社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (12)
2006／12／21

雑音低減のための複素周波数領域における参照再構成法

井原 健紘† 高木 一幸† 尾関 和彦†

† 電気通信大学 情報工学専攻
〒182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †{ihara,takagi,ozeki}@ice.uec.ac.jp

あらまし 雑音が入力に混入した場合の自動音声認識の性能を上げるべく，本稿では音声に雑音が重畳した単一チャネルの信号から音声のみを抽出する手法について述べる．このとき，入力信号と同じ話者の異なる発話を小規模のデータベースとして用いることができるものと仮定した．著者らは[1]において同様の問題に対し，ある尺度で入力フレームと類似しているフレームをデータベース内から抽出し，その抽出したフレームを参考にして出力を得るという手法を提案しているが，本稿ではさらにその類似尺度と出力方法の改良法を報告する．改良の要点は，短時間フーリエ変換後の位相情報をそのまま保持しておくことと，そこにバイナリマスクをかけることの二点である．従来，位相情報は絶対値処理により捨てられていたが，著者らは雑音低減には有用であろうと判断した．また，そこにバイナリマスクをかけることは，時間領域の信号から雑音成分を取り除くこととほぼ等価である．性能評価をするために雑音にSNR0dBの器楽曲および環境雑音を用いて単語認識実験をおこなったところ，約58%の単語正解率が得られた．ただし，有声音区間・無声音区間・無音区間の判定はまだ自動化されていない．

キーワード 雑音低減，参照再構成法，複素周波数，周波数マスク，最近傍法

Referential Reconstruction in Complex Frequency Domain for Noise Reduction

Takehiro IHARA†, Kazuyuki TAKAGI†, and Kazuhiko OZEKI†

† Department of Computer Science, the University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan
E-mail: †{ihara,takagi,ozeki}@ice.uec.ac.jp

Abstract This paper presents a method for extracting the speech signal from the single-channel speech signal contaminted by the noise in order to improve the performance of automatic speech recognition of the noise contaminated input signal. It is assumed that the small database of utterance by the same speaker of the input signal that differ from the input signal can be used. For the same problem, the authors presented a method in [1] that extracts frames similar to the input frames by some similarity measure from the small database, and then produces output frames by refering the similar frames. In this paper, an improved similarity measure and the production process of the outputs is reported. The main improved points are keeping the phase information of Fourier transformed frames instead of discarding it, and applying the binary mask to the frames. While the phase infomation is conventionally discarded by the process of obtaining absolute spectrum, the authors consider it as worth infomation for noise reduction. Applying the binary mask to the Fourier transformed frames has the meaning similar to removing the noise component from the signal in the time domain. For evaluation, words recognition experiments by using instrumental music and environmental noise of SNR of 0dB were performed. The correctness was approximately 58%. The judgement of voiced and unvoiced speech and silent part has not been automated yet.

Key words noise reduction, referential reconstruction, complex frequency, frequency mask, nearest neighbor

1. はじめに

何年前からなのかは知らないが，長らく雑音のせいで自動音声認識の性能が落ちるといわれ続け，雑音低減の研究，および雑音に頑健な音響モデルの研究などがおこなわれてきた．なぜそのような研究をする必要があるのかといえば，ひとえに現在の特徴量の主流であるケプストラムが雑音に弱くできているからである．フレーム内のすべてのサンプルの情報を用いてフーリエ変換を施し，対数周波数領域でのすべての帯域の情報を用いて包絡情報を抽出するのだから，雑音に弱いのは当然である．しかしながら，ほとんどの状況で我々人間は苦労せずに他人の声を聞きとることができることから，我々の耳が雑音に対して何か特別に複雑なこと—例えばプリプロッセッシングをしたりポストプロセッシングをしたり—をしているとは考えづらい．おそらく人間の耳は特徴抽出の段階で何か単純な方法を用いて雑音対策をとっているものと思われる．筆頭著者は，再び時間領域での信号処理の段階から雑音除去を見直すことにより，その単純な雑音対策の方法を探る足がかりを見つけたいと考えている．本稿で述べる手法は単なる雑音低減の手法に過ぎないが，背景にはこのような思惑が存在している．

さて，雑音低減の研究には様々な分類方法があるが，ひとまず入力チャネル数に関することから触れたい．一つ確かなのは，現段階では入力チャネルが複数のとき [2] と単一のときでは性能にかなりの差があるということである．それは空間情報が使えるかどうかという点に由来する差である．人間の聴覚特性を考えたときに，例えば単一のスピーカーしか持たないラジオの音も分離できることから，雑音低減に空間情報は必要ないものと思われる．ゆえに，本研究では入力は単一チャネルのみを用いることにする．単一チャネルの研究では，スペクトルサブトラクション [3] が古くから使われている．単純で実時間処理が可能であり性能が高いという点で有効性の観点からすれば今のところ最も使いやすい雑音低減の方法である．ただし，背景雑音の急激な変化に耐えられないという弱点がある．スペクトルサブトラクションは雑音を推定してそれを入力から減算するという方法であるが，雑音を推定するということに無理があると思われる．雑音とは入力から所望の信号を引いた残りであり，ネガティヴに定義される存在である．ネガティヴに定義されるものを推定するのは厳密に解こうとすれば相当難しい作業である．本研究では，音声のデータベースを使うことにより（つまり音声をポジティヴに定義することにより），雑音低減をする．単一入力チャネルという条件下で音声のデータベースを使って雑音低減をおこなっている研究も多数存在する．例えば [4] では 2 名分の音声を混合したものを入力とし，その分離をおこなっている．このときデータベースとしてその 2 名両方の音声を保持している．人間の耳でも 2 名分の音声を聞き取るのはかなりの集中力のいる作業であるので，問題が少々難しすぎるきらいもあるが，結果的には成功しているようである．また，単一チャネルという条件下で全くデータベースを使わない研究もある [5]．この研究では基本周波数を雑音低減のための主な手がかりとしているが，もしもヴィオラなどの楽器が音声と同じ基本周波数で演奏したら人間の声と区別がつかないのではないかと思われる．したがって，人間の声の特性を示す何らかの事前知識は必要なのではないかと思われる．ただし，これから本稿に書くようなその人物の音声そのものを知っているという仮定は機械に有利すぎる条件かもしれない．

著者らは [1] において参照再構成法 (Referential Reconstruction) という手法を提案した．これは，単一話者の音声に背景雑音が混ざった入力信号から，音声のみを取り出すための手法である．前もってその話者の異なる発話を用意しておき，フレームごとにその用意しておいた小規模のデータベースから似た音声波形を探し，その探した波形をそのまま出力フレームとして使用する．[1] では入力信号とデータベースとの距離尺度を相関係数としたが，それでは結局，白色雑音（つまり無相関雑音）のときのみにしか期待どおりに動作しなかった．本稿では距離尺度に改良を加えた手法について述べる．距離尺度には W-DO (W-Disjoint Orthogonality) [6] の仮定を用いた．また，出力方法の改良もおこなった．

なお，本手法は最近傍法 (Nearest Neighbor) [7] の一種と見ることができる．本稿では距離尺度に非線形処理を採り入れたために最近傍法のイメージからは遠ざかってしまったが，改良前の手法は紛れもなく最近傍法であった．そういった意味では，類似の手法として圧縮の手法であるベクトル量子化 (Vector Quantization) [8]，制御の手法である Lazy Learning [9] などが挙げられる．また，波形の素片接続をしているので，素片接続型の音声合成の一種 [10] と見ることもできる．

2. 参照再構成法

まず，本稿で提案する距離尺度を説明する前に，改良前の手法について説明する．

本稿では式 (1) のように，所望の信号 $s(t)$ に加法性雑音 $n(t)$ が加算され入力信号 $x(t)$ として観測されるとする．

$$x(t) = s(t) + n(t), \tag{1}$$

雑音低減の目的は，この $n(t)$ を取り除き $s(t)$ を復元することである．

提案手法は $x(t)$ をフレームごとに処理する．つまり式 (1) は次式のように書き表されることになる．

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}. \tag{2}$$

ここで，$\boldsymbol{x}$，$\boldsymbol{s}$，$\boldsymbol{n}$ はそれぞれ観測信号，所望の信号，加法性雑音のフレーム（ベクトル）表現である．本稿での雑音低減の問題は所望の信号フレーム $\boldsymbol{s}$ を観測フレーム $\boldsymbol{x}$ から復元するという作業となっている．推定フレーム $\hat{\boldsymbol{s}}$ が復元されたら，それらを接続することによりクリーン音声の推定信号 $\hat{s}(t)$ を得る．

なおこのとき，所望の音声の発話者が発話した別の内容の発話を小規模のデータベースとして扱うことができるものとする．このデータベースを今後「参照信号」と呼ぶことにし，r で表す．またこのフレーム表現を $\boldsymbol{r}$ とする．

2.1 基本的なアルゴリズム

提案手法の処理は，まず最初に参照信号中から入力フレーム

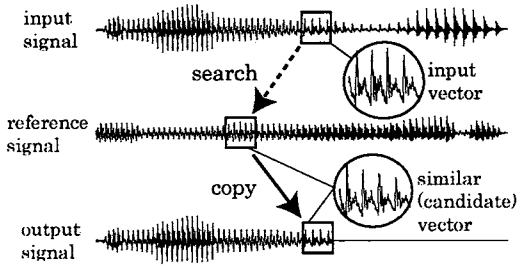

図 1　参照再構成法の概略

Fig. 1　Outline of referential reconstruction.

に似たフレームを探すことから始まる．この探したベクトルを類似フレームと呼ぶことにする．そして，類似フレームのパワーと位相を調節し出力フレームとする．図で示すと図 1 のようになる．提案手法は，所望のフレーム $\boldsymbol{s}$ と似ている参照フレームが参照信号中に存在するはずである，という仮定を用いている．

このアルゴリズムを詳細に書くと以下のようになる．

（1）　入力信号と参照信号をフレームに分割し，$\boldsymbol{x}_i$ と $\boldsymbol{r}_j$ を作る．ただし，$\boldsymbol{x}_i$ および $\boldsymbol{r}_j$ はそれぞれ $\boldsymbol{x}$ の i 番目のフレーム，$\boldsymbol{r}$ の j 番目のフレームを表す．

（2）　入力フレームと参照フレームとの類似度を計算する．改良前の手法では類似度として相関関数を用いていたのでそれにならって説明をすると次のようになる．$\boldsymbol{x}_i$ と $\boldsymbol{r}_j$ を短時間フーリエ変換し，$\boldsymbol{X}_i$ と $\boldsymbol{R}_j$ を作る．その積 $\boldsymbol{X}_i\mathrm{conj}(\boldsymbol{R}_j)$ の逆フーリエ変換を計算し相関関数 $C(\tau)_{i,j}$ を得る．ここで $\mathrm{conj}(\cdot)$ は複素共役を示す．$\frac{C(\tau)_{i,j}}{\|\boldsymbol{X}_i\|\,\|\boldsymbol{R}_j\|}$ を計算し，正規化相関関数 $NC(\tau)_{i,j}$ を得る．ここで $\|\cdot\|$ はノルムを示す．$NC(\tau)_{i,j}$ の τ に関する最大値を類似度とする．

（3）　類似度が最大の参照フレームを類似フレームとする．$\max_\tau NC(\tau)_{i,j}$ が最大のインデックス j を $maxj$ とおく．

（4）　音声でないと判断されたフレームは出力をゼロベクトルとするようにする．もし $\max_\tau NC(\tau)_{i,maxj}$ が閾値以下ならステップ (5) の処理を行わず，出力フレーム $\boldsymbol{y}_i$ を 0 ベクトルとする．

（5）　類似フレームの位相とパワーを調節して出力 $\boldsymbol{y}_i$ を得る．$NC(\tau)_{i,maxj}$ が最大となる τ を $max\tau$ とする．この $max\tau$ のサンプル数だけ参照信号 r 中の $\boldsymbol{r}_{maxj}$ の時刻をずらし，$\hat{\boldsymbol{r}}_{maxj}$ とする．次式によって，出力 $\boldsymbol{y}_i$ を算出する．

$$\boldsymbol{y}_i = \frac{\boldsymbol{x}_i \cdot \hat{\boldsymbol{r}}_{maxj}}{\|\hat{\boldsymbol{r}}_{maxj}\|^2}\hat{\boldsymbol{r}}_{maxj}. \tag{3}$$

ここで $\cdot$ は内積を示す．

もし $\hat{\boldsymbol{r}}_{maxj} \approx \alpha\boldsymbol{s}_i$ のように所望のフレームと類似した類似フレーム $\hat{\boldsymbol{r}}_{maxj}$ が見つかったとしたら，式 (3) の処理によって，以下のようにパワーが調節される．ここで α は調節されるべき定数である．また i 番目の雑音フレーム $\boldsymbol{n}_i$ は，i 番目の所望のフレーム $\boldsymbol{s}_i$ とは相関がないと仮定する．式 (2) を用いて，式 (3) は次のように変形できる．

$$\begin{aligned}\boldsymbol{y}_i &= \frac{\boldsymbol{x}_i \cdot \hat{\boldsymbol{r}}_{maxj}}{\|\hat{\boldsymbol{r}}_{maxj}\|^2}\hat{\boldsymbol{r}}_{maxj} \\ &\approx \frac{(\boldsymbol{s}_i + \boldsymbol{n}_i)\cdot\alpha\boldsymbol{s}_i}{\|\alpha\boldsymbol{s}_i\|^2}\alpha\boldsymbol{s}_i \\ &= \frac{\alpha\|\boldsymbol{s}_i\|^2}{\alpha^2\|\boldsymbol{s}_i\|^2}\alpha\boldsymbol{s}_i \\ &= \boldsymbol{s}_i. \end{aligned} \tag{4}$$

3. 改良手法

参照再構成法の処理過程を大まかに分類すると三つのステップとなる．一つ目は類似フレームの探索であり，二つ目は音声か否かの判定であり，三つ目は出力フレームの生成である．これらを本稿で説明する改良手法では，さらに有声音と無声音の二つに分類する．つまり六つの処理過程に分かれることになる．以下，これらを個別に説明していくことにする．ただし，処理の順番と説明の順番は異なる．

なお，有声音および無声音の判定については，まだ研究途中であるので音声データベースに記載されている答え（ラベルデータ）を用いている．

3.1 W-disjoint orthogonality

個々の処理過程を説明する前に改良手法で用いる "W-disjoint orthogonality (W-DO)" [6] という仮定について説明する．これは，信号を時間-周波数解析した際に，「ある時刻・ある周波数には一つの信号源に由来する成分しか含まれない」というものである．この仮定に従えば，入力に雑音が混ざっていたとしても時間-周波数解析をして適切なバイナリマスクを適用すれば雑音を取り除くことができるということになる．

[6] には人間の声の重畳に関する議論しかなされていないが，著者らが簡易的に実験をしたところ 0dB の SNR でクラシック音楽を雑音として重畳させた入力信号でも，適切に（理想的な）マスクをかければ 131 単語の単語音声認識実験に対して 97%の正解率を得ることができた．このことから，W-DO は雑音が音楽のときにも適用できる可能性があるといえる．

なお，以降の数式の記述を簡略化するため，任意の関数 $f(\omega)$ に対して次式のような $mask$ という関数を定義することにする．

$$mask(f(\omega)) = \begin{cases} 0 & if\ \ f(\omega) < threshold \\ 1 & otherwise \end{cases} \tag{5}$$

なお，$threshold$ には $f(\omega)$ の絶対値の平均を用いる場合と，マスクの 0 と 1 が一定の割合になるようにする場合の二種類がある．

3.2 無声音の出力

この処理過程は 2.1 節のステップ (5) の半分に相当する．無声音として選ばれた類似フレームのパワーと位相を調節するのが目的である．

まず，類似フレーム $\boldsymbol{r}_{maxj}$ を元にバイナリマスク $mask(\boldsymbol{R}_{maxj})$ を作る．そのマスクを入力フレームのフーリエ変換と類似フレームのフーリエ変換の双方に適用し，マスク後の入力フレームとマスク後の類似フレームを得，相関関数 $NC_{i,maxj}$ を計算する．

$$\boldsymbol{X}_{i,maxj}^{(masked)} = mask(\boldsymbol{R}_{maxj})\boldsymbol{X}_i. \quad (6)$$

$$\boldsymbol{R}_{maxj}^{(masked)} = mask(\boldsymbol{R}_{maxj})\boldsymbol{R}_{maxj}. \quad (7)$$

$$NC_{i,maxj} = ifft\left(\frac{(\boldsymbol{X}_{i,maxj}^{(masked)})\mathrm{conj}(\boldsymbol{R}_{maxj}^{(masked)})}{||\boldsymbol{X}_{i,maxj}^{(masked)}||\,||\boldsymbol{R}_{maxj}^{(masked)}||}\right). \quad (8)$$

相関関数が最大値となる時刻インデックスを得る．その時刻インデックスを用いて，類似フレームの時刻を調節し，時刻調節後の類似フレーム $\hat{\boldsymbol{r}}_{maxj}$ を得る．

マスク後の入力フレームとマスク後の類似フレームを逆フーリエ変換して時間域に戻し，雑音成分を消した入力フレーム $\boldsymbol{x}_{i,maxj}^{(masked)}$ と雑音成分を消した類似フレーム $\hat{\boldsymbol{r}}_{maxj}^{(masked)}$ を得る．それらに式 (11) を適用することにより，パワーを調節した出力フレーム $\boldsymbol{y}_i$ を得る．

$$\boldsymbol{x}_{i,maxj}^{(masked)} = ifft(\boldsymbol{X}_{i,maxj}^{(masked)}). \quad (9)$$

$$\hat{\boldsymbol{r}}_{maxj}^{(masked)} = ifft(\hat{\boldsymbol{R}}_{maxj}^{(masked)}). \quad (10)$$

$$\boldsymbol{y}_i = \frac{\boldsymbol{x}_{i,maxj}^{(masked)} \cdot \hat{\boldsymbol{r}}_{maxj}^{(masked)}}{||\hat{\boldsymbol{r}}_{maxj}^{(masked)}||^2}\hat{\boldsymbol{r}}_{maxj}. \quad (11)$$

3.3 有声音の出力

この処理過程は 2.1 節のステップ (5) の半分に相当する．有声音として選ばれた類似フレームを利用して適切なバイナリマスクを作り，そのマスクを入力に適用することにより出力を得ることが目的である．なお，3.2 節と全く同じ手法を有声音に適用して実験をしたところ，耳で聞いたかぎりでは雑音の低減性能は本節の手法よりもよくなったが，単語認識率が悪くなった．

まず，3.2 節と全く同じ手法によって，仮出力フレーム $\hat{\boldsymbol{y}}_i$ を作る．入力信号と仮出力フレームのパワーの差を比較し，入力信号が閾値よりも上回っているとき，雑音成分が多く混入していると見なし，その周波数成分を 0 とする．その入力にマスクをかけ逆フーリエ変換をし，出力フレーム $\boldsymbol{y}_i$ を得る．なお β は定数である．

$$\hat{\boldsymbol{X}}(\omega)_i = \begin{cases} 0 & if \quad \boldsymbol{X}(\omega)_i > \beta\hat{\boldsymbol{Y}}(\omega)_i \\ \boldsymbol{X}(\omega)_i & otherwise \end{cases}. \quad (12)$$

$$\boldsymbol{y}_i = ifft(\mathrm{mask}(\boldsymbol{R}_{maxj})\hat{\boldsymbol{X}}_i) \quad (13)$$

3.4 有声音の探索

この処理過程は 2.1 節のステップ (2) の半分に相当する．有声音の適切な類似ベクトルを得ることが目的である．この処理過程は二段階に分かれている．一段階目の処理のみでも雑音低減は可能であるが，二段階目の処理を適用した方が性能が高まる．

まず前処理として，参照信号全体から有声音のフレームのみを抽出する．参照信号には雑音が乗っていないので，パワーやゼロ交差回数などのルールベースで抽出することができる．また，有声音以外を有声音として抽出してしまうのは問題であるが，有声音のフレームを見逃してしまうのは問題ではないので粗い処理でよい．

まず，一段階目の処理として，有声音として選ばれた参照フレームからそれぞれのフレームに対応するバイナリマスクを作る．このバイナリマスクを入力フレームのフーリエ変換および参照フレームのフーリエ変換に適用する．それらから正規化相関関数を計算する．

$$\boldsymbol{X}_{i,j}^{(masked)} = mask(\boldsymbol{R}_j)\boldsymbol{X}_i. \quad (14)$$

$$\boldsymbol{R}_j^{(masked)} = mask(\boldsymbol{R}_j)\boldsymbol{R}_j \quad (15)$$

$$NC_{i,j}(\tau) = ifft\left(\frac{(\boldsymbol{X}_{i,j}^{(masked)})\mathrm{conj}(\boldsymbol{R}_j^{(masked)})}{||\boldsymbol{X}_{i,j}^{(masked)}||\,||\boldsymbol{R}_j^{(masked)}||}\right). \quad (16)$$

正規化相関関数の最大値が類似度となる．

$$similarityFirst_{i,j} = \max_{\tau}(NC_{i,j}(\tau)) \quad (17)$$

参照フレームにしたがって作ったマスクが間違っていれば，類似度は低くなることが期待される．また，マスクが適切なとき，入力から雑音の成分が取り除かれるので，この正規化相関関数は入力に雑音がない状態での正規化相関係数に相当する．

次に，二段階目の処理として，第一段階の類似度の高い方から N 個の参照フレームを選び (4. 節の実験では $N = 100$)，この中から類似フレーム一つを決定する．まず，3.2 節（無声音の出力）と同様の方法を用いて，それぞれの参照フレームから位相とパワーを調節した仮出力フレーム $\hat{\boldsymbol{y}}_{i,j}$ を N 個計算する．次に，入力フレーム $\boldsymbol{X}_i$ と仮出力フレーム $\hat{\boldsymbol{Y}}_{i,j}$ の低帯域（低い方から L%．4. 節の実験では $L = 20$）を 0 にする．その後，一段階目と同様にマスクをかける．これらから次の二つの相関係数を計算する．この第二段階目では要するに高帯域のみを頼りに，第一段階目とほぼ同様の尺度で類似度を測っている．

$$CC_{i,j}^{(masked)(X)} = \frac{(\boldsymbol{X}_i^{(masked)}) \cdot \mathrm{conj}(\hat{\boldsymbol{Y}}_j^{(masked)})}{||\boldsymbol{X}_i^{(masked)}||^2}. \quad (18)$$

$$CC_{i,j}^{(masked)(\hat{Y})} = \frac{(\boldsymbol{X}_i^{(masked)}) \cdot \mathrm{conj}(\hat{\boldsymbol{Y}}_j^{(masked)})}{||\hat{\boldsymbol{Y}}_j^{(masked)}||^2}. \quad (19)$$

もし，入力フレームで正規化した相関係数と仮出力フレームで正規化した相関係数の差が大きければ，入力フレームと参照フレームの類似度が低いということがいえる．逆に二つが同じくらいならば，入力フレームと参照フレームの類似度が高い可能性があるといえる．類似フレームを決定するための類似度としては，二つの相関係数の小さな方の値を用いる．

$$similarity_{i,j} = \min(CC_{i,j}^{(masked)(X)}, CC_{i,j}^{(masked)(\hat{Y})}). \quad (20)$$

最後に類似度の最も高い参照フレームを類似フレームとする．

3.5 無声音の探索

この処理過程は 2.1 節のステップ (2) の半分に相当する．無声音の適切な類似ベクトルを得ることが目的である．この処理も二段階に分かれる．

まず前処理として，参照信号全体から無声音のフレームのみを抽出する．参照信号には雑音が乗っていないので，パワーやゼロ交差回数などのルールベースで抽出することができる．また，無声音以外を無声音として抽出してしまうのは問題であるが，無声音のフレームを見逃してしまうのは問題ではないので粗い処理でよい．

一段階目の処理として，無声音のみとなった参照信号のそれぞれのフレームからバイナリマスクを作る．このマスクを入力フレームに適用する．入力フレームに適切なマスクが適用されたかどうかを次のような尺度を用いて判定する．マスクが適用された入力フレームを逆フーリエ変換し，時間域の信号に戻す．

$$\boldsymbol{x}_{i,j}^{(masked)} = ifft(\boldsymbol{X}_{i,j}^{(masked)}) \tag{21}$$

このとき，もし適切にマスクが適用されていたとしたら，マスク適用前と同じ時刻に信号の微細な極大値や極小値が残っているはずである．任意の関数 $g(t)$ に対して，

$$(g(t) - g(t+1))(g(t+1) - g(t+2)) < 0 \tag{22}$$

のとき極値であると判定することにする．マスク適用前 $\boldsymbol{x}$ に極値であったのにマスク適用後 $\boldsymbol{x}_{i,j}^{(masked)}$ に極値でなくなった時刻インデックス t の個数をペナルティとしてカウントし，-1 をかけたものをマスクの適切さの尺度の一つとする．また同様に，

$$(g(t) - g(t+2))(g(t+1) - g(t+3)) < 0 \tag{23}$$

についても適切さの尺度を測り，両者の和をマスクの適切さとした．なお，適切なマスクをかければ極値の時刻インデックス t の個数が増えるのは自然であるので，マスク適用前 $\boldsymbol{x}$ に極値でなかったのにマスク適用後 $\boldsymbol{x}_{i,j}^{(masked)}$ に極値になってもペナルティとしてはカウントしない．

二段階目の処理として，マスクの適切さが上位 M 番目までの参照フレームをもとにしたマスクを入力フレームと参照フレームの両方に適用する（4. 節の実験では $M = 50$）．それらの正規化相関係数を類似度とする．

$$similarity = \frac{|\boldsymbol{X}_{i,j}^{(masked)}| \cdot |\boldsymbol{R}_{j}^{(masked)}|}{||\boldsymbol{X}_{i,j}^{(masked)}|| \; ||\boldsymbol{R}_{j}^{(masked)}||} \tag{24}$$

類似度の最も高い参照フレームを類似フレームとする．

3.6 処理過程適用の順番

ここまでに紹介してきた処理過程は，以下の順序で実行される．

最初に有声音の探索（3.4 節）を行う．二番目に有声音の判定を行う．三番目に有声音の出力（3.3 節）を行う．四番目に無声音の探索（3.5 節）を行う．五番目に無声音の判定を行う．六番目に無声音の出力（3.2 節）を行う．なお，最後に無声音でも有声音でもなかったフレームにはゼロベクトルを出力するのであるが，全くの 0 であると特徴量が適切に計算できなくなるため，振幅の極めて小さな雑音を出力することとした．

4. 実　　験

性能の評価のために，単語認識実験をおこなった．認識機としては Julian [11] を用いた．ATR 音声データベース重要語 5420 単語 [12] から 40 単語おきに選出した 131 単語を実験に用いた．なお，女性話者一名のみである．雑音にはクラシック音楽 [13] およびパーティー会場での環境雑音 [14] を用いた．SNR は 0dB とした．また，本手法は雑音を除去しているというよりは音声を再構築しているという趣きがあるので，クリーン音声

表 1　Julian に用いたパラメータ．
Table 1　Configuration used for Julian.

sampling	16bit / 16kHz
window	length: 25ms, shift: 10ms
feature	MFCC(12) + ΔMFCC(12) + Δenergy
HMM	speaker independent
	32 mixtures
	triphone
	diagonal covariance
	3 states

表 2　単語認識率 (%)．ND: 雑音除去をしていない．RR: 参照再構成法（提案手法）．SS: スペクトルサブトラクション．
Table 2　Word recognition rate (%).ND: No Denoising. RR: Referential Reconstruction (presented method). SS: Spectral Subtraction.

	ND	RR	SS
symphony	16.0	59.5	42.0
party	3.8	57.3	34.4
clean	96.9	92.4	96.9

s に対する再構成性能も測った．参照信号には認識単語とは異なる単語 524 個を用い，前もって無音区間は削除した．従来手法として Julian に実装されているスペクトルサブトラクション法 [3] を用いた．スペクトルサブトラクションのパラメータにはデフォルトのものを用いた（減算計数 2.0，フロアリング係数 0.5）．これらの単語音声認識実験には 4130 話者 260 時間の発話から作られた「高精度成人モデル」[15] を用いた．Julian に用いたパラメータを表 1 に示す．また，提案手法による雑音低減時のフレーム長は有声音に対しては 32ms，無声音に対しては 8ms とした．またシフト幅は，入力信号に対してはフレーム長の 25%，参照信号に対しては 50%とした．

表 2 に実験結果を示す．クラシック音楽およびパーティーの環境雑音に対しては，提案手法は従来手法や何も処理をしていない場合よりも高い単語認識率を得ることができた．ただし，前述したが，まだ有声音・無声音の自動判別をしていないので，これを自動化したときにどの程度性能が低下するかは分からない．なお，再構成性能（表の clean）では提案手法は若干認識性能の妨げになった．再構成をしても性能に変化がないのが理想であるが，やはり元の音声とは異なるので性能が下がるのは仕方のないことである．また，図 2–4 に波形を示す．発話単語は「背広」である．クリーン信号 (a) と雑音低減後の信号 (c) がやや異なったので改善の余地はまだあると見られる．雑音重畳信号 (b) から復元したことを考慮すれば，さほど悪い結果ではない．

5. まとめと考察

本稿では，音声に雑音が重畳した単一チャネルの信号から音声のみを抽出する手法について述べた．このとき，入力信号と同じ話者の異なる発話を小規模のデータベースとして用いることができるものとした．著者らは [1] において同様の問題に対し，ある尺度で入力フレームと類似しているフレームをデータ

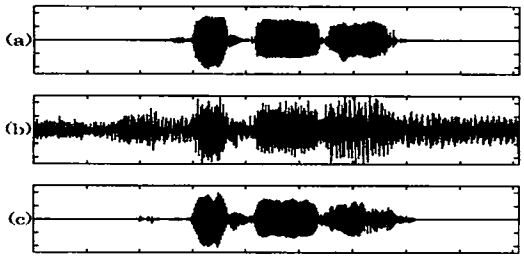

図 2 実験結果. (a) クリーン信号. (b) 雑音を混入させた信号. (c) 雑音低減後の信号.

Fig. 2 The result of the experiments. (a) Clean signal. (b) Signal contaminated by noise. (c) Noise-reduced signal.

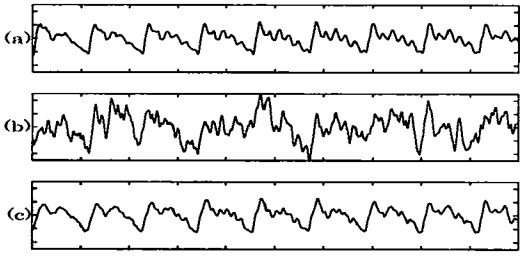

図 3 有声音部分の実験結果. (a) クリーン信号. (b) 雑音を混入させた信号. (c) 雑音低減後の信号.

Fig. 3 The result of the experiments of vowel part. (a) Clean signal. (b) Signal contaminated by noise. (c) Noise-reduced signal.

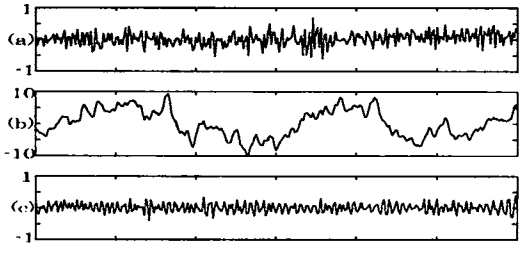

図 4 無声子音部分の実験結果. (a) クリーン信号. (b) 雑音を混入させた信号 (ただし縮尺が (a) と (c) とは異なる). (c) 雑音低減後の信号.

Fig. 4 The result of the experiments of unvoiced consonant part. (a) Clean signal. (b) Signal contaminated by noise (the range differs from both (a) and (c)). (c) Noise-reduced signal.

ベース内から抽出し，その抽出したフレームを参考にして出力を得るという手法を提案しているが，本稿ではさらにその類似尺度と出力方法の改良法を報告した．

クラシック音楽およびパーティーの環境雑音に対しては，提案手法は従来手法や何も処理をしていない場合よりも高い単語認識率を得ることができた．ただし，前述したが，まだ有声音・無声音の自動判別をしていないので，これを自動化したときにどの程度性能が低下するかは分からない．手法としての目的は音声認識性能の向上であるが，その目的を手段として見なした場合には筆頭著者の目的は雑音に頑健な特徴抽出の方法を探ることであるので，後者の目的はすでにある程度達成することができたといえる．

類似尺度および出力方法の両方に周波数領域におけるバイナリマスクを使用した．マスクを使用しなかったとき ([1]) に比べてマスクを使用した今回の方が性能が上がっていることから，マスクの有効性が言える．マスクが有効であるのは，音が周波数領域においてスパースであるせいだと考えられ，時間領域での音の周期性が音源分離において重要な手がかりであると思われる．今後，周期性についてさらなる検討をしたい．

また，類似尺度および出力方法の両方に，周波数領域の絶対値のみならず位相情報も用いた（逆フーリエ変換によって時間信号に戻した）．これまで人間は位相情報を聞いていないと言われていたが，機械にとっては有益な情報である可能性もある．今後，位相情報についてさらなる検討をしたい．

今後の検討課題としては，第一に有声音・無声音の判定手法の確立が挙げられる．そのほか，高速化，話者非依存化，伝達関数非依存化が挙げられる．

文　献

[1] T. Ihara, T. Nagai, K. Ozeki, and A. Kurematsu, "Noise reduction in time domain using referential reconstruction," IEICE Transactions on Information and Systems, Vol.E89-D, No.3, pp.1203-1213, Mar. 2006.

[2] 井原健紘, 半田正樹, 長井隆行, 榊松明, "周波数振分けによるマルチチャネル混合音声の分離と音源定位," 信学論 (A), Vol.J86-A, No.10, pp.998–1009, Oct. 2003.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.27, Issue.2, pp.113–120, Apr. 1979.

[4] S. T. Roweis, "One microphone source separation," Neural Information Processing Systems (NIPS), pp.793–799, 2000.

[5] L. W. DeLiang, and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. on Neural Networks, Vol.10, Issue.3, pp.684–697, May. 1999.

[6] S. Rickard and Ö. Yilmaz, "On the Approximate W-Disjoint Orthogonality of Speech," Proc. ICASSP 2002 pp.529–532, 2002.

[7] T. Cover, P. Hart, "Nearest neighbor pattern classification," IEEE Trans. on Information Theory, Vol.13, Issue.1, pp.21–27, Jan. 1967.

[8] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression, Springer, 1992.

[9] G. Bontempi, M. Birattari, and H. Bersini, "Lazy learning for modeling and control design," Int. Journal of Control, vol.72, no.7/8, pp.643-658 1999.

[10] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR–ν-TALK Speech Synthesis System," Proc. ICSLP-92, vol.1, pp.483–486, 1992.

[11] http://julius.sourceforge.jp/

[12] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech Database User's Manual," ATR Interpreting Telephony Research Laboratories, TR-I-0028, 1988.

[13] L. V. Beethoven, "Symphony No.7 in A major," コロムビアミュージックエンタテインメント, 2004.

[14] "Ambient Noise Database for Telephonometry 1996," NTT Advanced Technology Corporation, 1996.

[15] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository – Software of Continuous Speech Recognition Consortium –," In Proc. ICSLP, 2004.