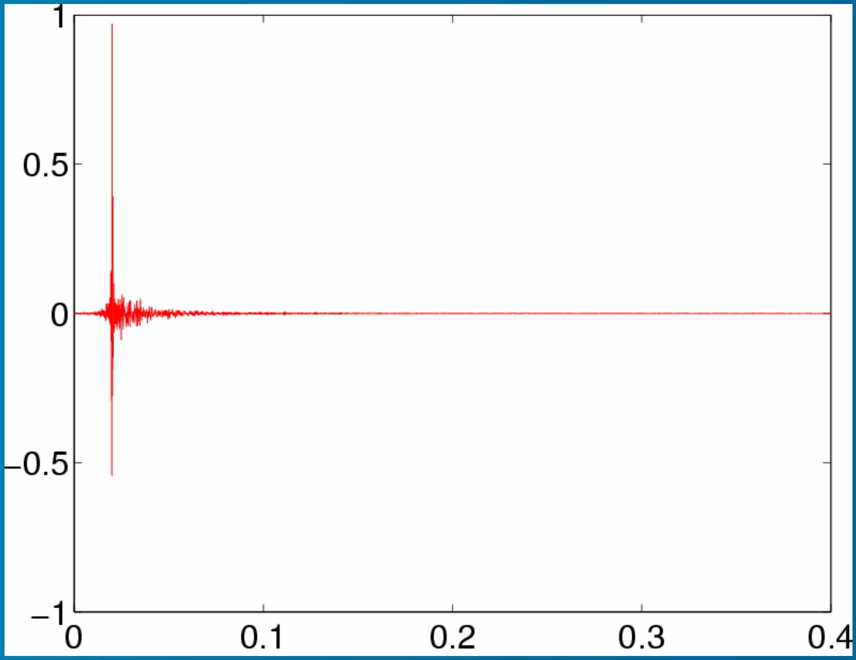
Dereverberated impulse response –ATF-HERB

Room impulse response $h(n)$
reverberation time : 1.0 sec.



Time (sec.)

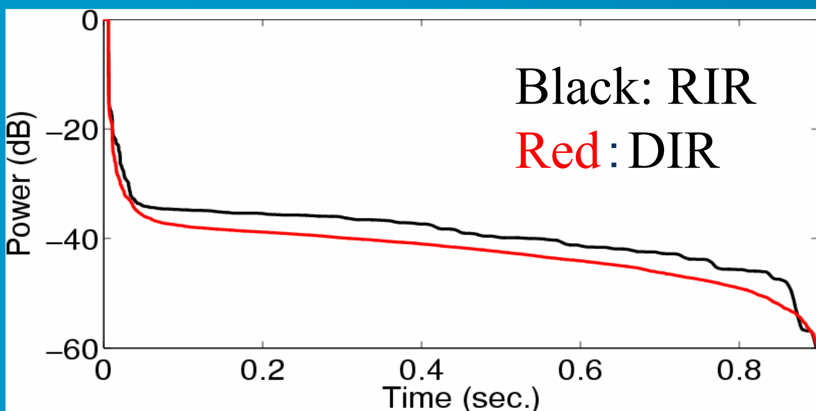
Dereverberated impulse response
$q(n) = w(n) * h(n)$
obtained with female utterances



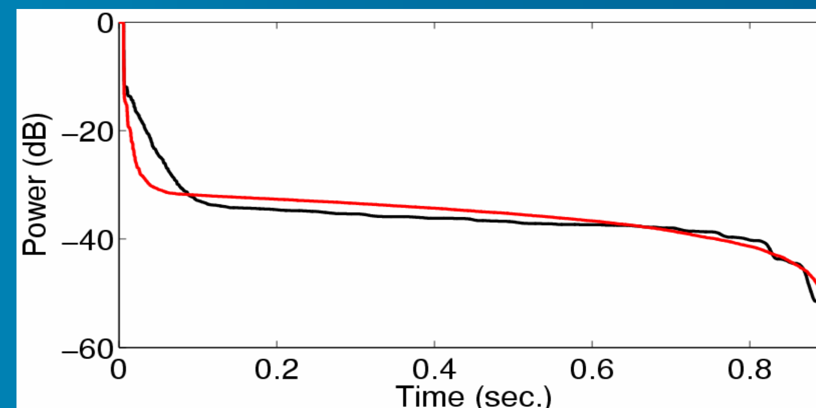
Time (sec.)

Reverberation curves (female) –ATF-HERB

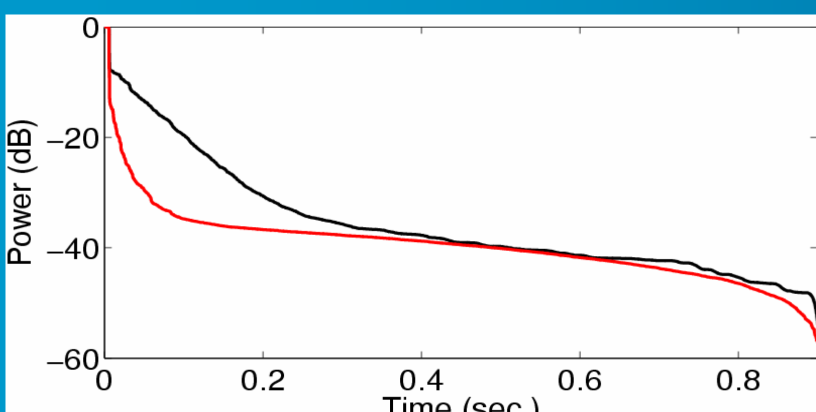
Energy of room impulse response (RIR)/dereverberated impulse response (DIR) decreasing with time



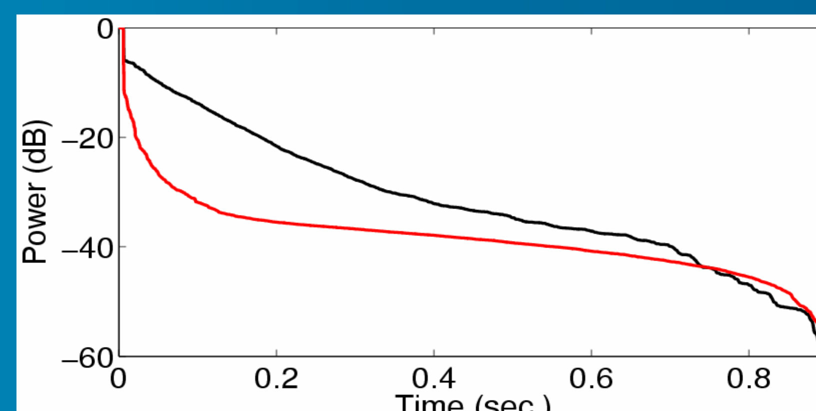
Reverberation time: 0.1 sec



Reverberation time: 0.2 sec



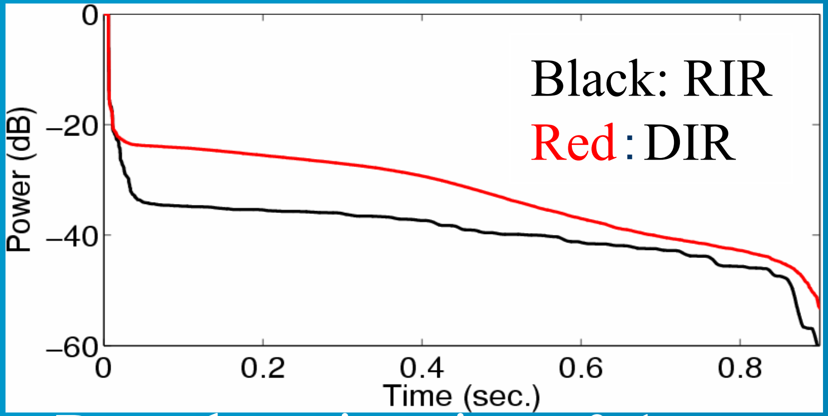
Reverberation time: 0.5 sec



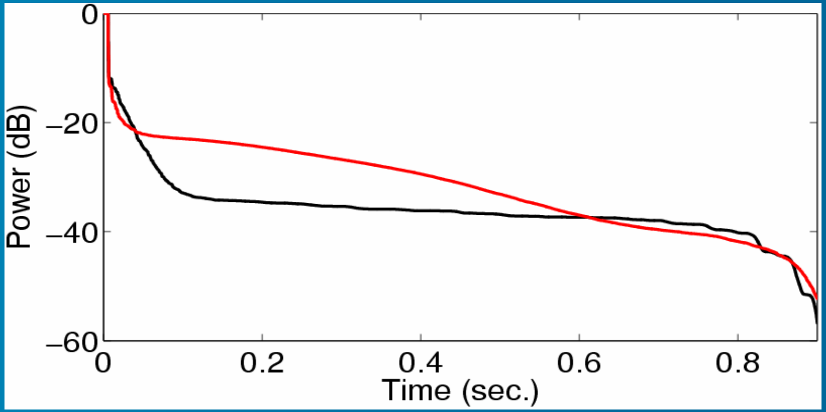
Reverberation time: 1.0 sec

Reverberation curves (male) –ATF-HERB

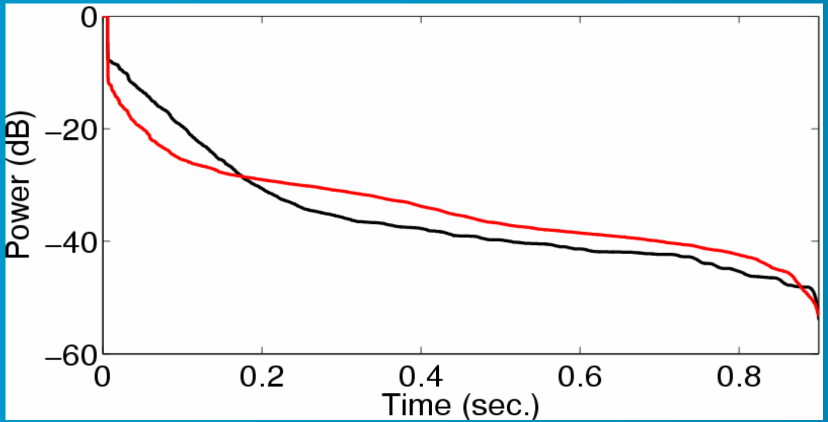
Energy of room impulse response (RIR)/dereverberated impulse response (DIR) decreasing with time



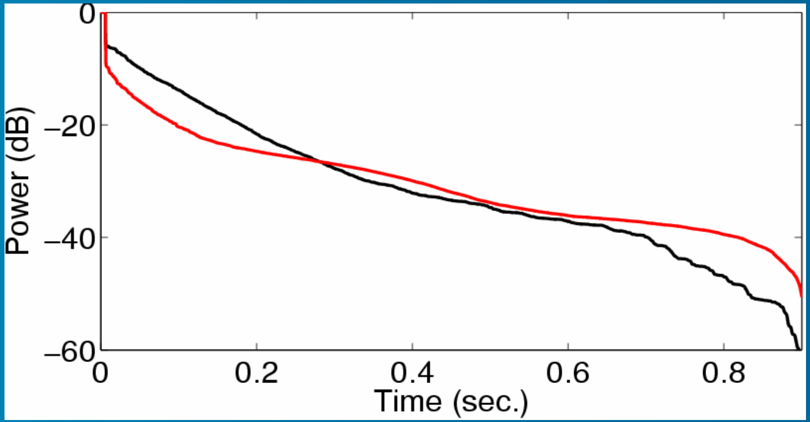
Reverberation time: 0.1 sec



Reverberation time: 0.2 sec



Reverberation time: 0.5 sec

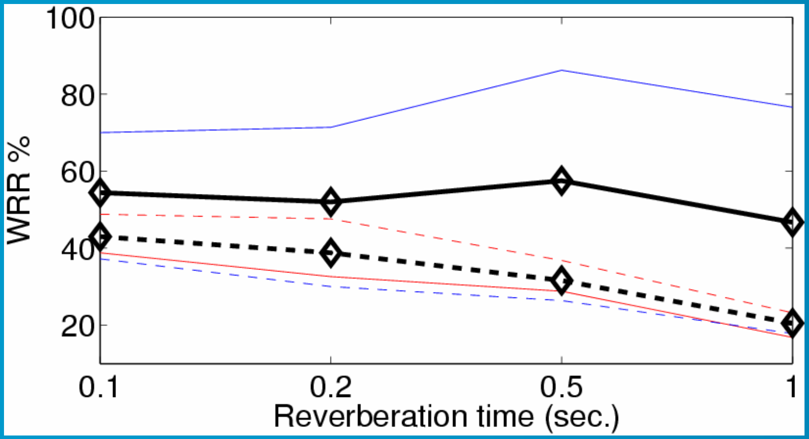


Reverberation time: 1.0 sec

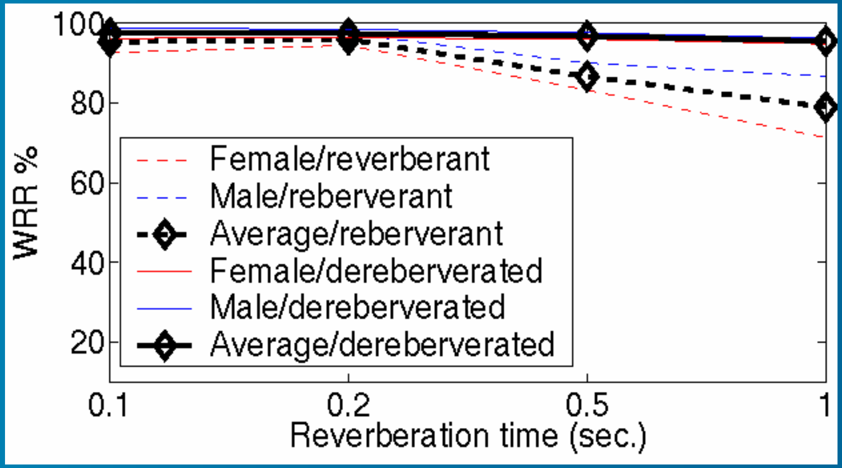
Word recognition rates (WRRs) –ATF-HERB

- Speaker dependent WRRs for reverberant/dereverberated signals
  - 4740 words for training, 500 words for testing, two speakers (MAU/FKM)
- With two types of acoustic monophone models:
  - Clean speech model: learned with original signal
  - Matched condition model: learned with reverberant/dereverberated signal

(1) WRRs with clean speech model



(2) WRRs with matched condition model



Analysis conditions: 12 order MFCCs, 12 order delta MFCCs, three state HMMs, five mixture Gaussian distributions, 25 msec frame length, 5 msec frame shift

Conclusion

- A new dereverberation principle based on quasi-periodicity of speech is presented
- Two dereverberation methods, ATF-HERB and MMSE-HERB, are implemented
- A dereverberation filter trained with 5240 words effectively reduces the reverberation
- Future work
  - Reduction of the data size required for training
  - Application to adaptive processing