G-015

雑音既知の条件における音声の調波構造を用いた雑音除去

Noise Reduction Based on the Harmonic Structure of the Speech Signal in the Known Noise Case

大田健紘† 柳田益造†
Kenko Ohta Masuzo Yanagida

１．はじめに

ハンズフリー音声認識では，周囲からの雑音により音声認識率が著しく低下する．また，操作対象がテレビなどそれ自身が音を発する物の場合には，それ自身が発する音が雑音となり音声認識率低下の原因となる．例えば，「たけまる君」[1]のようなシステムである．しかし，テレビが発する音は計算機にとっては既知雑音として扱うことができる．雑音情報が既知であればスペクトル減算法を用いることにより，雑音を除去できるが，それだけでは不十分である．なぜなら，雑音源と受音点間の伝達特性は時刻により変動するなどの場合，受音点での雑音の推定が不確かになるためである．そこで，適応的な伝達特性推定処理を導入することにより，雑音の推定精度を向上させ，音声認識率の向上を試みる．

２．雑音除去の方法

２．１ 一般的なスペクトル減算

提案法では，まず事前に取得した N 点の雑音源信号 $\boldsymbol{n}(t)$ とその受音点信号 $\boldsymbol{n}_r(t)$ を用いて，雑音源と受音点間の伝達特性 $\boldsymbol{h}_{nr}(t)$ を推定する．$\boldsymbol{n}_r(t)$ は

$$\boldsymbol{n}_r(t) = F^{-1}[F[\boldsymbol{n}(t)]F[\boldsymbol{h}_{nr}(t)]] \tag{1}$$

と記述できる．ただし，$F[\cdot]$はフーリエ変換，$F^{-1}[\cdot]$はフーリエ逆変換を表す．すると，$\hat{\boldsymbol{h}}_{nr}(t)$ は

$$\boldsymbol{h}_{nr}(t) \cong \hat{\boldsymbol{h}}_{nr}(t) = F^{-1}\left[\frac{F[\boldsymbol{n}_r(t)]}{F[\boldsymbol{n}(t)]}\right] \tag{2}$$

として推定することができる．

推定した伝達特性を用いて雑音源信号から受音点信号を推定する際，雑音源と受音点間の伝達特性が時刻により変動する場合，事前に推定した伝達特性を当該時刻の伝達特性に適応させる必要がある．つまり，ある時刻 k 以降の N 点の受音点信号 $\boldsymbol{n}(t)_k$ は，雑音源信号及び伝達特性を用いて

$$\boldsymbol{n}_r(t)_k = F^{-1}[F[\boldsymbol{n}(t)_k]F[\boldsymbol{h}_{nr}(t)_k]] \tag{3}$$

と書くことができる．ただし，

$$\boldsymbol{n}(t)_k = [n(k), n(k+1), \cdots, n(k+N-1)]^T$$
$$\boldsymbol{n}_r(t)_k = [n_r(k), n_r(k+1), \cdots, n_r(k+N-1)]^T$$
$$\boldsymbol{h}_{nr}(t)_k = [h_{nr}(k), h_{nr}(k+1) \cdots, h_{nr}(k+N-1)]^T$$

である．事前に推定した伝達特性は $\hat{\boldsymbol{h}}_{ns}(t)$ であるため，ある時刻に得られる推定雑音は

$$\hat{\boldsymbol{n}}_r(t)_k = F^{-1}[F[\boldsymbol{n}(t)_k]F[\hat{\boldsymbol{h}}_{nr}(t)]] \tag{4}$$

である．この推定雑音 $\hat{\boldsymbol{n}}_r(t)_k$ では十分な減算をすることができない．

２．２ 伝達特性の適応化方法

前節で適応処理の必要性を述べたが，本節では適応処理の方法について述べる．事前に推定した伝達特性を，ある時刻における伝達特性に適応させるには，ある時刻における伝達特性に関する情報を受音信号から抽出することができればよい．この情報は，受音点における雑音信号に含まれていることがわかる．しかし，受音点においては，目的音声と雑音が混合されている．そこで提案法では，まず，基本周波数を推定し，それを基に受音信号中に含まれている目的音声を推定する．そして，推定した目的音声を受音点信号から減算することにより，当該時刻における伝達特性の情報を含む雑音信号を推定することができる．最後に，事前に推定した伝達特性を畳み込んだ受音点雑音と，推定した雑音信号を用いて，適応処理を行う．これにより雑音除去性能の向上が期待できる．以下では具体的な方法について説明する．

２．２．１ 音声信号の推定

事前に推定した伝達特性をある時刻の伝達特性に適応させる方法について説明する．本研究では，雑音源信号は既知と仮定する．この条件下で受音信号から雑音を除去することを考える．まず受音信号 $\boldsymbol{r}_r(t)_k$ に含まれる目的音声 $\boldsymbol{s}(t)_k$ の基本周波数を推定する．雑音を含んだ $\boldsymbol{r}_r(t)_k$ から基本周波数を推定するのでは十分な精度を得ることができないため，$\boldsymbol{r}_r(t)_k$ から $\hat{\boldsymbol{n}}_r(t)_k$ をスペクトル上で減算を行い，その結果を用いて基本周波数を推定する．受音信号 $\boldsymbol{r}_r(t)_k$ は

$$\boldsymbol{r}_r(t)_k = \boldsymbol{s}(t)_k + \boldsymbol{n}_r(t)_k \tag{5}$$

と書けるので，目的音声 $\boldsymbol{s}(t)_k$ の推定値 $\hat{\boldsymbol{s}}(t)_k$ は

$$\hat{\boldsymbol{s}}(t)_k = F^{-1}[F[\boldsymbol{r}_r(t)_k] - \alpha F[\hat{\boldsymbol{n}}_r(t)_k]] \tag{6}$$

として求めることができる．ただし，α は減算率を表す．そして，$\hat{\boldsymbol{s}}(t)_k$ に対してケプストラムなどを用いて基本周波数の推定を行う．ただし，音声・非音声の判定を行い，残留雑音により非音声区間に誤って推定された基本周波数を取り除いている．そして，推定した基本周波数に基づいて $\boldsymbol{r}_r(t)_k$ に含まれる $\boldsymbol{s}(t)_k$ を推定する．しかし，音声区間は母音のように調波構造を有する区間と，子音のように調波構造を持たない区間が存在する．そのため，音声区間を摩擦音のように高周波数に強いスペクトル持つ区間と，母音の区間に分類する．そして，区間ごとに $\boldsymbol{s}(t)_k$ の推定方法を切り替える．

(ｉ)母音区間での $\boldsymbol{s}(t)_t$ の推定

母音区間は調波構造を有しているため，基本周波数の整数倍の正弦波の足し合わせにより推定することが可能である．そのため，母音区間での $\boldsymbol{s}(t)_k$ の推定は，式(7)に従い正弦波の足し合わせにより行う．

$$s_{est}(k) = \sum_{m=1}^{M} a_m \sin(2\pi f_0 mkT + \phi_m) \tag{7}$$

ただし，a_m は正弦波の振幅であり，f_0 は推定した基本周波数，m は高調波番号，T はサンプリング周期，ϕ_m は位相，M は帯域内の高調波の数で

† 同志社大学

$$M = \frac{sampling\_rate}{2f_0} \tag{8}$$

により決定される．現在，振幅のパワースペクトル上での減算を行っているので，位相 ϕ_m は 0 としている．また，振幅 a_m は，本来であれば音声のモデルを考慮して決めるべきであるが，式(7)で推定した信号の振幅は最大でも当該フレームの最大振幅にしかならないと考え，各周波数に対して一定の値として式(9)により推定する．

$$a = \frac{\max_{0 \le l < N-1} r_r(k+l)}{M} \tag{9}$$

ただし，N は FFT の点数である．

(ii)摩擦音区間での $\boldsymbol{s}(t)_k$ の推定

摩擦音区間は母音区間とは異なり，調波構造を有していない．そのため，式(7)による推定が行えないので，受音信号から雑音信号を減算することで摩擦音区間での $\boldsymbol{s}(t)_k$ が推定できると仮定している．つまり，式(10)に従い推定する．

$$\boldsymbol{s}_{est}(t)_k = F^{-1}[F[\boldsymbol{r}_r(t)_k] - F[\hat{\boldsymbol{n}}_r(t)_k]] \tag{10}$$

(iii)母音，摩擦音以外の区間での $\boldsymbol{s}(t)_k$ の推定

最後に母音，摩擦音以外の区間での $\boldsymbol{s}(t)_k$ の推定は，無音区間であると考えて，

$$\boldsymbol{s}_{est}(t)_k = \boldsymbol{0} \tag{11}$$

とする．

(i)～(iii)より推定した $\boldsymbol{s}_{est}(t)_k$ は

$$\boldsymbol{s}(t)_k \cong \boldsymbol{s}_{est}(t)_k \tag{12}$$

と考えることができるため，

$$\boldsymbol{n}_r(t)_k \cong \boldsymbol{n'}_r(t)_k = F^{-1}[F[\boldsymbol{r}_r(t)_k] - F[\boldsymbol{s}_{est}(t)_k]] \tag{13}$$

として，ある時刻の受音点雑音の近似値 $\boldsymbol{n'}_r(k)$ が得られる．ただし，

$$\boldsymbol{n'}_r(t)_k = [n'_r(k), n'_r(k+1), \cdots, n'_r(k+N-1)]^T$$

とする．

2．2．2　適応処理

式(13)で表される $\boldsymbol{n}_r(t)_k$ の近似値を用いて，事前に推定した伝達特性 $\hat{\boldsymbol{h}}_{nr}(t)$ をある時刻における伝達特性 $\boldsymbol{h}_{nr}(t)_k$ への適応化を行う．式(13)の近似が成り立つことから，$\boldsymbol{n'}_r(t)_k$ は $\boldsymbol{n}(t)_k$ と $\boldsymbol{h}_{nr}(t)_k$ の畳み込みであると考えられる．そのため，

$$\varepsilon(k) = n'_r(k) - \sum_{l=0}^{N-1} n(k+N-1-l) h_{ad}(k+l) \tag{14}$$

の 2 乗値が最小になるように $\boldsymbol{h}_{ad}(t)_k$ の適応を行う．ただし，適応処理の初期値には $\hat{\boldsymbol{h}}_{nr}(t)$ を用いる．つまり，

$$\frac{\partial \varepsilon^2(k)}{\partial h_{ad}(k+\tau)} = 2n(k+N-1-\tau)\varepsilon(k) = 0 \tag{15}$$

となるように $\boldsymbol{h}_{ad}(t)_k$ を

$$h_{ad}^{(p+1)}(k) = h_{ad}^{(p)}(k) + \beta n(k+N-1-\tau)\varepsilon(k) \tag{16}$$

として更新する．ただし，p は反復回数を表す．そして，適応処理を施した伝達特性を $\boldsymbol{n}(t)_k$ に畳み込み $\boldsymbol{n}_r(t)_k$ を推定しスペクトル減算を行う．適応処理を施した伝達特性 $\boldsymbol{h}_{ad}^{(p+1)}(t)_k$ は，次のサンプル点での適応処理の初期値 $\boldsymbol{h}_{ad}^{(0)}(t)_{k+1}$ となる．

3．実験

本研究では，テレビを音声により操作するという状況で実験を行った．図 1 に本研究で想定している状況を示す．

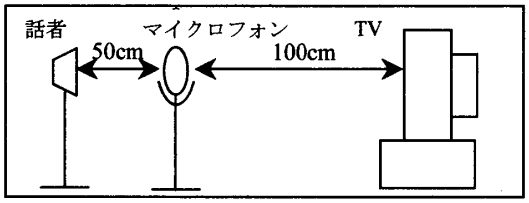

図 1　本研究で想定しているシステム利用状況

3．1　実験データ

実験には，ラインでの認識率が 100%の音声データを用い，それらを図 1 に示す状況においてスピーカから流した．そして，雑音には複数の人の声及び音楽が含まれている TV 音を用いた．S/N は 0dB，6dB，12dB および 18dB の 4 種類とする．被験者は男性 3 名，女性 1 名であり，各被験者は 50 個のテレビ操作コマンドを発話した．発話内容は「テレビ ON」や「音量大きくして」などである．音声認識のデコーダには"Julian"を用いている．

3．2　結果

被験者 4 名の音声認識率の平均値を図 2 に示す．図 2 は雑音除去処理を行っていない(Baseline)，スペクトル減算のみを行った(SS)そして提案法(Proposed)の 3 種類を比較した．

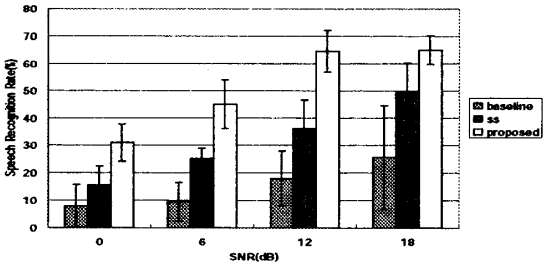

図 2　音声認識率の比較（エラーバーは標準偏差）

図 2 より，SS 法のみを用いた場合でも音声認識率は改善しているが，提案法を用いることによりさらに改善することがわかる．

4．考察

今回の実験では，ライン入力での認識率が 100%の音声データを認識対象としてスピーカから流し，複数の人の声及び音楽が含まれている TV 音を雑音とした．しかし，TV 音の種類としては，他にも砂嵐や人の声のみの場合などが考えられる．これらの雑音に対しても同様の改善がみられるか実験を行う必要がある．しかし，砂嵐の場合は，S/N が 12dB 程度であったとしても，テレビ操作コマンド音声のスペクトルがマスクされてしまうので，うまく動作しないものと考えられる．

参考文献

[1] 李 晃伸，山田真士，鹿野清宏，西村竜一，“公共音声情報案内システム「たけまるくん」の改善”，日本音響学会講演論文集，1-P-27, pp.215-216, Sept. 2004