

マスキング曲線を用いた音声強調とその応用

中西 崇 中島 栄俊 荻木 禎史 宇佐川 毅
(熊本大学)

1. はじめに

現在難聴者支援や音声認識などの多くの技術分野において音声強調システムに対する要求は極めて強い。これに対しさまざまな手法が提案されてきたが、問題点は残っている。音声強調システムには大別すると1入力システムと多入力システムがある。本報告では難聴者の補聴器などによる要求があること、また入力を増やすことにより処理が複雑になることから1入力システムに着目し音声強調を試みる。

1入力音声強調システムの手法として、雑音が定常であり音声信号と無相関であるという仮定に基づいてスペクトルの減算を行うスペクトルサブトラクション法^[1]があり、そのなかで最もシンプルな手法としてパワースペクトルサブトラクション法が知られている。しかしパワースペクトルサブトラクション法による強調信号にはスペクトルの引きすぎや雑音の引き残しによりミュージカルノイズが発生するという問題^[2]などがあり、音声認識などにおいて十分な効果が得られていない。

本報告では聴覚のマスキング現象を用いた音声強調の手法を提案する^{[3][4]}。マスキングとは、ある音に対する検知限が他の音の存在によって上昇する現象である。これは聴覚の周波数選択機能によって信号とマスカーとを分離できないときも生じる現象である。この現象を利用し、音声によるマスキング曲線より低い範囲にノイズを減衰させることでスペクトルの引きすぎにより発生するミュージカルノイズを抑える

ことが期待できる。本報告ではまずパワースペクトルサブトラクション法について、次に提案手法のアルゴリズムについて述べる。その後両手法を比較して評価を行う。

2. パワースペクトルサブトラクション法 (PSS)

PSS では雑音の統計的性質を用いて目的信号の強調を行う。観測信号を $y(t)$ とすると、 $y(t)$ は目的信号 $x(t)$ と雑音 $n(t)$ の和、すなわち

$$y(t) = x(t) + n(t) \quad (1)$$

と表せる。(1) 式の両辺をフーリエ変換すると

$$Y(\omega) = X(\omega) + N(\omega) \quad (2)$$

となる。更に (2) 式からパワースペクトルを求めると

$$\begin{aligned} |Y(\omega)|^2 &= |X(\omega) + N(\omega)|^2 \\ &= |X(\omega)|^2 + |N(\omega)|^2 \\ &\quad + X(\omega)N(\omega)^* + X(\omega)^*N(\omega) \\ &= |X(\omega)|^2 + |N(\omega)|^2 \end{aligned} \quad (3)$$

となる。ただし $*$ は複素共役とする。さらに雑音 $n(t)$ が定常で、かつ信号と無相関とする。今雑音のみが観測可能でその周波数特性が推定できるとする。この時強調信号 $\hat{x}(t)$ のパワースペクトル $|\hat{X}(\omega)|^2$ は

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha |N(\omega)|^2, & \text{if } |Y(\omega)| \geq |N(\omega)| \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

* Signal enhancement based on masking curve and the Application

By Takashi NAKANISHI, Hidetoshi NAKASHIMA, Fumiyoshi CHISAKI and Tuiyoshi USAGAWA (Kumamoto University)

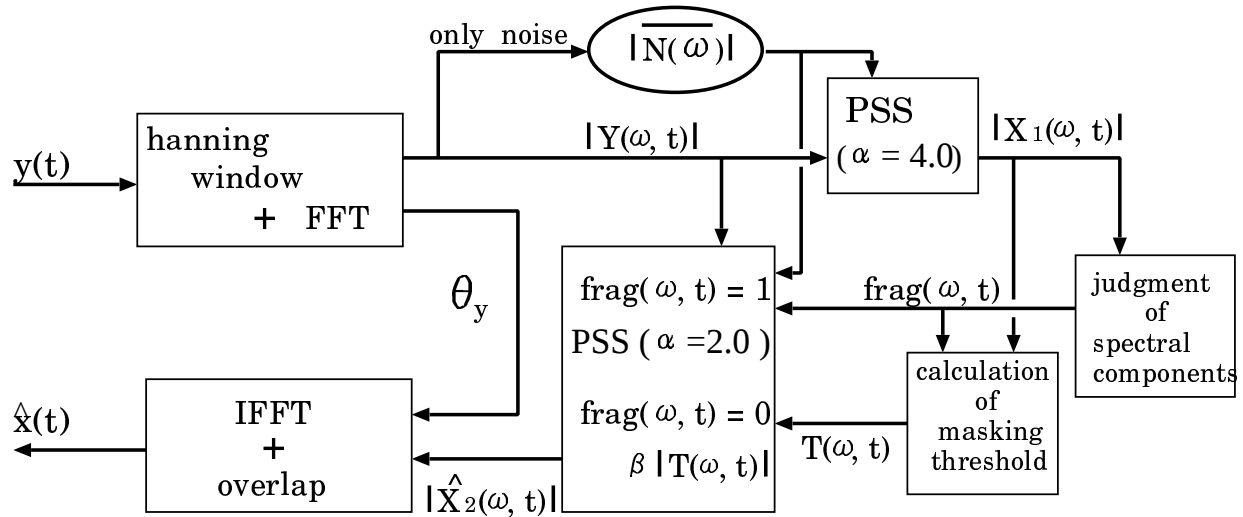


図 1 : ブロック図

とすることができる. ここで $|\overline{N(\omega)}|^2$ は雑音の平均パワースペクトルである. 位相に関しては位相歪が与える影響が聴覚上で少ないことから $y(t)$ の位相 $\theta_Y(\omega)$ を用いるとすれば推定音声信号 $\hat{x}(t)$ は

$$\hat{x}(t) = F^{-1}[\|\hat{X}(\omega)\|e^{j\theta_Y(\omega)}] \quad (5)$$

で与えられる.(ただし F^{-1} は逆フーリエ変換)

3. 提案アルゴリズム

PSS では雑音スペクトルの平均値を求め観測信号のスペクトルから減算することで強調信号のスペクトルを算出する. しかし一般に雑音スペクトルは分散を持つことから, 雑音スペクトルの引き残し又音声スペクトルの引きすぎが生じる恐れがある. この影響でミュージカルノイズや音声の歪みが発生することになり, 結果として音声明瞭度が低下する.

本報告で提案するアルゴリズムは PSS を用いて大幅にスペクトルを減衰させた信号に対してマスキング曲線の計算と音声スペクトルの検出を行い, 雑音部分のスペクトルをマスキング曲線以下にすることで減衰後のスペクトルの不連続性を回避することで, スペクトルサブトラクション法で問題となるミュージカルノイズを軽減できることが期待される.

図 1 に本手法のブロック図を示す. 処理手順は,

- (1) 観測信号を FFT 分析する
- (2) ノイズスペクトルを保持する

- (3) PSS を用い観測信号のスペクトルを減衰させることで signal, non-signal の検定を行い flag を立て, signal 成分のマスキング曲線を求める

- (4) signal 成分は PSS を, non-signal 成分はマスキング曲線を用いて減衰を行う

- (5) 4. で求めたスペクトルを IFFT し, 時間波形を求める

である. この手順をフレーム毎に行う.

3.1 マスキング曲線の計算

PSS の α の値を大きくし観測信号のスペクトルを大幅に減衰させることで得られた $|\hat{X}_1(\omega, t)|$ よりスペクトルを signal 成分と non-signal 成分に別け, マスキング曲線を求める. signal の定義は MPEG1 の心理音響モデルをベースにしている. ここでは MPEG1 の心理音響モデルでは周波数領域で隣接したスペクトルより大きく, さらに同一臨界帯域内において周辺のスペクトルより 7dB 大きいスペクトルをピッチを持つスペクトルと定義しているのに従いそのスペクトルを signal とする. さらに signal でないスペクトルを non-signal であると定義する.

signal と判定されたら $flag(\omega, t) = 1$ とし, non-signal と判定されたら $flag(\omega, t) = 0$ とする. $flag(\omega, t) = 1$ であるスペクトルをマスキング曲線を求める. フレーム毎に全ての signal と判定されたスペクトルによるマスキング曲線の線形和

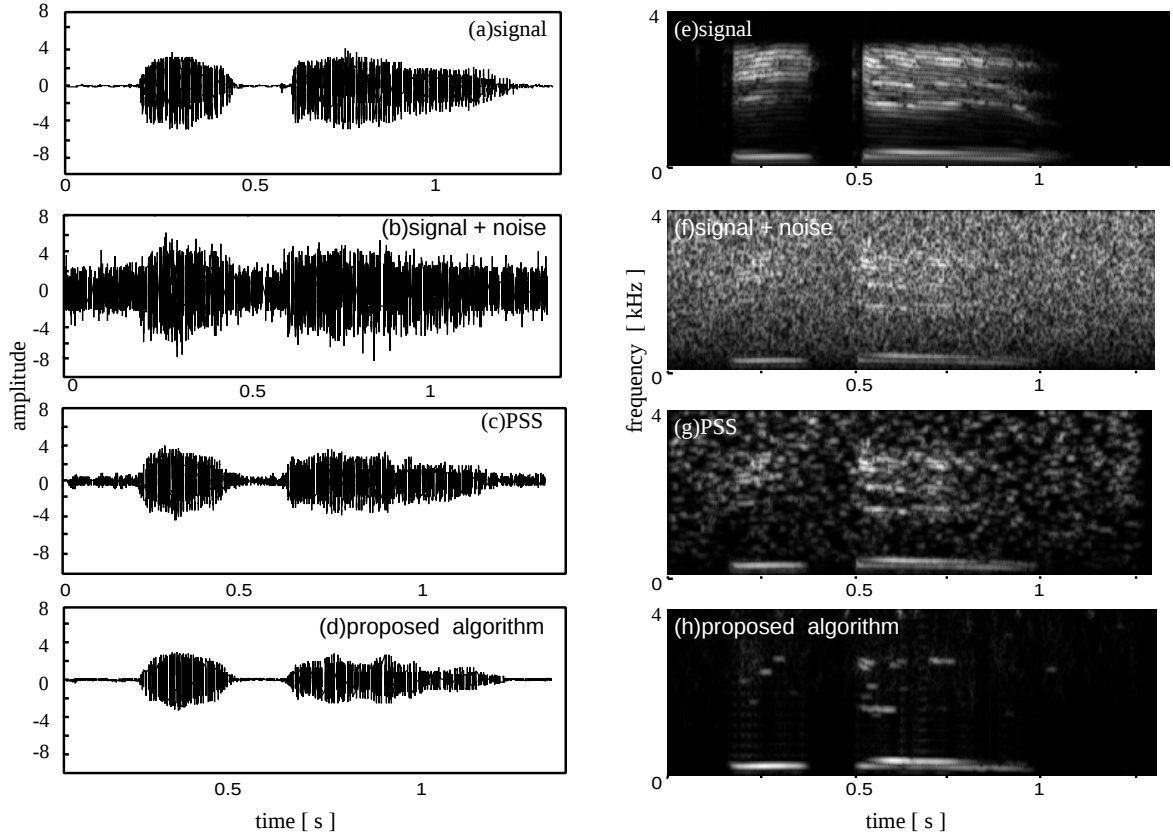


図2：波形とスペクトログラム：(a)(e) 原信号, (b)(f) 原信号+雑音 (ホワイトノイズ, $SNR = 0dB$), (c)(g) スペクトルサブトラクション法による強調信号, (d)(h) 提案手法による強調信号

をとることでそのフレームのマスキング曲線 $T(\omega, t)$ を得ることができる。

3.2 音声強調

3.1 において分類されたスペクトルに対して以下のような減衰処理を行う。

$$|\hat{X}_2(\omega, t)| = \begin{cases} \sqrt{|Y(\omega)|^2 - \alpha|N(\omega)|^2}, & \text{if } flag(\omega, t) = 1 \\ \beta T(\omega, t), & \text{if } flag(\omega, t) = 0 \end{cases} \quad (6)$$

ここで β は減衰係数であり

$$\beta = 10^{\frac{\gamma}{20}} \quad (7)$$

とした。ここで γ は減衰レベルであり、スペクトルを何 dB 減衰させるか決定するパラメータである。

上記の過程において強調されたスペクトル $|\hat{X}_2(\omega, t)|$ は観測信号の位相 $\theta_y(\omega)$ を用い逆フーリエ変換することで強調信号を作り出すことができる。すなわち

$$\hat{x}(t) = IFFT[|\hat{X}(\omega, t)|e^{j\theta_y(\omega)}] \quad (8)$$

を得る。

4. 評価

提案手法の評価を行うために、波形とスペクトログラムによる比較と、音声認識実験を行った。ここでは比較アルゴリズムとして PSS を用いた。音声認識実験ではノンリニアスペクトルサブトラクション法 (以下 NSS)^[5] も用いた。

4.1 波形とスペクトログラム

図2(a) ~ (d) に各々(a) 原信号, (b) 観測信号, (c) PSS を用いた強調信号, (d) 提案手法を用いた強調信号を示す。観測信号として女声を用い白色雑音を $SNR = 0dB$ で付加した。(c), (d) の比較より無音声区間において提案手法の場合は PSS 法を用いた場合に比べ雑音が減少しているのが分かる。有音声区間においては提案手法の場合 PSS 法に比べわずかに歪んでいるがほぼ同じ波形である。

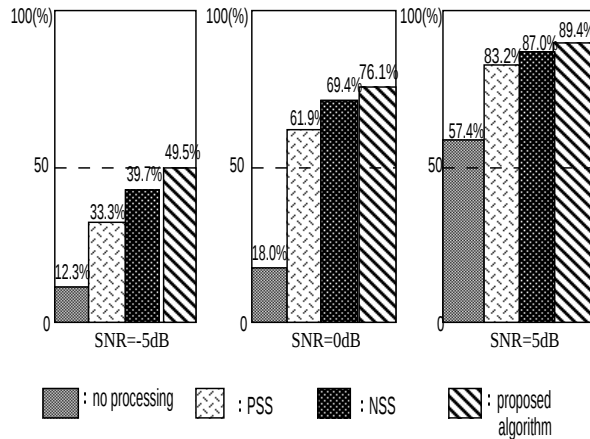


図 3：観測信号と PSS と NSS による強調信号と提案手法による強調信号の認識率

図 2(e)～(h) に信号 (a)～(d) のスペクトログラムを示す。(h) は (g) に比べ幾ど雑音が無くなっているのが見て取れる。しかし (h) を原信号のスペクトログラム (e) と比較すると音声の低周波数域の成分は残っているが高調波成分は大きく歪んでいるのが分かる。このため音声は歪んで聴こえる。

4.2 音声認識

認識装置には HMM、音声は女性 1 人が発音した 18 個の単語を用いた。背景雑音には白色雑音を $SNR = -5dB, 0dB, 5dB$ で付加した。この観測信号に提案手法、PSS、NSS、何も処理をしていない場合の 4 種類の信号を用い認識率の比較を行った。

この結果を図 3 に示す。シミュレーションの結果スペクトルサブトラクション法を用いた信号や観測信号より提案手法を用いた信号の方が認識率が高いといえる。特に $SNR = -5dB$ の場合は PSS を用いた場合は 33.3%、NSS を用いた場合は 39.7%、提案手法を用いた場合は 49.5% となり、改善が見られた。 $SNR = 5dB$ の場合においても僅かではあるが認識率は PSS、NSS を用いた場合に比べ改善されている。

5. まとめ

本報告ではマスキング曲線を用いた音声強調の手法を提案した。音声認識装置のフロントエンドとしての有用性は評価実験の結果より、PSS、NSS よりも提

案手法の方が高いことが分かる。

提案手法の問題点として挙げられるのは先に PSS を用いることで音声の高調波の成分をノイズと判定してしまうことや、引き残したノイズを音声成分として扱ってしまうことなどがある。低い周波数領域においては音声成分と雑音成分を間違えて判定したとしても音声の低周波数域の成分の影響であまり問題にはならないが、高い周波数領域においては、低周波数域の成分に比べ高調波成分はエネルギーが小さいためこの問題が顕著に現れる事になる。この問題点を解決するために現段階では高調波成分を雑音成分として扱うことにより強調信号の高調波成分を減衰させている。

今後の課題として音声成分と雑音成分の誤認を防ぐことが挙げられる。

参考文献

- [1] Steven F.Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-27, no.2, pp113-120, 1979
- [2] Saeed V.Vaseghi, "Advantage Signal Processing and Digital Noise Reduction," Wiley & Son Ltd, 1995
- [3] INTERNATIONAL STANDARD ISO/IEC 11172-3, 1993(E)
- [4] Nathalie Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," IEEE Trans. Speech and Audio Processing., vol. 7, no.2, pp126-137, 1999
- [5] Lockwood and J. Boudy, "Non-linear Spectral Subtractor(NSS): Robust HMMs using NSS and the Projection," Speech communication, no.11, pp215-228, 1992