MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Suppression of Acoustic Noise in Speech Using Spectral Substraction

Haryo Akbarianto Wibowo

About This Paper…

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-27, NO. 2, APRIL 1979

Suppression of Acoustic Noise in Speech Using Spectral Subtraction

STEVEN F. BOLL, MEMBER, IEEE

About This Paper…

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-27, NO. 2, APRIL 1979

Suppression of Acoustic Noise in Speech Using Spectral Subtraction

STEVEN F. BOLL, MEMBER, IEEE

This paper was published when I was not in this world yet!

Introduction

- Background Noise in speech can **degrade the performance** of audio tasks!
- (IN THE PAST) Noise-cancelling microphones is useless to noise reduction more than 1 kHz

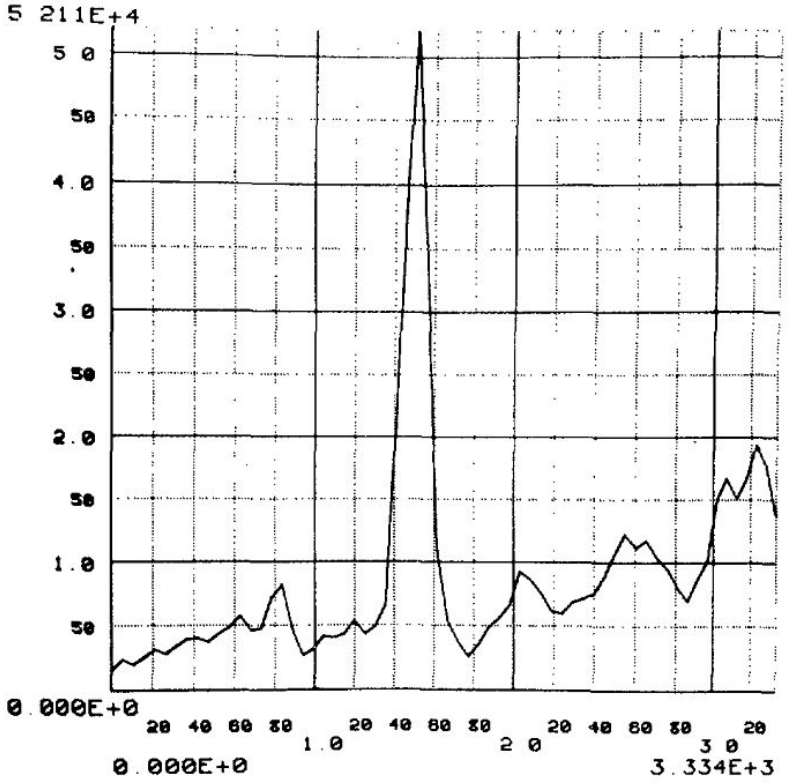


Fig. 5. Average noise magnitude of helicopter noise.

Contributions

Develop a noise suppression technique and efficient algorithm.

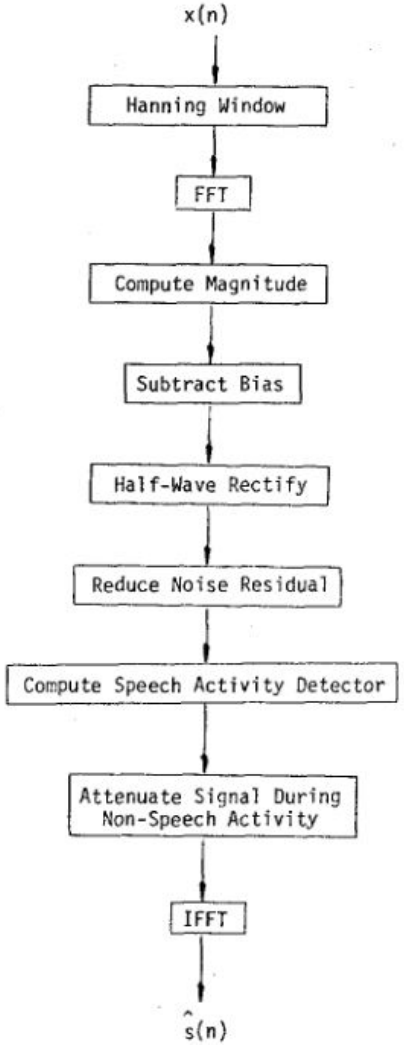
Main Idea

Subtracting the noise magnitude spectrum from the noise speech spectrum

Previous approaches: obtain the noise estimate from a second microphone

This paper: Approximated using average noise magnitude from non-speech activity

Substractive Noise Suppression


x(n)
Hanning Window
FFT
Compute Magnitude
Subtract Bias
Half-Wave Rectify
Reduce Noise Residual
Compute Speech Activity Detector
Attenuate Signal During Non-Speech Activity
IFFT
ŝ(n)

Assumption

- *The background noise environment remains locally stationary to the degree that its spectral magnitude expected value just prior to speech activity equals its expected value during speech activity.*
- If the environment changes to a new stationary state, it has time (300 ms) to estimate the new noise.
- It assumes that significant noise reduction is possible by removing the noise from the magnitude spectrum only

Additive Noise Model (Discrete Time Fourier Transform [DTFT])

$$x(k) = s(k) + n(k).$$

Taking the Fourier transform gives

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega})$$

where

$$x(k) \longleftrightarrow X(e^{j\omega})$$

$$X(e^{j\omega}) = \sum_{k=0}^{L-1} x(k)e^{-j\omega k}$$

$$x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega k}\, d\omega.$$

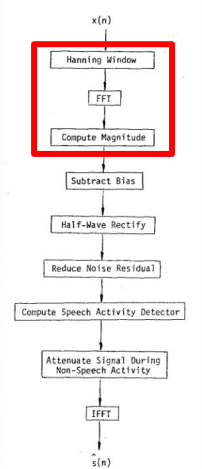
phase

n(k) = windowed noise signal

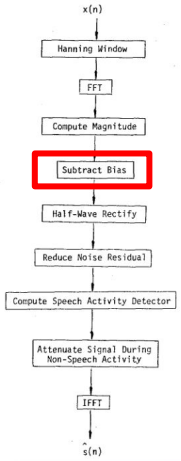
s(k) = windowed speech signal

x(k) = windowed speech with noise signal

N, S, X = Fourier transformed signal.



Spectral Subtraction Estimator



n(k) = windowed noise signal

s(k) = windowed speech signal

x(k) = windowed speech with noise signal

N, S, X = Fourier transformed signal.

The spectral subtraction filter $H(e^{j\omega})$ is calculated by replacing the noise spectrum $N(e^{j\omega})$ with spectra which can be readily measured. The magnitude $|N(e^{j\omega})|$ of $N(e^{j\omega})$ is replaced by its average value $\mu(e^{j\omega})$ taken during nonspeech activity, and the phase $\theta_N(e^{j\omega})$ of $N(e^{j\omega})$ is replaced by the phase $\theta_x(e^{j\omega})$ of $X(e^{j\omega})$. These substitutions result in the spectral subtraction estimator $\hat{S}(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = [|X(e^{j\omega})| - \mu(e^{j\omega})]\, e^{j\theta_x(e^{j\omega})}$$

Equivalent!

or

$$\hat{S}(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$$

with

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|}$$

$$\mu(e^{j\omega}) = E\{|N(e^{j\omega})|\}.$$

Spectral Error

$$\epsilon(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega}) - \mu(e^{j\omega})\, e^{j\theta_x}.$$

n(k) = windowed noise signal

s(k) = windowed speech signal

x(k) = windowed speech with noise signal

N, S, X = Fourier transformed signal.

$$\hat{S}(e^{j\omega}) = [|X(e^{j\omega})| - \mu(e^{j\omega})]\, e^{j\theta_x(e^{j\omega})}$$

We have subtract our noise, now what?

$$\epsilon(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega}) - \mu(e^{j\omega})\, e^{j\theta_x}.$$

To reduce the spectral error further:

1. Magnitude Averaging
2. Half-wave Rectification
3. Residual Noise Reduction
4. Additional Signal Attenuation during Non Speech Activity

Magnitude Averaging

n(k) = windowed noise signal

s(k) = windowed speech signal

x(k) = windowed speech with noise signal

N, S, X = Fourier transformed signal.

Miu(e^jw) = Average non speech noise signal

Previously, Spectral Error

$$\epsilon(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega}) - \mu(e^{j\omega})\, e^{j\theta_x}$$

We can also replace magnitude of X with its local average

$$\overline{|X(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |X_i(e^{j\omega})|$$

$X_i(e^{j\omega}) = i$th time-windowed transform of $x(k)$

gives

$$S_A(e^{j\omega}) = [\overline{|X(e^{j\omega})|} - \mu(e^{j\omega})]\, e^{j\theta_x(e^{j\omega})}.$$

The rationale behind averaging is that the spectral error becomes approximately

$$\epsilon(e^{j\omega}) = S_A(e^{j\omega}) - S(e^{j\omega}) \cong \overline{|N|} - \mu$$

Thus, the sample mean of $|N(e^{j\omega})|$ will converge to $\mu(e^{j\omega})$ as a longer average is taken.

Magnitude Averaging (Problem)

- Speech is non-stationary and therefore only limited time averaging is allowed.
- The results show that averaging over more than three half-overlapped windows with a total time duration of 38.4 ms will decrease intelligibility.
- Different approach: Half-Wave Rectification

Half-Wave Rectification

For each frequency ω where the noisy signal spectrum magnitude $|X(e^{j\omega})|$ is less than the average noise spectrum magnitude $\mu(e^{j\omega})$, the output is set to zero. This modification can be simply implemented by half-wave rectifying $H(e^{j\omega})$. The estimator then becomes

$$\hat{S}(e^{j\omega}) = H_R(e^{j\omega})X(e^{j\omega})$$

where

$$H_R(e^{j\omega}) = \frac{H(e^{j\omega}) + |H(e^{j\omega})|}{2}.$$

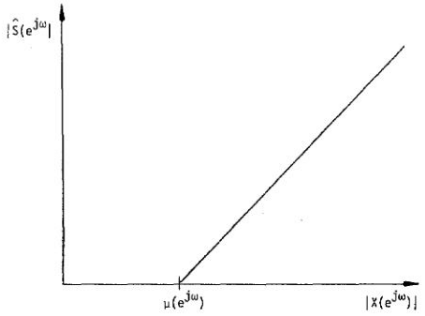


Fig. 1. Input-output relation between $X(e^{j\omega})$ and $\hat{S}(e^{j\omega})$.

n(k) = windowed noise signal

s(k) = windowed speech signal

x(k) = windowed speech with noise signal

N, S, X = Fourier transformed signal.

Miu(e^jw) = Average non speech noise signal

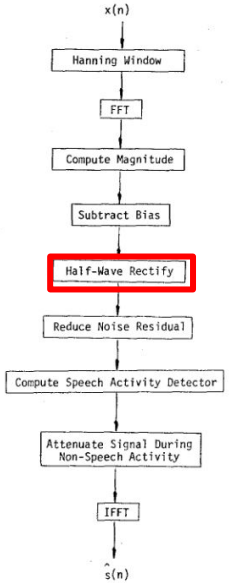
or

$$\hat{S}(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$$

with

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|}$$

$$\mu(e^{j\omega}) = E\{|N(e^{j\omega})|\}.$$



Half-Wave Rectification (Advantages)

- Noise floor is reduced by miu(e^jw)
- Low variance coherent noise tones are eliminated

Half-Wave Rectification (Disadvantages)

- **If speech + noise is lower than miu, its gone.**

Residual Noise Reduction

- Speech + Noise still exists > miu
- The noise that is also still there is called **noise residual**, will also form.

$$N_R = N - \mu e^{j\theta n},$$

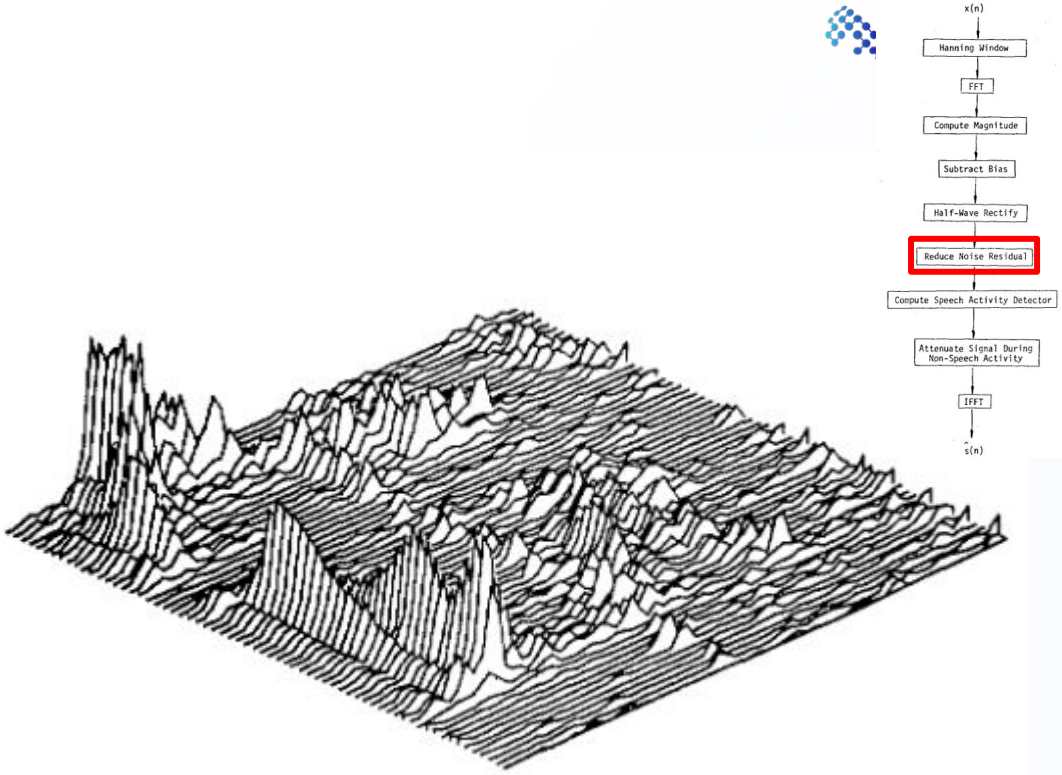


Fig. 7. Short-time spectrum using bias removal and half-wave rectification.

Residual Noise Reduction

- The residual can be reduced by taking advantage of its frame-to-frame randomness.
- Mitigated by replacing its current value with its minimum value chosen from the adjacent frames.

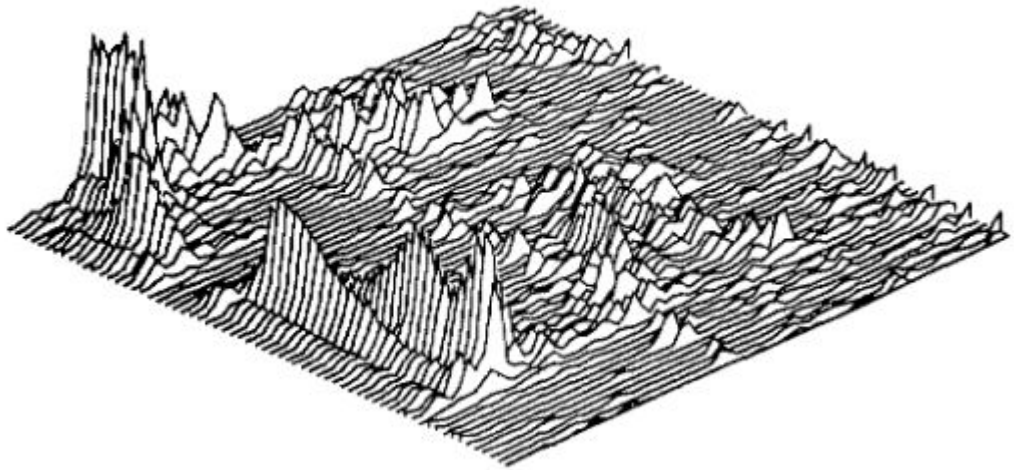


Fig. 7. Short-time spectrum using bias removal and half-wave rectification.

Residual Noise Reduction

$\hat{S}(e^{j\omega})$

- If the amplitude of S(tilde) lies below max noise residual, then it fluctuates, the spectrum at that frequency is due to noise. Suppress it by taking the minimum
- If S(tilde) lies below the maximum, but constant, it has low energy **speech** at that frequency.
- If S(tilde) is higher than the maximum residual, leave it.d

$$|\hat{S}_i(e^{j\omega})| = |\hat{S}_i(e^{j\omega})|, \quad \text{for} \quad |\hat{S}_i(e^{j\omega})| \geq \max |N_R(e^{j\omega})|$$

$$|\hat{S}_i(e^{j\omega})| = \min \{|\hat{S}_j(e^{j\omega})| \, j = i-1, i, i+1\},$$
$$\text{for} \quad |\hat{S}_i(e^{j\omega})| < \max |N_R(e^{j\omega})|$$

where

$$\hat{S}_i(e^{j\omega}) = H_R(e^{j\omega}) X_i(e^{j\omega})$$

and

$\max |N_R(e^{j\omega})|$ = maximum value of noise residual measured during nonspeech activity.

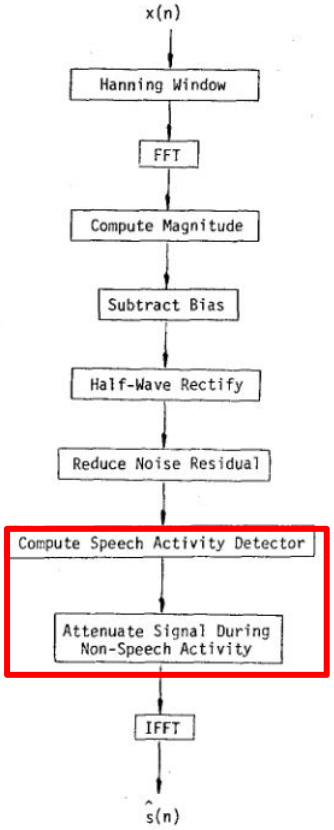
Amount of noise with this schema is equivalent to that obtained by averaging over three frames.

Additional Signal Attenuation During Nonspeech Activity

- If Speech is absent $\hat{S}(e^{j\omega})$ is noise.

$$T = 20 \log_{10} \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\hat{S}(e^{j\omega})}{\mu(e^{j\omega})} \right| d\omega \right].$$

- This paper empirically state that if T < -12 DB, it is noise.
- What to do?
  - Leave it as is
  - Set it to zero
  - Attenuate it?



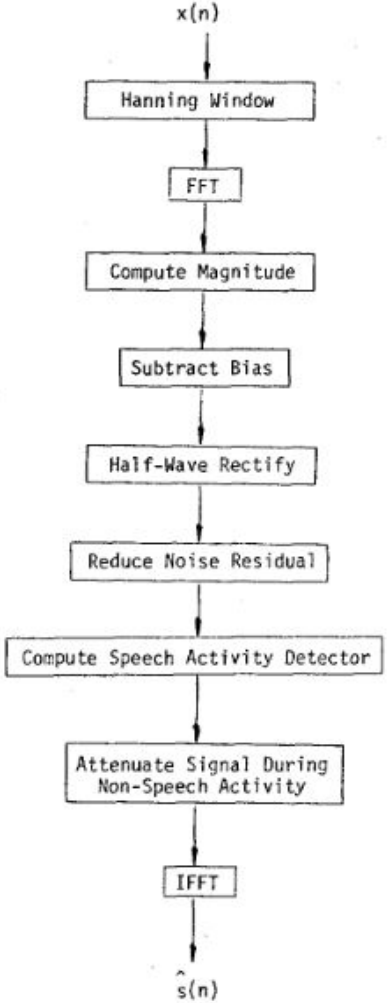
Additional Signal Attenuation During Nonspeech Activity

- Turns out having non-speech signal improves the result!
- Again empirically, optimum value of -30dB is set.

$$\hat{S}(e^{j\omega}) = \begin{cases} \hat{S}(e^{j\omega}) & T \geq -12 \text{ dB} \\ cX(e^{j\omega}) & T \leq -12 \text{ dB} \end{cases}$$

where $20 \log_{10} c = -30$ dB.

Inverse Fast Fourier Transform to Reconstruct it



Experiment Results

Evaluation

- Diagnostic Rhyme Test (DRT) from Dynastat Inc
  - Consists of 192 words recorded in a helicopter environment

Results are presented

- Short-time amplitude spectra of helicopter speech
- DRT intelligibility and quality score on **LPC vocoder speech**
- Short-Time Spectra Using Residual Noise Reduction and Non Speech Signal Attenuation

Short Time Spectra of Helicopter Speech

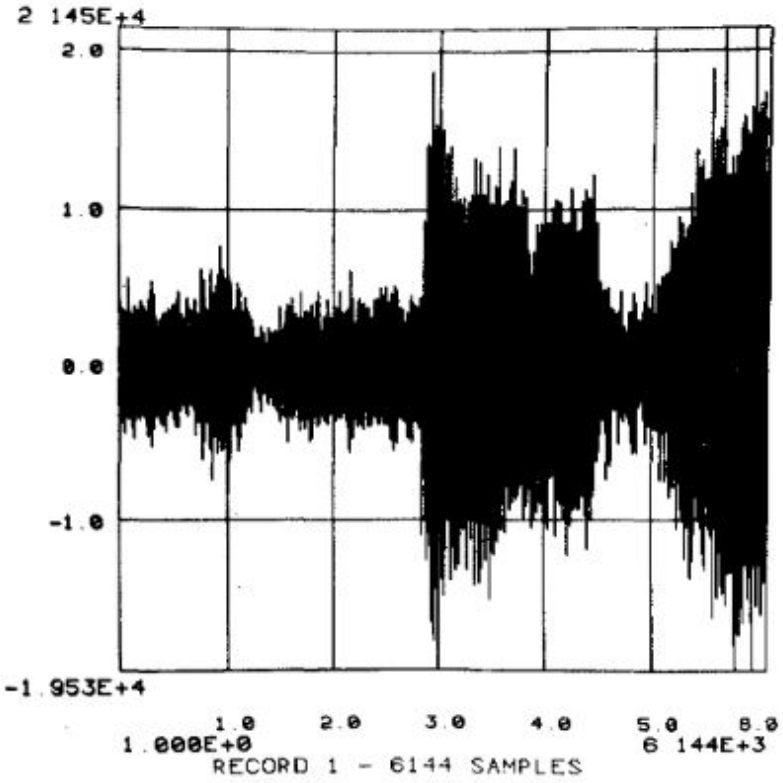


Fig. 4. Time waveform of helicopter speech. "Save your".

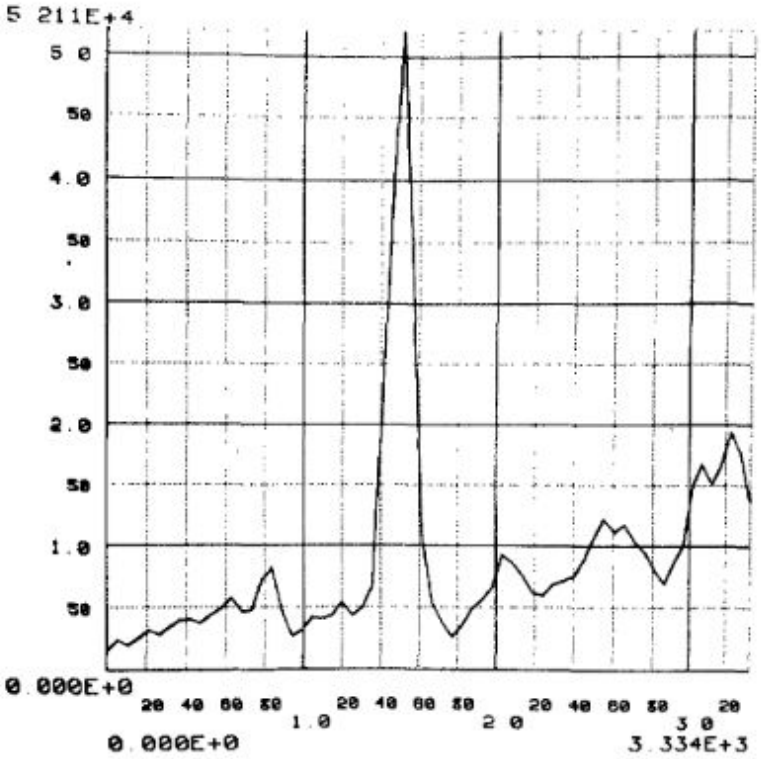


Fig. 5. Average noise magnitude of helicopter noise.

Short Time Spectra of Helicopter Speech

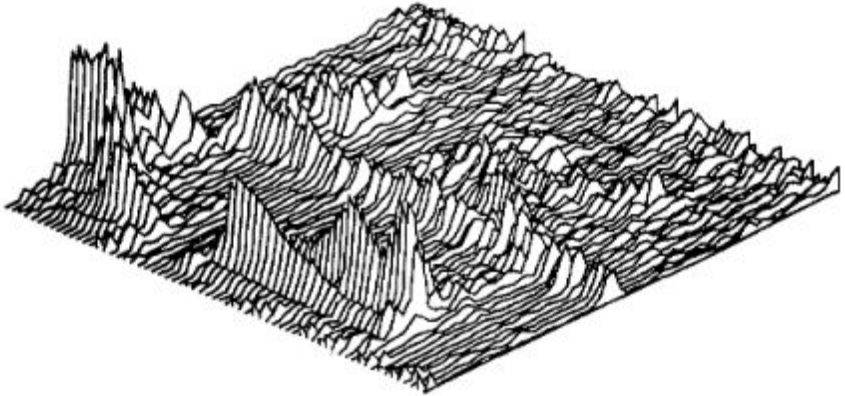


Fig. 6. Short-time spectrum of helicopter speech.

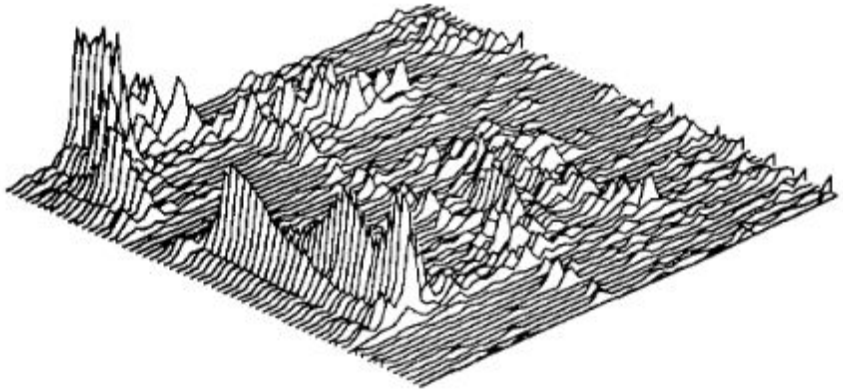


Fig. 7. Short-time spectrum using bias removal and half-wave rectification.

Short Time Spectra of Helicopter Speech

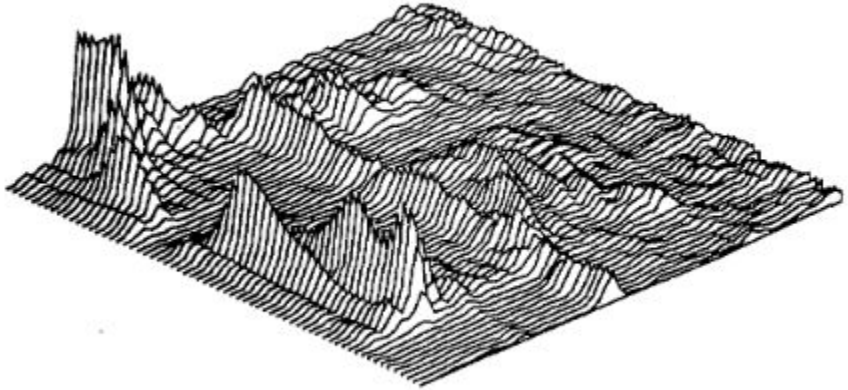


Fig. 8. Short-time spectrum of helicopter speech using three frame averaging.

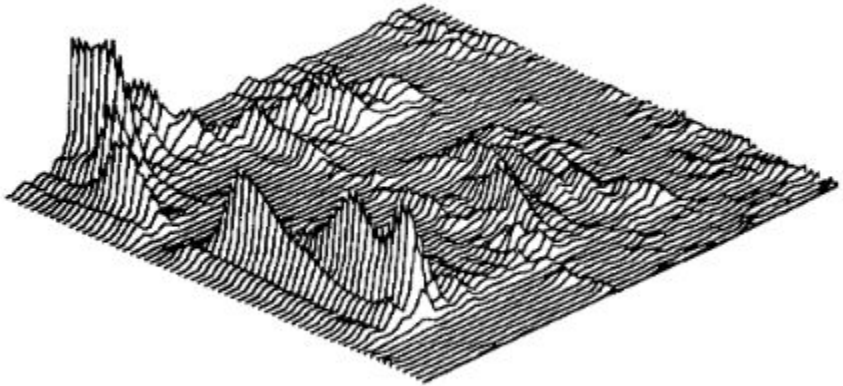


Fig. 9. Short-time spectrum using bias removal and half-wave rectification after three frame averaging.

Intelligibility and Quality Results using the DRT

NON LPC Vocoded

TABLE I
DIAGNOSTIC RHYME TEST SCORES

	Original	$\hat{S}$ (No Average)	$\hat{S}$ (Three Average)
Voicing	95	92	91
Nasality	82	78	77
Sustention	92	87	86
Sibilation	75	83	84
Graveness	68	70	66
Compactness	88	87	88
Total	84	83	82

TABLE II
QUALITY RATINGS

	Original	$\hat{S}$ (No Average)	$\hat{S}$ (Three Averages)
Naturalness of Signal	63	60	61
Inconspicuousness of Background	36	38	42
Intelligibility	30	32	33
Pleasantness	20	31	25
Overall Acceptability	27	33	29
Composite Acceptability	26	32	29

It does not **decrease intelligibility**, but does **increase quality**, like **pleasantness and inconspicuousness of background**!

Intelligibility and Quality Results using the DRT

LPC Vocoded

TABLE III
DIAGNOSTIC RHYME TEST SCORES

	LPC on Original	LPC on $\hat{S}$ without averaging	LPC on $\hat{S}$ with averaging
Voicing	84	90	86
Nasality	56	63	52
Sustention	49	52	56
Sibilation	61	70	88
Graveness	67	62	59
Compactness	83	83	93
Total	66	70	72

Improve the intelligibility

TABLE IV
QUALITY RATINGS

	LPC on Original	LPC on $\hat{S}$ without averaging	LPC on $\hat{S}$ with averaging
Naturalness of Signal	53	49	58
Inconspicuousness of Background	34	36	39
Intelligibility	28	30	28
Pleasantness	15	28	20
Overall Acceptability	24	28	26
Composite Acceptability	23	29	25

Short-Time Spectra Using Residual Noise Reduction and Nonspeech Signal Attenuation

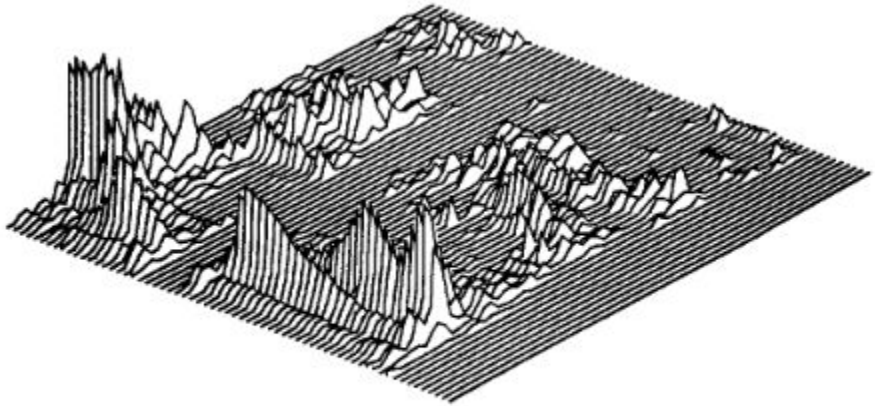


Fig. 10. Short-time spectrum using bias removal, half-wave rectification, residual noise reduction, and nonspeech signal attenuation (helicopter speech).

My take

- Positive, efficient, and elegant algorithm at that time
- Simple and easy to understand paper.
- Many assumptions and arbitrary threshold are given.
- This only works well on noise that has uniform distribution or close to it.
- Lack of Benchmark (i.e.: other model and type of noise)

Thank you!