論　文

音声認識のためのマルチレートシステムを用いたスペクトルサブトラクション法

今井　卓[†]　中垣　淳[†]　柴田　孝次[†]　宮永　喜一[††]

A Study on Spectral Subtraction Using Multirate Systems for Speech Recognition

Suguru IMAI[†], Atsushi NAKAGAKI[†], Koji SHIBATA[†], and Yoshikazu MIYANAGA[††]

あらまし　Gülzow らによりマルチレートシステムを用いたスペクトルサブトラクション法が提案されている[6]．音声認識システムの耐雑音性向上のためにこの手法を前処理として組み込み，音声認識実験を行ったところ広帯域雑音に対しては認識率が大きく向上したが，狭帯域雑音に対しては逆に低下することが分かった．本論文では，広帯域雑音と狭帯域雑音に対する強調処理結果を比較することで認識率低下の原因を解明し，その改良法を提案する．原因解明の結果，狭帯域雑音に対しては帯域分割が適切に行えていないことが判明した．そこで提案法では従来の「分割」による帯域分割手順と新たに導入する「併合」による帯域分割手順を組み合わせることでこの問題を解決する．音声認識実験により提案法を用いることで狭帯域雑音に対する認識率低下を改善できることを示す．提案法で QMF（Quadrature Mirror Filter）を用いた変換を使用すると計算量の増加が見られたが，QMF を用いた変換を DFT ベースの変換に変更することで計算量を低減できた．

キーワード　スペクトルサブトラクション，マルチレートシステム，音声認識，耐雑音性向上

1. まえがき

近年，音声認識技術が発達しカーナビゲーションシステムなどの制御や音声からテキストへの自動変換を行うソフトウェア，音声入力を行える携帯ゲーム機などの音声認識システムを用いた製品が数多く登場するようになった．音声認識を用いる利点として，手や目が塞がった状態でも機器を制御できたり，データ入力などを効率的に行えるなどが挙げられる．音声認識技術は今後も様々な分野で用いられていくと予想される．

音声認識システムは雑音のない環境では十分なレベルの認識性能を示すが，背景雑音が存在するような実環境においては認識性能が著しく劣化する．しかし，音声認識システムを搭載した製品は実環境内で用いられることが多く，実環境内での認識性能をいかに向上させるかが重要な課題となる．この問題への対処法の一つとして音声認識システムの前処理に音声強調処理を組み込み，入力される観測信号の雑音成分を低減することで認識性能を向上させる方法がある[1],[2]．

定常的なスペクトル特性をもった加法性の雑音により劣化した音声を強調するための手法としてスペクトルサブトラクション（SS）法がある．その中でも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よく知られている[3]．しかし，パワースペクトルサブトラクション法では音声スペクトルの引きすぎや雑音スペクトル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がある．そのため，ミュージカルノイズの発生を抑えて強調を行うための手法が検討されている[4],[5]．

Gülzow らはマルチレートシステムを用いた SS 法を提案している[6]．この手法は，推定した雑音との SNR に応じて観測信号の帯域を分割し，観測信号に非一様な時間・周波数分解能をもたせて SS を行うものである．この手法を音声認識システムの前処理として組み込み，数種類の雑音に対して孤立単語音声認識実験を行ったところ，広帯域雑音に対しては他の音声強調法に比べて認識率が大きく向上した．しかし，狭帯域雑音に対しては認識率が逆に低下してしまうという結果を得た．

† 北見工業大学，北見市
Kitami Institute of Technology, Kitami-shi, 090–8507 Japan
†† 北海道大学大学院情報科学研究科，札幌市
Graduate School of Information Science and Technology, Hokkaido University, Kita 14 Nishi 8, Kita-ku, Sapporo-shi, 060–0814 Japan

本論文では，広帯域雑音を付加した音声信号と狭帯域雑音を付加した音声信号を Gülzow らの手法を用いて強調処理したときの結果を比較し，狭帯域雑音に対して認識率が低下する原因を探った．その結果，Gülzow らが用いている帯域分割の手順に問題があることが判明した．そこで，新たに「併合」による帯域分割手順を導入し，従来の「分割」による帯域分割手順と組み合わせて各フレームの時間・周波数分解能を決定する手法を提案する．更に，各フレームの SNR を判定基準とした VAD を用いて有音区間と無音区間を判定し，それぞれの区間でより適切な時間・周波数分解能が得られるように処理を分ける．また，提案法を用いることにより増加した計算量を低減するために DFT ベースの変換法を検討する．

以下，**2.** で Gülzow らが提案したマルチレートシステムを用いた SS 法を簡単に説明する．**3.** では狭帯域雑音に対する認識率低下の原因を示し，その改良法を提案する．また，提案法を用いた場合の孤立単語音声認識実験の結果も示す．**4.** では提案法を用いることにより増加した計算量の低減方法について説明する．

2. マルチレートシステムを用いた SS 法

2.1 概　　要

SS は，観測信号に一様な時間・周波数分解能をもたせ，周波数軸上で等間隔に配置された成分に対して行われることが多い．これに対し，Gülzow らは観測信号に非一様な時間・周波数分解能をもたせた場合を考え，様々なパターンに対して実験を行った．その結果，図 1 に示すような観測信号と推定雑音の SNR が高い帯域ほど周波数分解能が細かく，SNR が低くなるにしたがって分解能が粗くなるように帯域分割をしたときの時間・周波数分解能が最も効果的に強調を行えることが分かった [6]．これは音声スペクトルの概形が残っているような高 SNR の帯域では周波数分解能を細かくすることでその概形を維持し，音声スペクトルが雑音スペクトルに埋もれているような低 SNR の帯域では周波数分解能を粗くすることによりその帯域をまとめて処理することで雑音スペクトルの推定誤差を小さくしようとするものである．

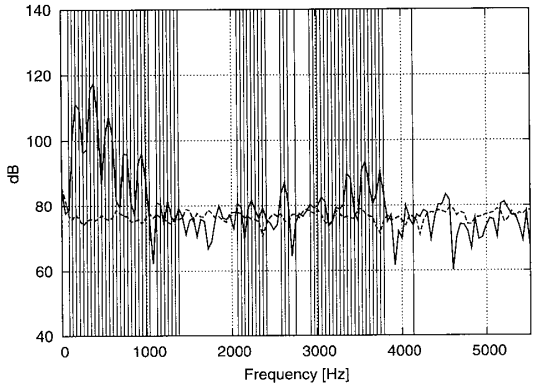


図 1　観測信号と推定雑音のスペクトル：観測信号（実線），推定雑音（破線），周波数分解能（縦線）

Fig. 1　Observation signal spectrum (solid line), estimated noise spectrum (dashed line) and frequency resolution (vertical line).

マルチレートシステムを用いた SS 法の処理手順を図 2 のブロック図を用いて説明する．はじめに，観測信号 $x_p(l)$ を分析フィルタバンクを用いて Q 個のチャネル信号 $x_{p,q}(l_q)$ $(q = 0, 1, \cdots, Q-1)$ に分割する．ここで，p はフレーム番号を表す．また，$x_p(l)$ は音声信号 $s_p(l)$ とそれと無相関で定常的なスペクトル特性をもった雑音信号 $n_p(l)$ により

$$x_p(l) = s_p(l) + n_p(l) \qquad l = 0, 1, \cdots, L \qquad (1)$$

のように表されると仮定する．次に，分割されたチャネル信号に対し以下の式を用いて SS を行う．

$$\hat{s}_{p,q}(l_q) = w_{p,q}(l_q) x_{p,q}(l_q) \qquad (2)$$

ここで，$\hat{s}_{p,q}(l_q)$ は強調音声信号のチャネル信号を表し，$w_{p,q}(l_q)$ は次式で与えられる重み係数を表す．

$$w_{p,q}(l_q) = \frac{\sqrt{\pi}}{2} \sqrt{\left(\frac{1}{1+R^{(\mathrm{post})}_{p,q}(l_q)}\right)\left(\frac{R^{(\mathrm{prio})}_{p,q}(l_q)}{1+R^{(\mathrm{prio})}_{p,q}(l_q)}\right)} \times M\left[\left(1+R^{(\mathrm{post})}_{p,q}(l_q)\right)\left(\frac{R^{(\mathrm{prio})}_{p,q}(l_q)}{1+R^{(\mathrm{prio})}_{p,q}(l_q)}\right)\right] \qquad (3)$$

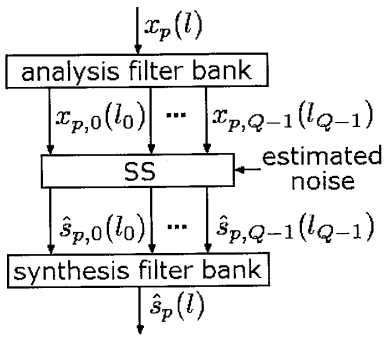


図 2　マルチレートシステムを用いた SS 法のブロック図

Fig. 2　Block diagram of the SS using multirate systems.

式 (3) の $R_{p,q}^{(\mathrm{post})}(l_q)$ と $R_{p,q}^{(\mathrm{prio})}(l_q)$ は以下のように定義され，

$$R_{p,q}^{(\mathrm{post})}(l_q) = \frac{\{x_{p,q}(l_q)\}^2}{\{\hat{n}_q(l_q)\}^2} - 1 \tag{4}$$

$$R_{p,q}^{(\mathrm{prio})}(l_q) = (1-\beta)\max\left(R_{p,q}^{(\mathrm{post})}(l_q), 0\right) + \beta\frac{\{w_{p,q}(l_q-1)x_{p,q}(l_q-1)\}^2}{\{\hat{n}_q(l_q-1)\}^2} \tag{5}$$

関数 $M[\cdot]$ は，

$$M[u] = \exp\left(-\frac{u}{2}\right)\left[(1+u)I_0\left(\frac{u}{2}\right) + uI_1\left(\frac{u}{2}\right)\right] \tag{6}$$

のように定義される．ここで，$\{\hat{n}_q(l_q)\}^2$ はサブバンド q の推定した雑音を表し，$I_0(\cdot)$，$I_1(\cdot)$ はそれぞれ 0 次，1 次の第 1 種変形ベッセル関数を表す．最後に，合成フィルタバンクを用いてチャネル信号 $\hat{s}_{p,q}(l_q)$ を強調音声信号 $\hat{s}_p(l)$ に合成する．

マルチレートシステムを実現するためのスペクトル分析・合成部の構成は様々なものが考えられるが，比較的任意の時間・周波数分解能が得られる QMF を用いたフィルタバンクを使用する．

2.2 QMF を用いたフィルタバンク

QMF を用いた分析フィルタバンクは，高域通過フィルタ $H(z)$ とそれと鏡像関係をもつ低域通過フィルタ $L(z)$，それらに続くダウンサンプラからなる 2 分割 QMF フィルタバンクのトリー構造によって構成され，図 3 のようになる．ここで，$y_q^s(m_s)$ $(m_s = 0, 1, \cdots, 2^{S-s}-1)$ はステージ s におけるサブバンド q のチャネル信号を表す．また，S は全ステージ数を表す．合成フィルタバンクも同様に，アップサンプラとそれに続くフィルタ $\tilde{H}(z) = -H(z)$ と $\tilde{L}(z) = L(z)$ よりなるフィルタバンクのトリー構造で構成される．

このフィルタバンクを用いて，観測信号 $x_p(l)$ を分析すると，各ステージで異なる帯域幅のチャネル信号が生成される．これらのチャネル信号を帯域が重ならないように各ステージから選択し，フィルタバンクからの出力とすることで任意の時間・周波数分解能を得ることができる．

2.3 SNR に応じた時間・周波数分解能の決定

QMF を用いたフィルタバンクから得られるチャネル信号を観測信号と推定雑音の SNR に応じて選択する必要がある．具体的な手順として「分割」による帯域分割手順が Gülzow らにより提案されている．

2.3.1 「分割」による帯域分割手順

「分割」による帯域分割は以下の手順で行う．まず，周波数分解能が最も粗いステージ 0 から処理を始め，チャネル信号 $y_0^0(m_0)$ に対する SNR を計算し，しきい値 η との比較を行う．このとき，チャネル信号 $y_q^s(m_s)$ に対する SNR は次式を用いて計算する．

$$\mathrm{SNR} = \frac{\sum_{m_s=0}^{2^{S-s}-1}\{y_q^s(m_s)\}^2}{\sum_{m_s=0}^{2^{S-s}-1}\{\hat{n}_q^s(m_s)\}^2} \tag{7}$$

ここで，$\{\hat{n}_q^s(m_s)\}^2$ はステージ s におけるサブバンド q の推定した雑音を表す．次に，SNR が η 以上であるなら $y_0^0(m_0)$ を $y_0^1(m_1)$ と $y_1^1(m_1)$ へ分割し，η より小さいなら分割を終了して $y_0^0(m_0)$ をフィルタバンクからの出力として選択する．最後に，分割が行われた場合はそれぞれのチャネル信号に対して，途中のステージで分割が終了するか，最終ステージに到達するまでこの手順を繰り返し適用する．

この手順を適用したときの例を図 4 に示す．図 4

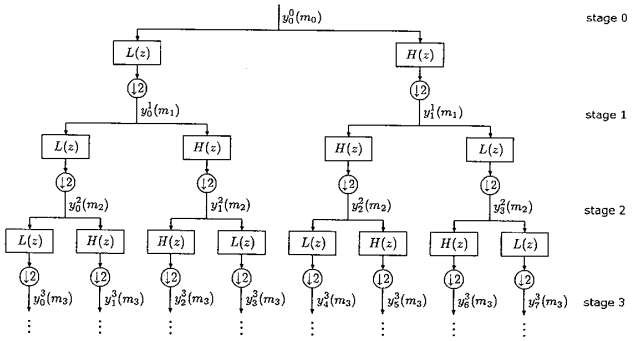


図 3 QMF を用いた分析フィルタバンク
Fig. 3 Analysis filter bank using QMF.

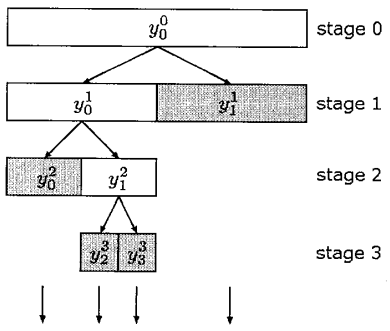


図 4 「分割」による帯域分割手順 ($L=8$, $S=3$ の場合)
Fig. 4 Band splitting procedure (e.g., $L = 8$, $S = 3$).

は分割を行った結果，チャネル信号 $y_0^2(m_2)$，$y_2^3(m_3)$，$y_3^3(m_3)$，$y_1^1(m_1)$ がフィルタバンクからの出力として選択されたことを表している．

3. 認識率低下の原因とその改良法

Gülzow らの手法 [6] を用いると認識率がどのように変化するのかを調べるため認識実験を行った．実験条件は **3.3.1** のとおりである．図 15 を見ると，広帯域雑音である白色雑音と工場雑音に対してはマルチレートシステムを用いることで SS のみのときより認識率が大きく向上しているが，狭帯域雑音である走行自動車内雑音に対してはベースラインよりも低下していることが分かる．この原因を調べるために白色雑音と走行自動車内雑音に対する強調処理結果の比較を行う．

3.1 問題点 1

観測信号の各フレームに「分割」による帯域分割を行うといくつの帯域に分割されるのかを図 5 に示す．音声波形は発話 /wakou/ の /wa/ の部分を表す．また，図 5 で用いた音声のスペクトログラムを図 6，図 7 に示す．音声スペクトルの雑音への埋もれ具合が同程度となるもので比較を行ったので両者の SNR が異なっていることに注意せよ．

白色雑音が付加された場合，有音区間では帯域が分割されているが無音区間では分割されていない．一方で，走行自動車内雑音が付加された場合では，有音・無音にかかわらず突発的に帯域が分割されている．有音区間または無音区間で帯域分割される場合とされない場合に観測信号のスペクトルがどのように処理されるのかを考える．

走行自動車内雑音が付加された音声の無音区間で帯域分割されない場合は，強調処理後のスペクトルのパワーがすべての帯域で減衰され，雑音なしのスペクトルとよく一致している．しかし，図 8 に示すように帯域分割される場合を示す 0.23 秒付近では，周波数分解能が粗い帯域でパワーが比較的減衰されているものの，周波数分解能が細かい帯域で余り減衰されず，強調処理後のスペクトル全体の概形が大きく乱れている．無音区間では帯域全体にわたって SNR が低くなっているので，周波数分解能が粗くなるように分割されるべきである．白色雑音の場合は無音区間の大部分で適切な時間・周波数分解能を得ることができている．しかし，走行自動車内雑音の場合は図 8 のように適切な時間・周波数分解能が得られていない．

次に，走行自動車内雑音が付加された音声の有音区間で帯域分割されない場合を示す 0.42 秒付近（図 9）

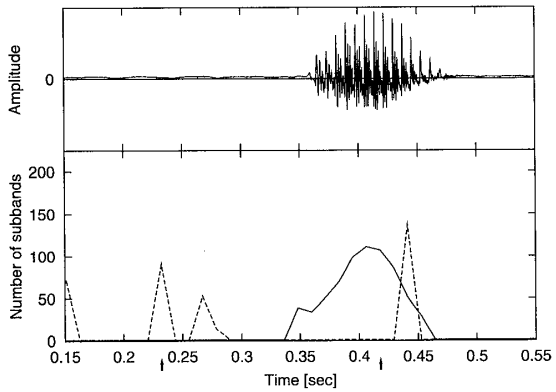


図 5 音声波形（雑音なし，上段）と帯域数の時間変化（雑音あり，下段）：白色雑音の場合（SNR 0 dB，実線）と走行自動車内雑音の場合（SNR −15 dB，破線）

Fig. 5 Speech waveform (clean, top) and the number of subbands (noisy, bottom). White noise (SNR 0 dB) : solid line, car noise (SNR −15 dB) : dashed line.

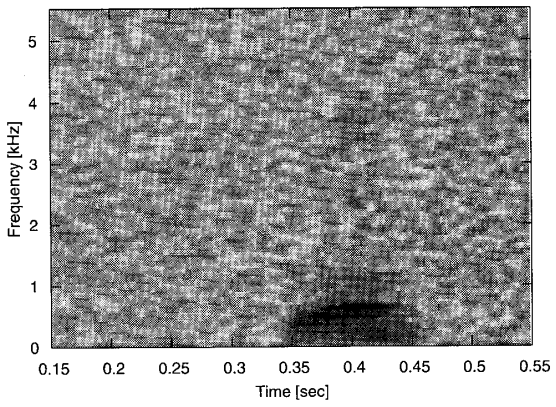


図 6 白色雑音の場合のスペクトログラム（SNR 0 dB）

Fig. 6 Spectrogram under white noise environment (SNR 0 dB).

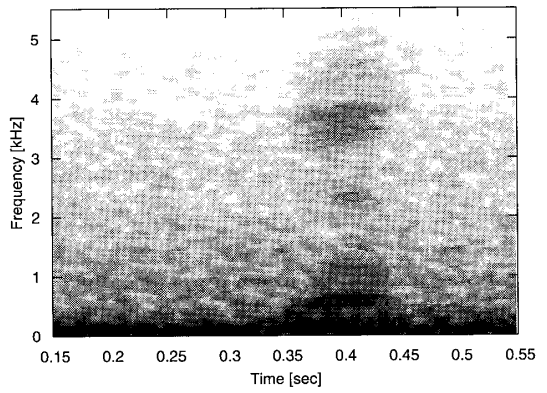


図 7 走行自動車内雑音の場合のスペクトログラム（SNR −15 dB）

Fig. 7 Spectrogram under car noise environment (SNR −15 dB).

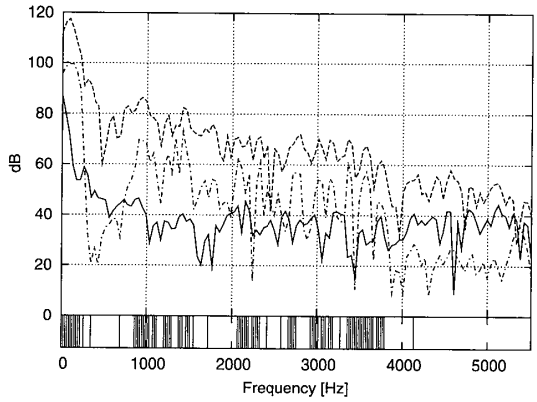


図 8 約 0.23 秒の時刻でのスペクトル（上段）と周波数分解能（下段）：雑音なし音声のスペクトル（実線），走行自動車内雑音が付加された音声の強調処理前のスペクトル（SNR −15 dB，破線），強調処理後のスペクトル（鎖線）

Fig. 8 Spectrum (top) and frequency resolution (bottom) at about 0.23 seconds. Clean : solid line, car noise (SNR −15 dB) : dashed line, enhanced speech : chain line.

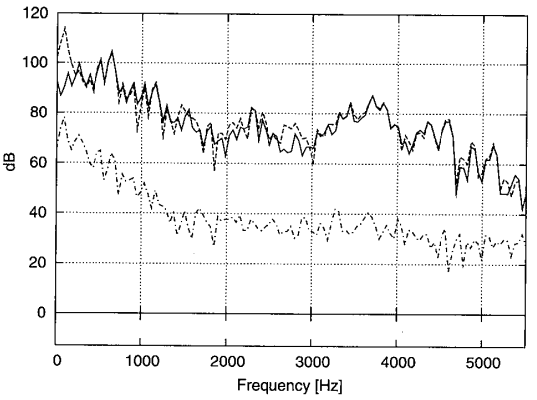


図 9 約 0.42 秒の時刻でのスペクトル（上段）と周波数分解能（下段）：雑音なし音声のスペクトル（実線），走行自動車内雑音が付加された音声の強調処理前のスペクトル（SNR −15 dB，破線），強調処理後のスペクトル（鎖線）

Fig. 9 Spectrum (top) and frequency resolution (bottom) at about 0.42 seconds. Clean : solid line, car noise (SNR −15 dB) : dashed line, enhanced speech : chain line.

では，全体的にパワーが減衰され雑音なしのスペクトルから大きくかけ離れている．有音区間では帯域ごとの SNR に従って周波数分解能の粗い部分と細かい部分の両方が存在する時間・周波数分解能が得られるべきであるが，図 9 では帯域が全く分割されていないために適切な時間・周波数分解能が得られていない．

分割手順が期待どおりに動作しない原因として，分割の判定で用いる SNR を式 (7) のようなサブバンド全体のパワーの比として求めていることが挙げられる．広帯域雑音が付加された音声の場合，あるサブバンドにおいて式 (7) による SNR が低くなったとすると雑音の影響はそのサブバンド内の帯域全体にわたって出ていると考えられるので，この SNR を用いた分割終了の判定は妥当だといえる．しかし，狭帯域雑音が付加された音声の場合には，あるサブバンドでこの SNR が低くなったとしても雑音の影響がサブバンド内の一部の帯域だけに集中している可能性があるので分割を終了すると問題が生じる場合がある．例えば，図 9 ではステージ 0 で分割が終了しているが雑音の影響で SNR が低くなっている帯域は低域の一部分だけであり，それ以外の帯域では雑音なしのスペクトルからほぼ変化しておらず SNR が高い状態になっている．このような場合には高 SNR の帯域で分割を行う必要がある．この問題点を改良するための方法としては，しきい値を下げるなどの分割の判定基準を変更することが挙げられる．しかし，判定基準の変更では，狭帯域雑音に対する改善効果が得られたとしても広帯域雑音に対して過剰に帯域分割されるなどのデメリットが生じる．そこで本論文では，狭帯域雑音に対する改善効果を示しつつ，広帯域雑音に対する過剰な分割の度合いが小さかった「併合」による帯域分割手順を導入し，従来の「分割」による帯域分割手順と組み合わせることで広帯域雑音と狭帯域雑音の両方に対して有効な改良法を提案する．

3.1.1 「併合」による帯域分割手順

「併合」による帯域分割手順は，図 4 で示した「分割」による帯域分割手順とは逆に，周波数分解能が最も細かい最終ステージから処理を始める．まず，隣り合う二つのチャネル信号 $y^S_{2q}(m_S)$，$y^S_{2q+1}(m_S)$ $(q = 0, 1, \cdots, L/2-1)$ を一組と考え，組内のそれぞれのチャネル信号に対して SNR を計算する．次に，両方の SNR がしきい値 η より小さいなら $y^S_{2q}(m_S)$ と $y^S_{2q+1}(m_S)$ を $y^{S-1}_q(m_{S-1})$ へ併合し，どちらか一方でも η 以上であるなら併合を終了して $y^S_{2q}(m_S)$ と $y^S_{2q+1}(m_S)$ をフィルタバンクからの出力として選択する．最後に，この手順をステージが 0 となるか，組となるチャネル信号がなくなるまで繰り返し適用する．この手順を適用したときの例を図 10 に示す．

この手順は，帯域内に高 SNR の部分が存在する場合にはその周波数分解能を細かいまま維持することで狭帯域雑音に対しても適切な分割を行うことができる．「併合」による帯域分割手順を用いると図 5 の帯域数の時間変化と図 9 の有音区間で帯域分割されない場合

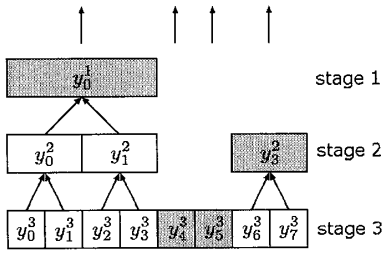


図 10 「併合」による帯域分割手順（L=8, S=3 の場合）
Fig. 10 Band merging procedure (e.g., $L=8$, $S=3$).

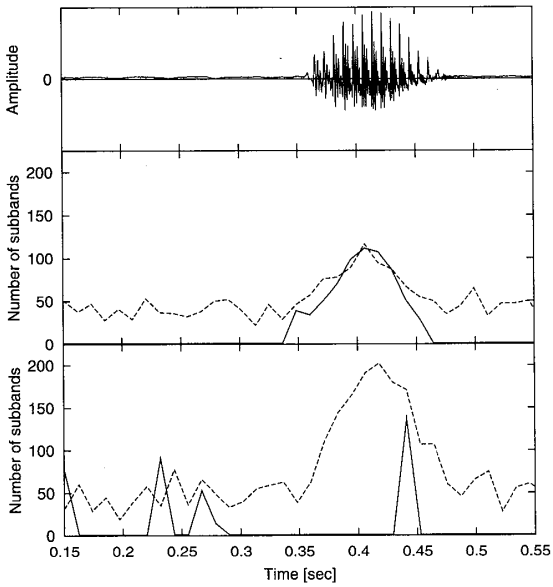


図 11 音声波形（雑音なし，上段）と白色雑音が付加された音声の帯域数の時間変化（SNR 0 dB，中段），走行自動車内雑音が付加された音声の帯域数の時間変化（SNR −15 dB，下段）：「分割」による帯域分割手順を用いたとき（実線），「併合」による帯域分割手順を用いたとき（破線）
Fig. 11 Speech waveform (clean, top) and the number of subbands (white noise with SNR of 0 dB : middle, car noise with SNR of −15 dB : bottom). Band splitting procedure : solid line, band merging procedure : dashed line.

の結果はそれぞれ図 11，図 12 のようになる．

図 11 を見ると，走行自動車内雑音が付加された音声の有音区間で分割されるようになったことが分かる．また，図 12 では図 9 と比べて高 SNR の帯域が分割されており，より適切な時間・周波数分解能が得られるようになったことが分かる．これにより強調処理後のスペクトルが雑音なしのスペクトルに一致するようになった．しかし，図 11 の白色雑音を付加した場合と走行自動車内雑音を付加した場合の両方で，無音区間でも分割されるようになっている．また，白色雑音を付加した場合の有音区間では，帯域が過剰に分割され低 SNR の帯域でスペクトルの乱れが発生する場合

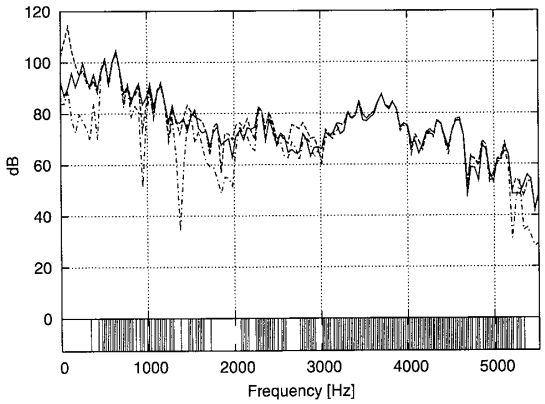


図 12 「併合」による帯域分割手順を用いたときの約 0.42 秒の時刻でのスペクトル（上段）と周波数分解能（下段）：雑音なし音声のスペクトル（実線），走行自動車内雑音が付加された音声の強調処理前のスペクトル（SNR −15 dB，破線），強調処理後のスペクトル（鎖線）
Fig. 12 Spectrum (top) and frequency resolution (bottom) at about 0.42 seconds when the band merging procedure is used. Clean : solid line, car noise (SNR −15 dB) : dashed line, enhanced speech : chain line.

があった．以上のことから，「併合」による帯域分割手順は狭帯域雑音に対する問題点への改善効果を示すが，デメリットもあることが分かる．そこで「併合」による帯域分割手順と従来の「分割」による帯域分割手順を組み合わせて互いに補い合うように用いることを考える．「分割」による帯域分割手順と「併合」による帯域分割手順を組み合わせる際に，フレームが有音区間と無音区間のどちらに属しているかにより処理を分ける．そのために VAD（Voice Activity Detection）を用いて有音区間と無音区間の判定を行う．

3.1.2 VAD

VAD を行うための判定尺度として，フレーム内の各周波数成分ごとの SNR の平均値を表す次式を用いる．

$$E = \frac{1}{L}\sum_{q=0}^{L-1}\left[\frac{\{y_q^S(0)\}^2}{\{\hat{n}_q^S(0)\}^2} - 1\right] \tag{8}$$

ここで，$\{\hat{n}_q^S(0)\}^2$ は最終ステージ S におけるサブバンド q の推定した雑音を表す．

有音・無音の判定は E としきい値 λ との比較で行い，$E \geq \lambda$ であるなら有音区間と判定し，$E < \lambda$ であるなら無音区間と判定する．更に，3 フレーム分の判定結果を用いて，有音フレームに挟まれた無音フレームを有音フレームに，無音フレームに挟まれた有音フレームを無音フレームに変更する訂正処理を行う．

3.1.3 有音区間・無音区間での処理

有音区間ではVADの判定尺度である式(8)のEを用いて「分割」と「併合」のどちらの帯域分割手順で得られた時間・周波数分解能を用いるか決定する．もし$E \geq \gamma$であるならば，そのフレームはSNRが比較的高い状態だと判断できるので音声スペクトルが雑音スペクトルに埋もれていない部分が多いと考えられる．したがって,「分割」と「併合」による帯域分割手順のうち帯域数が多い方の時間・周波数分解能を用いる．逆に$E < \gamma$であるならば，音声スペクトルが雑音スペクトルに埋もれている部分が多いと考えられるので過剰な帯域分割をしない「分割」による帯域分割手順で得られた時間・周波数分解能を用いる．無音区間では帯域全体のパワーを減衰したいので，分割を行わず一つの帯域とする．

3.2 問題点2

3.3 の認識実験では特徴パラメータとしてΔ対数パワーや$\Delta\Delta$対数パワーを用いており，それらを求めるために線形予測分析で得られる残差信号のパワーを使用している．強調処理後の各フレームの残差信号パワーの時間変化を図13に示す．音声波形は図5と同じものである．白色雑音が付加された音声の軌跡は，強調処理を行うことで雑音なしの軌跡に近くなっている．一方で，走行自動車内雑音が付加された音声の軌跡は強調処理を行うことで非常に乱れている．有音区間での乱れは **3.1** で説明した帯域分割手順による問題が原因である．無音区間の乱れでも帯域分割手順による問題が原因に含まれるが，雑音自体の影響も存在する．本論文では雑音が定常であることを仮定しているが，走行自動車内雑音は白色雑音に比べてパワーの変動が若干大きく，強調処理を行ったときにその変動が増幅されていると考えられる．したがって，改良した帯域分割手順を用いても無音区間の一部にはパワーの乱れが残ってしまう．そこで無音区間での残差信号パワーが一定となるように調整を行う．具体的には無音区間と判定されたフレームの残差信号パワーをあらかじめ設定しておいた定数値に置き換えることにより調整を行う．

3.3 孤立単語音声認識実験1

帯域分割手順の改良とパワー調整を組み込んだ提案法の有効性を確認するために孤立単語音声認識実験を行った．

3.3.1 実験条件

音声信号として電子協日本語共通音声データより地名100単語を量子化ビット数16ビット，サンプリング周波数11025 Hzとして用い，雑音信号として白色雑音と電子協騒音データベースからの工場雑音と走行自動車内雑音を用いた．雑音を付加した音声の聞き取り難さが同程度となるようにした結果，走行自動車内雑音に対してSNRを20 dB低く設定した．更に，狭帯域雑音に対する提案法の有効性を確認するため，図14に示すような狭帯域雑音を作成し実験に用いた．分析条件としてフレーム長Lを256サンプル，フレームシフトΔlを128サンプルとし，窓関数にはハニング窓を用いた．また，伝達関数$1-0.97z^{-1}$を用いプリエンファシスを行った．推定雑音スペクトルは観測信号の先頭から10フレーム分を平均化して求めた．

音声認識には32状態1混合のHMM [7] を用い，特徴ベクトルをLPCメルケプストラム12次元，ΔLPCメルケプストラム12次元，$\Delta\Delta$LPCメルケプストラム12次元，残差信号のΔ対数パワー1次元，$\Delta\Delta$対数パワー1次元の計38次元とした．学習には男性話者40人一人当り3回発話分の雑音を付加していない音声を用い，評価には学習に用いなかった別の男性話

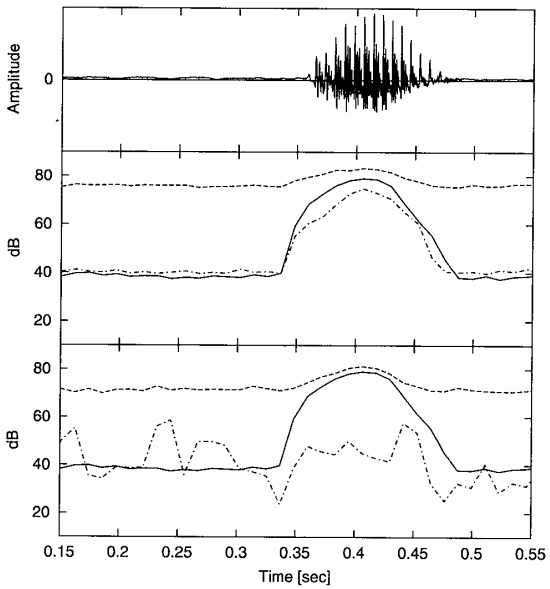


図13 音声波形（雑音なし，上段）と白色雑音が付加された音声の残差信号パワー（SNR 0 dB，中段），走行自動車内雑音が付加された音声の残差信号パワー（SNR −15 dB，下段）：雑音なしの音声（実線），強調処理前（破線），強調処理後（鎖線）

Fig. 13 Speech waveform (clean, top) and the LP-residual signal power of noisy speech (white noise with SNR of 0 dB : middle, car noise with SNR of −15 dB : bottom). Clean : solid line, before processing : dashed line, after processing : chain line.

者 10 人一人当り 1 回発話分を用いた．

従来の帯域分割手順におけるしきい値 η を 1.259 とし，提案した帯域分割手順における「併合」による手順のためのしきい値 η_m を 6.310,「分割」による手順のためのしきい値 η_s を 1.413，VAD のためのしきい値 λ を 0.447，有音区間における時間・周波数分解能を選択するためのしきい値 γ を 1.259 とした．しきい値は，白色雑音，工場雑音，走行自動車内雑音の 3 種類の雑音に対する認識率が全体的に良くなるように経験的に定めた．従来法におけるしきい値 η に関して，η の値を小さくすると走行自動車内雑音に対する有音区間の周波数分解能を上げることが可能だが，同時に無音区間の周波数分解能が過剰となるため全体として認識率が向上しない．η の値を更に小さくすると SS のみの場合と同等の処理になるので，その場合と同程度に認識率が向上するがマルチレートシステムを用いている意義がなくなる．また，η を小さな値にすると白色雑音と工場雑音に対しては周波数分解能が過剰となるので認識率が向上しなくなる．したがって，白色雑音と工場雑音に対してマルチレートシステムを用いた SS 法の効果が現れる範囲内で 3 種類の雑音に対する認識率が全体的に良くなるように η を定めた．

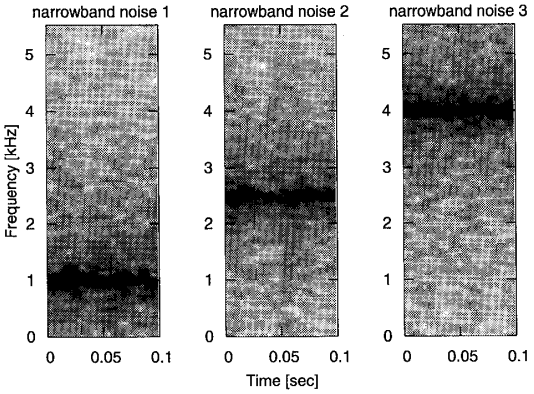


図 14　狭帯域雑音 1（左），狭帯域雑音 2（中央），狭帯域雑音 3（右）のスペクトログラム

Fig. 14　Noise spectrogram (narrowband noise 1 : left, narrowband noise 2 : center, narrowband noise 3 : right).

3.3.2　認識結果

白色雑音，工場雑音，走行自動車内雑音に対する認識結果を図 15 に示す．認識は，前処理を行っていないベースラインとなるもの，前処理として SS [4]（パワー調整なし）を組み込んだもの，SS にマルチレートシステム [6] を用いたもの，そして提案法を用いたもので行った．また，提案法におけるパワー調整の有効性を確認するため，パワー調整を行わない場合の結果も示す．

白色雑音に対してはマルチレートシステムを用いることで SS のみのときよりも低 SNR の条件下で認識率が特に大きく向上している．また，提案法を用いても従来と同程度の認識率を示している．工場雑音に対しても同様な結果が得られた．走行自動車内雑音に対しては，SS のみのときの認識率はベースラインよりも向上しているが，マルチレートシステムを用いると大きく低下している．これは提案法を用いることで SS のみの場合と同程度の認識率に改善できた．パワー調整の有無による認識率の違いを見ると，白色雑音と工場雑音の場合ではパワーの乱れが生じないのでほぼ変化していない．走行自動車内雑音の場合では調整を行わないときに認識率が低下しておりパワー調整の効果が示されている．

3 種類の狭帯域雑音に対する認識結果を図 16 に示す．狭帯域雑音 2 に対しては SS のみを用いた場合に

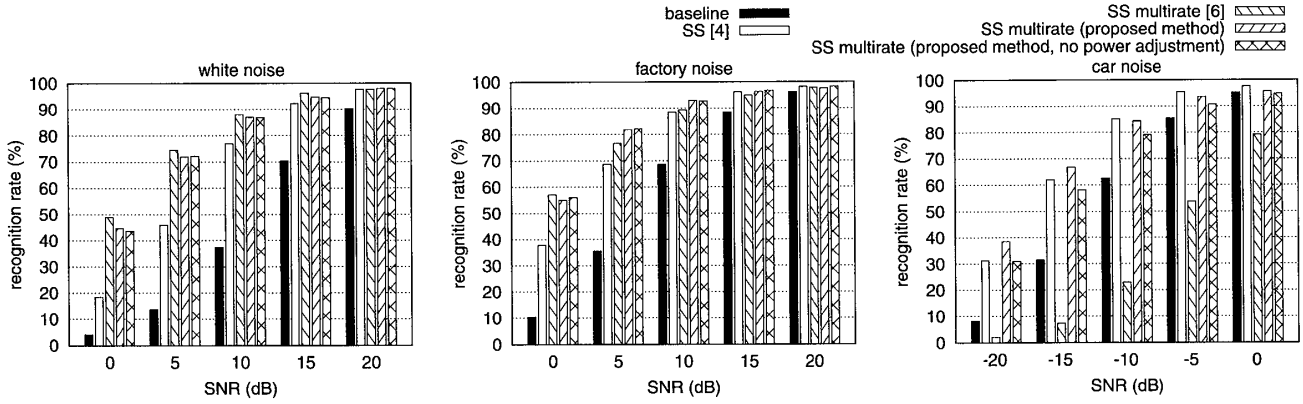


図 15　白色雑音（左），工場雑音（中央），走行自動車内雑音（右）に対する認識率

Fig. 15　Recognition rate under noise environment (white noise : left, factory noise : center, car noise : right).

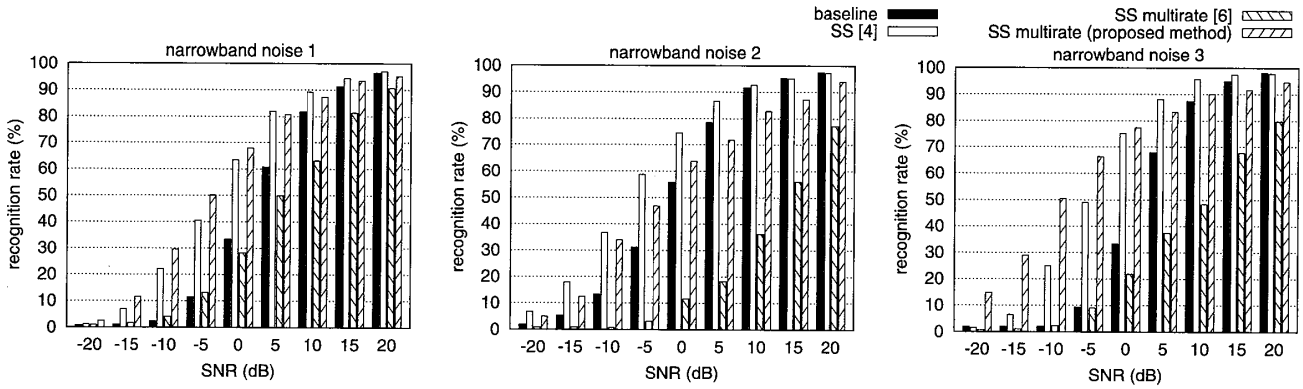


図 16 狭帯域雑音 1（左），狭帯域雑音 2（中央），狭帯域雑音 3（右）に対する認識率
Fig. 16 Recognition rate under noise environment (narrowband noise 1 : left, narrowband noise 2 : center, narrowband noise 3 : right).

比べて提案法を用いた場合の認識率が低いが，3 種類の狭帯域雑音全体としては走行自動車内雑音の場合と同様の傾向を示している．このことから，提案法は狭帯域雑音全般に対して有効であるといえる．

4. 提案した改良法の計算量低減

提案法には従来法に比べて計算量が増加してしまう問題がある．主な要因は，「併合」による帯域分割手順で必ず最終ステージのチャネル信号を必要とすることである．なぜなら，従来の「分割」による帯域分割手順では SNR による判定によっては最終ステージに到達する前に分割が終了するので，決定した時間・周波数分解能に対して無駄な分割を行わなかった．しかし，「併合」による帯域分割手順では最終ステージから処理を始めるので，決定した時間・周波数分解能にかかわらずすべての分割を行う必要がある．そのため，分割を実行する回数が増え計算量が増加する．

Gülzow らは文献 [6] において DFT ベースの手法も検討していた．しかし，SNR に応じた帯域分割を行わずに時間・周波数分解能を固定にしたものや SNR に応じた帯域分割を行うが時間分解能をもたせずに処理を行うものだった．後者の手法は少ない計算量で良好な音声強調を行えるが，音声認識システムの前処理として用いると QFM の場合に比べて認識率の改善具合が小さかった．そこで本論文では，前者の手法に SNR に応じた帯域分割を適用することで，認識率に大きな影響を与えずに QMF の場合に問題となる「併合」の導入による計算量の増加を抑えることを行った．以下，計算量低減のための手法を解説する．

4.1 DFT ベースの変換を用いたスペクトル分析部

DFT ベースの変換を用いた手法では，QMF を用いて構成したスペクトル分析・合成部を DFT ベースの変換を用いたものに変更する．DFT ベースの変換を用いたスペクトル分析では，まず，観測信号を DFT を用いて変換し，周波数軸を任意に分割する．そして，それぞれのサブバンドを逆変換することでチャネル信号の生成を行う．図 17 はオクターブ分析となるように周波数軸を分割してチャネル信号を生成した場合の例である．スペクトル合成も同様の考えで逆向きの処理を行えばよい．実信号をフーリエ変換するとそのスペクトルは折返し周波数で対称となるので帯域分割やチャネル信号の生成などは折返し周波数までの半分の帯域だけに行う．このように DFT ベースの変換を用いても QMF を用いたときのような任意の時間・周波数分解能を得ることができる．

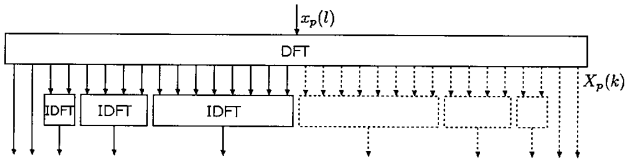


図 17 DFT ベースの変換を用いたスペクトル分析部（オクターブ分析の場合）
Fig. 17 Spectrum analyzer using the DFT-based procedure (e.g., octave analysis).

4.2 DFT ベースの変換のための SNR に応じた時間・周波数分解能の決定

DFT ベースの変換を用いた場合でも時間・周波数分解能の決定には，「分割」と「併合」による帯域分割手順を組み合わせた提案法を用いる．帯域分割の手順は基本的に同じであるが，分割操作の対象となるデータがチャネル信号 $y_q^s(m_s)$ から観測信号 $x_p(l)$ のフーリエ変換 $X_p(k)$ に変わる．したがって，式 (7) で表さ

れる SNR の計算式を

$$\mathrm{SNR} = \frac{\sum_{k=k_l}^{k_u} |X_p(k)|^2}{\sum_{k=k_l}^{k_u} |\hat{N}(k)|^2} \tag{9}$$

に変更する．また，式 (8) で表される VAD の判定尺度 E も

$$E = \frac{2}{L} \sum_{k=0}^{L/2-1} \left[\frac{|X_p(k)|^2}{|\hat{N}(k)|^2} - 1 \right] \tag{10}$$

に変更する．ここで，$|\hat{N}(k)|^2$ は推定した雑音スペクトルを表し，k_l，k_u は分割対象となる帯域の上限と下限の周波数値に対応するインデックスを表す．

帯域が分割される場合，QMF を用いた変換ではチャネル信号を実際に二つの信号に分割したが，DFT ベースの変換では周波数軸上に分割点を示す印を入れるだけとなる．更に，$X_p(k)$ はそれ自身が既に最大の周波数分解能をもっているので，そのまま「併合」による帯域分割手順に用いることができる．提案法において計算量が増加する主な要因は，「併合」による帯域分割手順のために最終ステージまで処理しなければならないことだというのは先に述べたが，DFT ベースの変換では $X_p(k)$ を用いることによりそれが必要なくなり，決定した周波数軸上の印に従って必要最小限のチャネル信号を生成するだけで良くなる．これにより「併合」による帯域分割手順を導入するための計算量を大きく低減できる．

また，折返し周波数までの半分のデータを処理するだけでよいことも計算量の低減につながる．しかし，QMF を用いた変換と DFT ベースの変換でフレーム長を同じにした場合，QMF を用いた変換では L 個のサンプルで折返し周波数までの帯域を表現しているのに対し，DFT ベースの変換では $L/2$ 個のサンプルで同じ帯域を表現することになり最大の周波数分解能が半分になる．そこで，フレーム長をそろえた場合に加えて，最大の周波数分解能をそろえた場合についても考える．最大の周波数分解能は DFT ベースの変換において各フレームの L 個のサンプルにゼロ詰めを行いフレーム長を倍の $2L$ として処理することでそろえる．

4.3 実行時間の計測

QMF を用いた変換を DFT ベースの変換に変更した場合の実行時間の変化を調べた．実験はゼロ詰めをして最大の周波数分解能をそろえた場合についても行った．

表 1 実験環境
Table 1 Experimental environment.

CPU	Intel Pentium4 630 (3 GHz,HTT)
メモリ	PC4200 DDR2-SDRAM 2048 MB
OS	Vine Linux 3.2
カーネル	2.4.31 (非 SMP なので HT は無効)
コンパイラ	gcc 3.3.2

4.3.1 実験条件

実験に用いる音声データや QMF を用いた変換のためのパラメータの値などは **3.3.1** と同じとする．DFT ベースの変換を用いたときのパラメータ値は，「併合」による帯域分割手順のためのしきい値 η_m を 3.162,「分割」による帯域分割手順のためのしきい値 η_s を 1.259，VAD のためのしきい値 λ を 0.282，有音区間における時間・周波数分解能を選択するためのしきい値 γ を 0.631 とした．また，フレーム長やサブバンド長を 2 のべき乗の長さに制限することですべての DFT/IDFT に FFT/IFFT を使用した．実装は C 言語で行い，実験環境には表 1 に示すものを用いた．実行時間は 1 単語当り 4 回発話分の計 400 個の音声データを処理したときの合計時間で計測した．計測では強調処理と特徴ベクトル生成までの処理時間を計ったので認識のための処理時間は含まれていない．

4.3.2 計測結果

白色雑音，工場雑音，走行自動車内雑音に対する計測結果を図 18 に示す．QMF を用いた変換で従来法を用いたものと提案法を用いたものを比較すると，すべての場合において提案法を用いた方の実行時間が大きく増加している．このとき増加量は 80 ～ 130%だった．次に，QMF を用いた変換と DFT ベースの変換で従来法を用いたものを比較すると，すべての場合において DFT ベースの変換を用いた方の実行時間が減少している．このとき減少量は 39 ～ 57%だった．これは，折返し周波数までの半分の帯域だけを処理していることによる効果が表れたためである．また，DFT ベースの変換を用いたもので従来法を用いたものと提案法を用いたものを比較すると，すべての場合において提案法を用いた方の実行時間が増加している．このとき増加量は 14 ～ 24%だった．しかし，QMF を用いた変換の場合では 80 ～ 130%の増加だったので，DFT ベースの変換を用いることにより「併合」による帯域分割手順を導入するための計算量を大きく低減できて

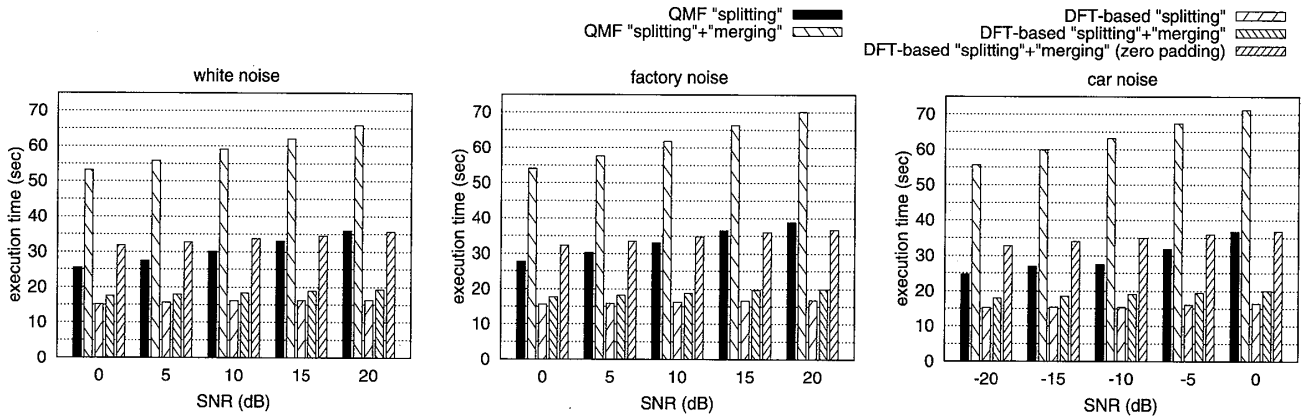


図 18 白色雑音（左），工場雑音（中央），走行自動車内雑音（右）に対する実行時間
Fig. 18 Execution time under noise environment (white noise : left, factory noise : center, car noise : right).

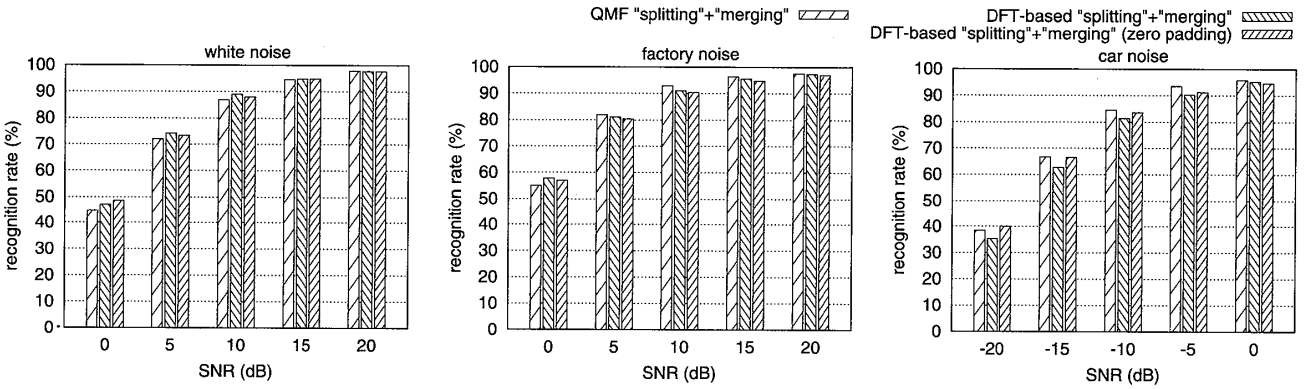


図 19 白色雑音（左），工場雑音（中央），走行自動車内雑音（右）に対する認識率
Fig. 19 Recognition rate under noise environment (white noise : left, factory noise : center, car noise : right).

いることが分かる．最後に，DFT ベースの変換で最大の周波数分解能をそろえるためにゼロ詰めした場合について考える．ゼロ詰めしたものとしていないものを比較すると，処理すべきサンプル数が増えている分ゼロ詰めした方の実行時間が増加している．このとき増加量は 81 ～ 84%だった．しかし，同じ最大の周波数分解能をもつ QMF を用いた変換の場合と比較すると，DFT ベースの変換でゼロ詰めした方の実行時間が 40 ～ 48%減少している．以上のことから，DFT ベースの変換を用いることで計算量の低減を行えることが確認できた．

4.4 孤立単語音声認識実験 2

QMF を用いた変換を DFT ベースの変換に変更したことによる認識率への影響を調べるために孤立単語音声認識実験を行った．

4.4.1 実験条件

使用する音声データやパラメータの値などの実験条件は **3.3.1**，**4.3.1** と同じである．

4.4.2 認識結果

白色雑音，工場雑音，走行自動車内雑音に対する認識結果を図 19 に示す．白色雑音や工場雑音に対しては DFT ベースの変換を用いてもほぼ同等の認識率を示している．走行自動車内雑音に対しては DFT ベースの変換を用いると認識率がわずかに低下しているものの，QMF を用いた変換と比べて大きな差は生じていない．また，ゼロ詰めを行い最大の周波数分解能を上げることで回復している．

5. むすび

本論文では，Gülzow らにより提案されたマルチレートシステムを用いた SS 法 [6] を音声認識システムの耐雑音性を向上させるための前処理として用いたとき，広帯域雑音に対しては認識率が向上するが狭帯域雑音に対しては低下することを示した．これに対し，強調処理結果を比較することで認識率低下の原因を解明し，その改良法を提案した．

提案法では，新たに導入した「併合」による帯域分割手順を従来の「分割」による帯域分割手順と組み合わせることで，狭帯域雑音に対して適切な帯域分割を行えるようにした．また，各フレームのSNRを判定基準としたVADにより有音区間と無音区間を判定し，それぞれの区間でより適切な帯域分割が行えるように処理を分けた．有音区間では各フレームのSNRを用いて「分割」と「併合」のどちらの分割手順を用いるか決定した．無音区間では帯域分割を行わず，強調処理後のフレームの残差信号パワーを一定にする調整を行った．これにより広帯域雑音に対する認識率を維持したまま，狭帯域雑音に対する認識率を向上できた．

提案法でQMFを用いた変換を使用すると帯域分割の回数が増えるために計算量が大きく増加したが，QMFを用いた変換をDFTベースの変換に変更することにより「併合」による帯域分割手順を導入するための計算量を低減した．また，孤立単語音声認識実験により変換方法の変更を行っても認識率に大きな影響を与えないことを確認した．

今後の課題としては，しきい値などの各パラメータ値を自動で決定できるようにすることが挙げられる．なぜなら，最も良い認識性能を示すパラメータ値が雑音の種類などで変化するからである．現在は実験結果に基づいて，全体的に良い認識性能を示すパラメータ値を経験的に定め，すべての雑音に対して同じ値を用いているが，これを雑音の種類やSNRに応じて決定できれば認識性能の更なる向上が期待できる．

文　献

[1] ATR国際電気通信基礎技術研究所（編），ATR先端テクノロジーシリーズ自動翻訳電話，オーム社，1994.

[2] G.M. Davis, Noise Reduction in Speech Applications, CRC Press, 2002.

[3] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, April 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109–1121, Dec. 1984.

[5] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.345–349, April 1994.

[6] T. Gülzow, T. Ludwig, and U. Heute, "Spectral-subtraction speech enhancement in multirate systems with and without non-uniform and adaptive bandwidths," Signal Process., vol.83, no.8, pp.1613–1631, Aug. 2003.

[7] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, Feb. 1989.

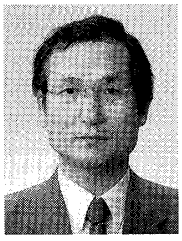
（平成19年11月26日受付，20年4月17日再受付）

今井　卓　（学生員）

2005北見工大・情報システム卒．2007同大大学院工学研究科博士前期課程了．同年同研究科博士後期課程入学，現在に至る．ディジタル信号処理の研究に従事．

中垣　淳　（正員）

1988北大・工・電子卒．1993同大大学院博士後期課程了．同年北見工業大学講師．ディジタル信号処理の研究に従事．工博．



柴田　孝次　（正員）

1969北大・工・電子卒．1971同大大学院修士課程了．同年電電公社（現NTT）．1974北見工大・電子工学科講師．1975助教授．1991同大情報システム工学科教授．工博．信号処理，ワイヤレス通信理論などの研究に従事．情報理論とその応用学会，IEEE各会員．

宮永　喜一　（正員）

1979北大・工・電子卒．1981同大大学院工学研究科修士課程了．1984米国イリノイ大学客員研究員．1987北海道大学工学部電子工学科講師．1988同助教授．現在，北大大学院情報科学研究科教授．主として，並列信号処理，並列計算機アーキテクチャ，適応信号処理の研究に従事．工博．