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 $x(n) = h(n) \ast s(n)$

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

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

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

Desired condition:

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

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

$q(n) = w(n) \ast h(n)$: 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

\[ x(n) = \sum_{m} h(m)s(n-m) = h(0)s(n) + \ldots + h(m)s(n-m) + \ldots \]

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)

\[ X(\omega) = H(\omega)S(\omega) = D(\omega)S(\omega) + R(\omega)S(\omega) \]

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

Direct signal
Reverberation (should be removed)
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 $q(0) \neq 0$.

$$y(n) = \sum_k w(m)x(n-m) = \sum_k q(m)s(n-m) = q(0)s(n) + \ldots + q(m)s(n-m) + \ldots$$

- 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 \( S(\omega) \)
- **Reverberant observed signal:** \( X(\omega) = H(\omega)S(\omega) \)
- **Approx. direct QP-signal:** \( \hat{X}(\omega) \approx D(\omega)S(\omega) \)

**Reverberation curve**

- **Frequency**
- **Time**

**Room transfer function:**

\[
H(\omega) = \frac{1}{SHX(\omega)} \approx \frac{1}{\hat{H}(\omega)}
\]

**Transfer function**

\[
W = \frac{\hat{X}(\omega)}{X(\omega)} \approx \frac{D(\omega)}{H(\omega)}
\]

**Utilize as inverse filter**

- **Reduce**
- \( H/H' \)
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

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

$$H = D + R$$

$$= DS_h + (RS_h + HS_n)$$

Direct QP-signal

Non-periodic signal

Harmonic filtering

Approximated direct QP-signal

$$\hat{X} = DS_h + (\hat{R}S_h + \hat{N})$$

Reduced non-periodic signal
ATF-HERB: property of ATF

- ATF is mathematically shown to be:
  \[
  E\left( \hat{X}'(\omega) / X(\omega) \right) = \frac{D(\omega) + \hat{R}(\omega)}{H(\omega)} P\left( |S_h(\omega)|^2 > |S_n(\omega)|^2 \right)
  \]

  \[
  \left( D + \hat{R} \right) / H \quad : \text{dereverberation operator}
  \]

  - reduces reverberation by multiplying \( X \) by the operator

  \[
  \left( D + \hat{R} \right) / H \cdot X = DS + \hat{R}S
  \]

- This operator can **dereverberate both periodic and non-periodic parts** of signals because an inverse filter is independent of signal characteristics.

- \( P\left( |S_h|^2 > |S_n|^2 \right) \): probability of \(|S_h|^2 > |S_n|^2\) \( (0.0 < P(\cdot) < 1.0) \)

  - only affects filter gain **without degrading dereverberation effects**
Derivation:

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

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

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

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

\[
E\left(\frac{1}{1+S_n/S_h}\right) = P\{|S_n| > |S_\hat{n}|\}
\]

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

Then, the next equation is derived:

\[
E\left(\frac{\hat{X}}{X}\right) = \frac{D + \hat{R}}{H} P\{|S_n|^2 - |S_\hat{n}|^2\}
\]

Lemma-1:

Let \(f(z) = \frac{1}{1+z} = \frac{1}{1+re^{i\theta}}\), where \(z = re^{i\theta}\), and assume \(\theta\) is uniformly distributed, \(\theta\) and \(r\) are independent, and \(r \neq 1\).

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

Proof:

\[
E(f(z)) = P|z| < 1)E(f(z)|_{|z| < 1} + P|z| > 1|E(f(z)|_{|z| > 1}.
\]

\[
E(f(z)|_{|z| < 1}) = E\left(\frac{1}{1+re^{i\theta}}\right)_{r<1} \quad \text{Taylor expansion}
\]

\[
= 1 + \sum_{n=1}^\infty E((-r)ne^{in\theta})_{r<1}
\]

\[
= 1.
\]

\[
E(f(z)|_{|z| > 1}) = E\left(\frac{1}{1+re^{i\theta}}\right)_{r>1} \quad r' = 1/r
\]

\[
= E\left(\frac{re^{-i\theta}}{1+re^{-i\theta}}\right)_{r'>1} \quad \text{Taylor expansion}
\]

\[
= \sum_{n=1}^\infty E((-r')ne^{-in\theta})_{r'>1}
\]

\[
= 0.
\]

\[
E(f(z)) = P(|z| < 1).
\]
Processing flow

**STEP1**
Reverberant signal $X$

- F0 estimation $F_0$
- Harmonics filtering $\hat{X}_1$
- Derev. filter calculation $W_1$
- Dereverberation $W_1X$

**STEP2**

- Dereverberation $W_2W_1X$
- Dereverberation $\hat{X}_2$
- Derev. filter calculation $W_2$
- Harmonics filtering $W_1X$
- F0 estimation $F_0$
Robust F0 estimation with reverberation

Reverberant signal $X$ → Reduction of reverberation influence → F0 estimation → F0

[F0 estimation in STEP1]

Direct harmonic sound
Freq. changes with time
Filter reducing sound components continuing at same freq.
Not reduced
F0 estimation

Reverberation
Freq. does not change
Reduced

[F0 estimation in STEP2]

Reverberant signal $X$
Dereverberation filter $W_1$
obtained in STEP1
$W_1X$
F0 estimation

[Nakatani, et al., ICSLP-2002]
**MMSE-HERB: basic idea & problems**

**• MMSE criterion to evaluate QP of target signal**

\[ C(w) = \sum_{n} (y(n) - \hat{y}(n))^2 \]

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

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

\[ y(n) \text{ is expected to be a QP-signal by minimizing } C(w) \]

**• 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

\[ C(W(\omega)) = \sum_n (Y(\omega) - \hat{X}(\omega))^2 \]

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

\[ W(\omega) = \frac{E(\hat{X}(\omega)X^*(\omega))}{E(X(\omega)X^*(\omega))} \]

- The solution again approaches the dereverberation operator

\[ W(\omega) \approx O(\hat{R}(\omega), \omega) \frac{E(|S_n(\omega)|^2)}{E(|S_n(\omega)|^2) + E(|S_n(\omega)|^2)} \]
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)
Demonstration – ATF-HERB

(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.

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

Black: RIR
Red: DIR
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