論　文

フレームワイズな音声検出に基づく適応フィルタを利用した自動車内でのロバスト音声認識

庄境　誠[†]　中村　哲[†]　鹿野　清宏[†]

A Robust Speech Recognition Using Adaptive Filter Based on Frame-Wise Voice Activity Detection in Car Environments

Makoto SHOZAKAI[†], Satoshi NAKAMURA[†], and Kiyohiro SHIKANO[†]

あらまし　加法性雑音が存在する環境下では，音声認識性能の大幅な劣化が起こる．本論文は，音源未知の加法性雑音に埋もれた，音源既知の加法性雑音を除去するアルゴリズムとして，フレームワイズの音声検出を利用した適応フィルタアルゴリズム NLMS-VAD を提案する．次に，音源既知および音源未知の加法性雑音，乗法性ひずみが存在する実環境におけるロバストな音声認識手法として，既提案の E-CMN/CSS 法に NLMS-VAD 法を組み合わせる方法を提案する．最後に，この提案法の性能を自動車内の不特定話者，大語い，単語認識タスクで評価し，スピーカ出力に起因する誤認識率がたかだか 2%程度に抑えられることを示す．

キーワード　音声認識，加法性雑音，乗法性ひずみ，適応フィルタ

1. まえがき

カーナビゲーションシステム（以後，カーナビと略す）は，将来自動車におけるマルチメディア情報端末に進化すると期待され，カーナビに対し，道路交通情報を提供するサービスも既に始まっている．カーナビのルート案内機能を利用するために目的地を登録する場合やカーナビを介して走行中に道路交通情報を動的に利用する場合の安全で快適なヒューマンマシンインタフェースとして音声認識技術のニーズが高い．

音声認識技術の実用化が広く進むために解決されるべき課題の一つとして，加法性雑音に対するロバスト性の向上が挙げられる．加法性雑音には，音源未知の場合と音源既知の場合がある．走行中の自動車内での，音源未知の加法性雑音としては，ロードノイズと呼ばれる道路面とタイヤとの摩擦音，風切り音，対向車との擦れ違い音，道路の継ぎ目やマンホールの蓋の上などで発生する揺動音（Bump 音），クラクション音，車内の人間の会話音声などが考えられる．音源未知の場合，ある程度定常的な加法性雑音に対しては，スペクトル減算法 [1], [2] や最小平均 2 乗誤差推定法 [3], [4] などの方法が提案されている．筆者らも，連続スペクトル減算法 [5] がスペクトル減算法や最小平均 2 乗誤差推定法よりもよりロバストな音響パラメータを生成することを見出した [6].

一方，自動車内では，カーオーディオシステム（以後，カーオーディオと略す）を利用する形態が既に一般化しているが，カーナビを利用する際もカーオーディオと同時に利用する形態が一般的であると言われている．従って，マイクの入力信号に混入した，カーオーディオからのスピーカ出力音が，非定常の加法性雑音として作用し，音声認識性能の大幅な性能低下を引き起こすことは想像にかたくない．スピーカ出力音の原信号をカーオーディオから直接取り出せる場合，マイク入力に回り込むスピーカ出力音は音源既知の加法性雑音ととらえることができる．また，カーナビのルートガイダンス音声や道路交通情報サービスのガイダンスの音声も同様に音源既知の加法性雑音とみなせる．音源既知の加法性雑音に対しては，適応フィルタを用いた雑音除去法が適用できる．これまでにも，音声応答システムのガイダンス音声出力中に発声された音声を認識できる機能，いわゆる Barge-In（Talk-Through）機能の実現に向けて，適応フィルタを用いる手法が試みられている [7], [8]．しかしながら，音源未知の加法性雑音と音源既知の加法性雑音が重畳

† 奈良先端科学技術大学院大学情報科学研究科，生駒市
Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630-0101 Japan

する環境における，ロバストな音声認識に関する研究はあまり多くない．

本論文は，音源既知および音源未知の加法性雑音に加え，乗法性ひずみが存在する環境におけるロバストな音声認識手法について論じる．まず，**2.** で，自動車内における音声認識システムのマイク入力信号のモデル化を線形スペクトル領域で行う．次に，**3.** で適応フィルタアルゴリズムの中でよく用いられるNLMS（Normalized Least Mean Square error）法（学習同定法とも呼ばれる）について概観する．**4.** では，音源未知の加法性雑音が存在する場合の適応フィルタの問題点を明らかにし，その解決方法として，フレームワイズにエネルギー，スペクトルの定常性およびピッチ性を用いて音声/非音声の検出を行うVAD（Voice Activity Detection）を利用したアルゴリズムNLMS-VAD（NLMS with frame-wise VAD）法を提案する．更に，**5.** では，音源未知の加法性雑音が存在する環境での音声認識に有効な連続スペクトル減算法CSS（Continuous Spectral Subtraction）および乗法性ひずみに有効な話者依存の音声/非音声分離型のケプストラム平均正規化法E-CMN（Exact Cepstrum Mean Normalization）を組み合わせる手法E-CMN/CSS法[6]について述べる．そして，この手法の前処理として，NLMS-VAD法を加えた，ロバストな音声認識手法を提案する．最後に，**6.** で，この提案法の性能を自動車内の不特定話者，大語い，単語認識タスクで評価する．

2. 自動車内の環境雑音のモデル化

個人の発声器官で生成される，時刻 t における周波数 ω での短時間スペクトル $S(\omega;\ t)$ の音声フレームにおける長時間平均を話者の個人性 $H_{Person}(\omega)$ と呼び，

$$H_{Person}(\omega)=\frac{1}{T}\cdot\sum_{t=1}^{T}S(\omega;\ t) \qquad (1)$$

と定義する．ここで，T は十分大きな自然数である．また，$S(\omega;\ t)$ を $H_{Person}(\omega)$ で除したもの

$$S^*(\omega;\ t)=S(\omega;\ t)/H_{Person}(\omega) \qquad (2)$$

を正規化音声スペクトルと定義する．このとき，音声スペクトル $S(\omega;\ t)$ は，正規化音声スペクトル $S^*(\omega;\ t)$ に乗法性ひずみ $H_{Person}(\omega)$ が重畳して生成されると解釈することができる．

$$S(\omega;\ t)=H_{Person}(\omega)\cdot S^*(\omega;\ t) \qquad (3)$$

一般に，時刻 t における周波数 ω での観測スペクトル $O(\omega;\ t)$ は，次式でモデル化できる．

$$H^*(\omega)=H_{Mic}(\omega)\cdot H_{Trans}(\omega)\cdot H_{Style(N)}(\omega)\cdot H_{Person}(\omega) \qquad (4)$$

$$\tilde{N}(\omega;\ t)=H_{Mic}(\omega)\cdot N(\omega;\ t) \qquad (5)$$

$$\tilde{E}(\omega;\ t)=H_{Mic}(\omega)\cdot E(\omega;\ t) \qquad (6)$$

$$O(\omega;\ t)=H^*(\omega)\cdot S^*(\omega;\ t)+\tilde{N}(\omega;\ t)+\tilde{E}(\omega;\ t) \qquad (7)$$

ここで，$N(\omega;\ t)$, $E(\omega;\ t)$ はそれぞれマイクに入力される直前の音源未知の加法性雑音のスペクトル，音源既知の加法性雑音のスペクトルを表す．また，$H_{Style(N)}(\omega)$, $H_{Trans}(\omega)$, $H_{Mic}(\omega)$ は，それぞれ加法性雑音に依存する発話様式（発話速度，発話の大きさ，Lombard効果など）に固有の周波数伝達特性，口からマイクまでの空間的な周波数伝達特性，マイクなどの入力系の電気的な周波数伝達特性を表す[9],[10].

3. 学習同定法（NLMS）による適応フィルタ

スピーカからの出力音響の直接音やダッシュボードや窓ガラスなどで反射した反射音は常時音声認識用のマイクに回り込む．ここでは，スピーカからマイクへの直接音および反射音をまとめて回り込み音と呼ぶことにする．また，スピーカ出力音から回り込み音が生成される経路を回り込み音生成経路と呼ぶことにする．一般に，回り込み音生成経路の特性は，FIRフィルタでモデル化できるが，自動車内の状況（人間の動作，人数，窓の開閉などの要因）により変化すると考えられる．この場合，適応フィルタの利用により，最適なフィルタ係数を推定する方法が有効であると考えられる．自動車内の状況の変化がほとんど起こらない場合には，あらかじめ最適なフィルタ係数を求めておき，フィルタ係数を固定して，回り込み音をキャンセルする方法で良いと思われる．しかしながら，自動車内の状況の変化がいつ発生するかは，一般に予測が困難である．また，音声認識アルゴリズムの観点から見ると，わずかな非定常加法性雑音の存在が誤認識につながることが知られている．適応的にフィルタリングを行う場合と固定的にフィルタリングを行う場合とで，どの程度の認識性能の差があるかについての評価は興味深い研究課題であるが，本論文では，適応フィルタにより継続的にフィルタ係数を推定するアプローチに絞り，音声認識性能をどの程度改善できるかを評価する．図1に適応フィルタによる回り込み音のキャンセルのブロック図を示す．スピーカから出力される

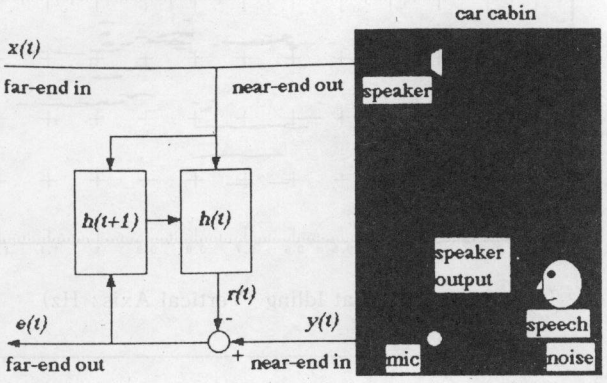


図 1 スピーカ音のキャンセル
Fig. 1 Canceller of speaker output.

音楽ソースの適応フィルタへの入力を遠端入力（far-end in），マイク入力を近端入力（near-end in），スピーカ出力を近端出力（near-end out），回り込み音のキャンセル後の出力信号を遠端出力（far-end out）と呼ぶ．また，遠端入力と近端出力は全く等価であると仮定し，遠端入力から近端出力が生成される系の特性（スピーカ特性など）は，回り込み音生成経路の特性に含めるものとする．

適応フィルタのアルゴリズムとして，これまでにLMS（Least Mean Square error），NLMS（Normalized Least Mean Square error）（学習同定法とも呼ばれる），RLS（Recursive Least Squares）などが提案されている[11]．時刻 t におけるFIRフィルタの係数，FIRフィルタへの入力データを

$$\boldsymbol{h}(t)=[h_1(t), h_2(t), \cdots, h_M(t)]^T$$

$$\boldsymbol{x}(t)=[x(t), x(t-1), \cdots, x(t-M+1)]^T$$

で表現する．ここで，T は転値を示す．また，時刻 t のマイク入力信号を $y(t)$ とすると，NLMSは，一般に以下の式で与えられる．

$$r(t)=\boldsymbol{h}(t)^T\boldsymbol{x}(t) \quad (8)$$

$$e(t)=y(t)-r(t) \quad (9)$$

$$\boldsymbol{h}(t+1)=\boldsymbol{h}(t)+\frac{\mu}{a+\|\boldsymbol{x}(t)\|^2}\boldsymbol{x}(t)e(t) \quad (10)$$

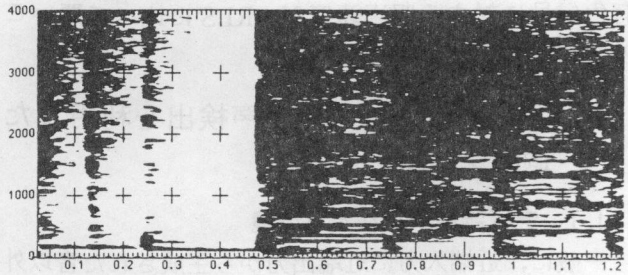
ここで，$\|\cdot\|^2$ はベクトルのエネルギーを表す．μ は，フィルタ係数の更新速度を決定する定数で，フィルタ係数が収束するために，$0<\mu<2$ を満たす必要がある．a は，$\|\boldsymbol{x}(t)\|^2$ が微小値の場合に式(10)の右辺第2項が発散するのを防止するための正の定数である．$r(t)$ を擬似回り込み音信号，$e(t)$ を回り込み音キャンセル信号と呼ぶことにする．NLMSは演算量が少なく実用的な方法としてよく使われるが，音声などの有色信号に対する収束速度が，RLSに比べて悪いことが知られている．

4. フレームワイズの音声検出を利用した適応フィルタ

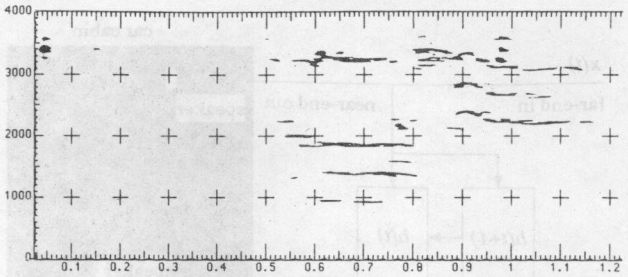
4.1 走行雑音が存在する環境

一般に，近端入力に近端出力から生成された音以外の音が混入する（以下，近端入力が存在すると言う）状況で係数の適応化を継続した場合，フィルタ係数の推定精度が劣化し，回り込み音のキャンセル性能が悪化する．そこで，遠端入力が存在し，かつ近端入力が存在する状態（ダブルトーク状態と呼ぶ）では，式(10)によるフィルタ係数の更新を停止させることが一般に行われる．遠端入力が存在するかどうかの判断は，遠端入力のエネルギーとあらかじめ定められたしきい値との単純な比較で可能である．一方，近端入力が存在するかどうかの判断を同様に行うと，回り込み音の影響で近端入力が存在すると判断する場合が多くなり，式(10)によるフィルタ係数の更新を頻繁に停止して，結果的にフィルタ係数の推定精度が劣化するという不具合が生じる．そこで，近端入力信号 $y(t)$ ではなく，回り込み音キャンセル信号 $e(t)$ のエネルギーを用いて，近端入力が存在するかどうかを判断するという方法が考えられる．近端出力から生成された音以外で近端入力に混入する音としては，大きく分けて走行雑音などの音源未知の加法性雑音と人間の音声の二つが考えられるが，いずれも適応フィルタで除去されずに遠端出力に残存する．一般に，走行中の自動車環境では，音源未知の加法性雑音のエネルギーレベルは大きく変動するため，近端入力の存在を判断するための最適なしきい値を一意に決めることは難しいという問題がある．

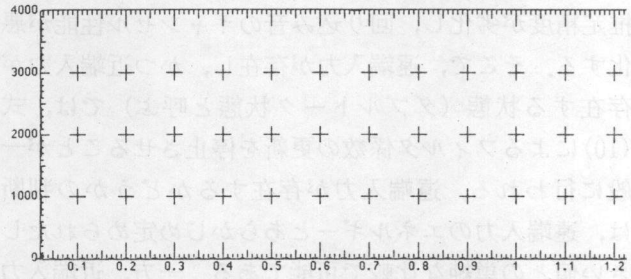
そこで，まず走行雑音のみが存在する場合での，NLMSによる回り込み音のキャンセル性能について評価を行う．図2の(a)，(b)，(c)，(e)，(f)に，それぞれ遠端入力信号（ポップス音楽）のスペクトログラム，アイドリング時での近端入力信号のスペクトログラム，アイドリング時での回り込み音キャンセル信号のスペクトログラム，時速100 km走行時の近端入力信号のスペクトログラム，時速100 km走行時の回り込み音キャンセル信号のスペクトログラムを示す．カーオーディオの音量は，アイドリング時と時速100 km走行時で，男性1名が快適と感じるレベルにセットした．従って，時速100 km走行時の方が，



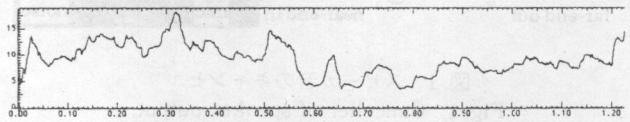
(a) Far End In (Vertical Axis: Hz)



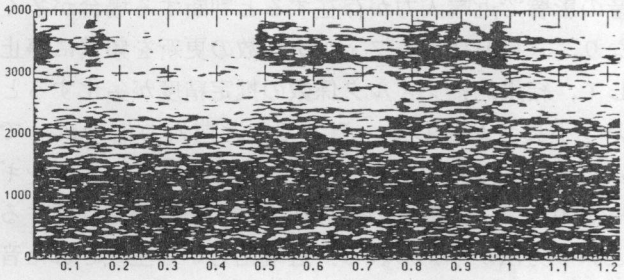
(b) Near End In at Idling (Vertical Axis: Hz)



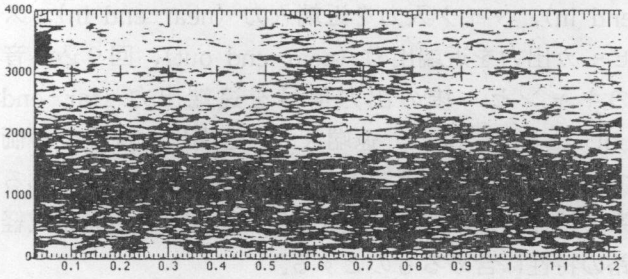
(c) Far End Out at Idling (Vertical Axis: Hz)



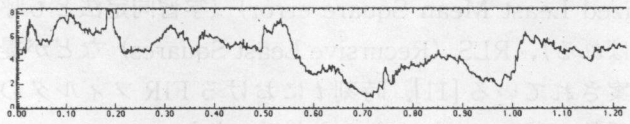
(d) ERLE at idling (Vertical Axis: dB)



(e) Near End In at 100 kmph (Vertical Axis: Hz)



(f) Far End Out at 100 kmph (Vertical Axis: Hz)



(g) ERLE at 100 kmph (Vertical Axis: dB)

図 2 NLMS の性能（横軸：s）
Fig. 2 Performance of NLMS (Horizontal Axis: s).

スピーカ出力レベルは大きく，回り込み音のレベルも大きい．近端入力信号は，2000 cc の乗用車の運転席サンバイザに単一指向性マイクを設置して収録した．フィルタ係数の初期値はすべて 0.0 とし，時刻 0 秒から継続的に式(8)～(10)によりフィルタ係数を更新しながら回り込み音キャンセル信号を求めた．サンプリング周波数は 8 kHz であり，回り込み音の最大遅延は 32 ms まで考慮した．従って，FIR のタップ数は 256 である．また，図 2 の(d)，(g)に，それぞれアイドリング時，時速 100 km 走行時での ERLE (Echo Return Loss Enhancement) の推移を示す．ERLE は近端入力信号の減衰量を表し，適応フィルタの性能を評価する尺度としてよく用いられ，次式で定義される [12].

$$ERLE=10\cdot\log_{10}\frac{E[y(t)^2]}{E[e(t)^2]} \tag{11}$$

$E[\cdot]$ は，推定値を表し，次式により求めた．

$$E[z(t)^2]=(1-\lambda)\cdot E[z(t-1)^2]+\lambda\cdot z(t)^2 \tag{12}$$

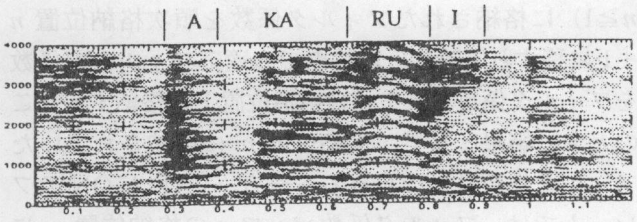
但し，$\lambda=1/256$ とした．ERLE の単位は，dB である．アイドリング時の ERLE の最大値，平均値はそれぞれ 18.80 dB，10.13 dB である．また，時速 100 km 走行時の ERLE の最大値，平均値はそれぞれ 9.33 dB，5.89 dB である．近端入力の音源未知の加法性雑音のレベルが大きいほど，式(11)で与えられる ERLE は低い値になる．

図 2(c)，(f)からアイドリング時，時速 100 km 時いずれの場合も回り込み音をほぼキャンセルできていることがわかる．近端入力に人間の音声が含まれな

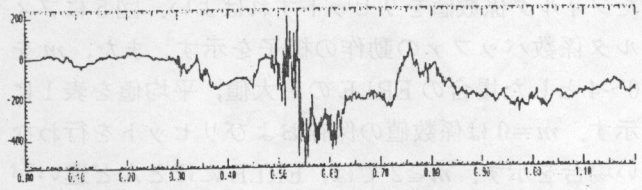
い場合は，フィルタ係数を継続的に更新することにより回り込み音の大部分はキャンセル可能であると思われる．すなわち，音源未知の加法性雑音の中で定常的かつ音声と無相関である走行雑音は，フィルタ係数の推定に与える影響が小さいと考えられる．

4.2 走行雑音および音声が存在する場合

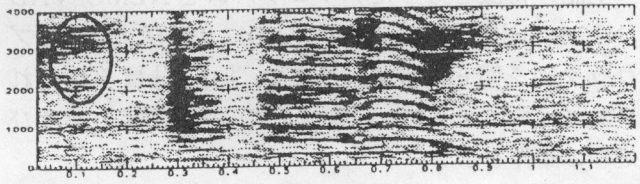
次に，近端入力に人間の音声が含まれる場合について調べる．2000 cc の自動車でカーオーディオからポップス音楽を再生しながら市街地を時速 60 km で走行し，運転席サンバイザに設置した単一指向性マイクで加法性雑音データを収録した．このとき，音楽のボリュームは女性1名が快適と感じるレベルにセットした．次に，停止中（エンジンオフ）の同一自動車内で同一女性1名が発声した音声データ（「明るい」）を同一録音レベルで収録した．そして，加法性雑音データと音声データとを計算機上で加算した信号のスペクトログラムを図3(a)に示す．図3(b)にフィルタ係数の初期値をすべて 0.0 とし，時刻0秒から連続的にフィルタ係数を更新した場合の回り込み音キャンセル信号のスペクトログラムを示す．また，図3(c)にフィルタ係数の 10 番目の係数の値の変化を示す．このときの ERLE の最大値，平均値はそれぞれ 8.48 dB, 4.18 dB である．特に，時刻 0.5 秒あたりから 0.15 秒の間，フィルタ係数値が激しく振動し，不安定になっている様子がわかる．また，時刻 1.0 秒以降の回り込み音（図3(b)の円で囲まれた部分）を十分にキャンセルできていない．



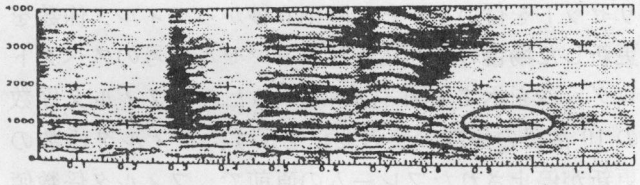
(a) Near End In (Vertical Axis: Hz)



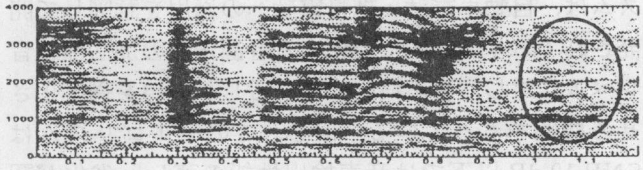
(c) 10th Coeff. of FIR filter without VAD (Vertical Axis: Coeff. value)



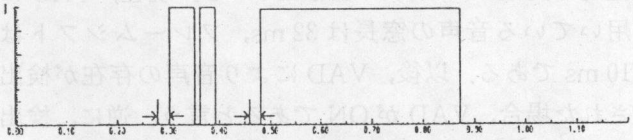
(e) Far End Out by NLMS-VAD (Vertical Axis: Hz)



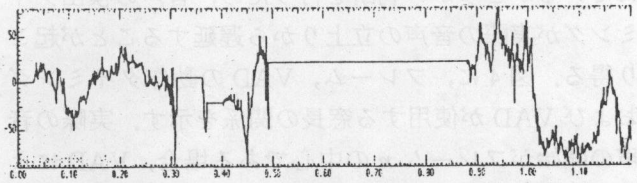
(g) Far End Out with previously estimated FIR (Vertical Axis: Hz)



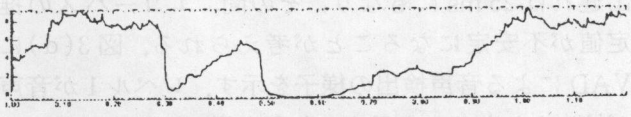
(b) Far End Out by conventional NLMS (Vertical Axis: Hz)



(d) VAD result (Vertical Axis: VAD value)



(f) 10th Coeff. of FIR filter with VAD (Vertical Axis: Coeff. Value)



(h) ERLE in (g) (Vertical Axis: dB)

図 3 NLMS-VAD の効果（横軸：s）
Fig. 3 Effect of NLMS-VAD (Horizontal Axis: s).

4.3 フレームワイズの音声検出の利用

近端入力に音声が存在する間はフィルタ係数の更新を停止し，近端入力に音声が存在しない間は，定常的な加法性雑音の存在のいかんにかかわらずフィルタ係数の更新を継続する必要がある．そのためには，音源未知の加法性雑音が混入する近端入力に音声が含まれているかどうかを判定する音声検出アルゴリズムVAD（Voice Activity Detection）が必要となる．VADとして，これまでにもさまざまなアルゴリズムが提案されているが，サンプルワイズにエネルギーを計算する方法では，加法性雑音と音声を分離することは困難であると思われる．そこで，本論文では，LPC分析の残差エネルギー，スペクトルの定常性およびピッチ性を利用して，フレームワイズにVADを行うアルゴリズム[13]に注目をした．このアルゴリズムは，自動車電話・携帯電話への適用を意図して開発されており，動的に変化する音源未知の加法性雑音レベルに対し，音声検出に用いるしきい値を適応化させる機能をもっている．但し，このアルゴリズムはSNR 10 dB以下では音声検出性能の著しい劣化が認められたため，音声検出のしきい値の適応化の部分のパラメータの変更などを行い，低SNRでの音声検出性能を向上させた方法を用いている．現在，VADに用いている音声の窓長は32 ms，フレームシフトは10 msである．以後，VADにより音声の存在が検出された場合，VADがONであると言う．逆に，検出されなかった場合，VADがOFFであると言う．このVADは，1フレームに1回近端入力に音声が含まれているかどうかの判断を行うため，音声の検出タイミングが実際の音声の立上りから遅延することが起こり得る．図4に，フレーム，VADの動作タイミングおよびVADが使用する窓長の関係を示す．実際の音声の開始がフレームnの中心である場合，VADによりその音声の開始を検出できるのは，フレーム$n+1$以降である可能性が高い．仮に，フレーム$n+2$のVADで検出できた場合，実際の音声の開始からの検出遅れは25 msにもなり，その間，エコーパスの推定値が不安定になることが考えられる．図3(d)にVADによる音声検出の様子を示す．レベル1が音声が検出されたことを示す．矢印を用いて示すように，2フレーム程度の音声検出遅れが認められる．不安定になったフィルタ係数値をより精度の高い値に回復することができれば，回り込み音のキャンセル性能の低下を避けることが可能だと考えられる．

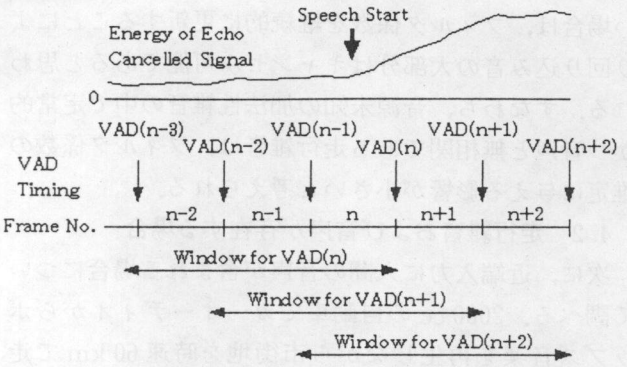


図4 VADの動作タイミング
Fig. 4 Timing of VAD operation.

そこで，m個分のフィルタ係数を格納できるバッファ（フィルタ係数バッファと呼ぶ）を用意する．VADがOFFのフレームでは，格納位置n（$m-1 \geqq n \geqq 1$）に格納されたフィルタ係数を順次格納位置$n+1$に移すと同時に，現時点での適応フィルタの係数をフィルタ係数バッファの格納位置1に格納する．このとき，結果として，格納位置mに格納されていたフィルタ係数は捨てられる．一方，VADがONのフレームでは，フィルタ係数バッファの格納位置mに格納されたフィルタ係数を取り出し，その値で劣化したフィルタ係数値をリセットすればよい．図5にフィルタ係数バッファの動作の様子を示す．また，mを0〜4とした場合のERLEの最大値，平均値を表1に示す．$m=0$は係数値の保存およびリセットを行わない場合を示す．$m \geqq 2$では，ERLEにほとんど違いが見られないため，$m=2$を選択する．これは，VADの検出遅れ（2フレーム程度）と対応している．

上記の特徴をもったアルゴリズムをNLMS-VAD（NLMS with frame-wise VAD）と呼び，全体のブロック図を図6に示す．ここで，$[s]$,$[f]$はそれぞれサンプルワイズ，フレームワイズの信号の流れおよび処理の動作を示す．

図3(e)に，フィルタ係数の初期値をすべて0.0とし，VADを動作させ，フィルタ係数値の格納およびリセットを行いながら，時刻0秒からフィルタ係数を更新した場合の回り込み音キャンセル信号のスペクトログラムを示す．図3(f)にそのときのフィルタ係数の10番目の係数の値の変化を示す．フィルタ係数の更新が停止されたフレームの直前で，フィルタ係数値が不安定になっているが，上記フィルタ係数の格納およびリセットにより，フィルタ係数が回復されている様子が示されている．これにより，時刻1.0秒以降の

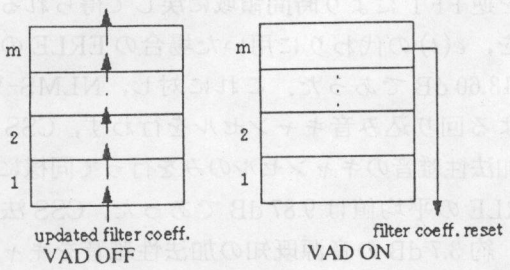


図 5 フィルタ係数バッファの動作
Fig. 5 Operation of filter coeff. buffer.

表 1 フィルタ係数バッファのサイズと ERLE の関係
Table 1 Relationship between filter coeff. buffer size and ERLE.

buffer size m	max ERLE(dB)	average ERLE(dB)
0	8.80	4.18
1	9.06	4.25
2	9.15	4.35
3	9.14	4.36
4	9.14	4.36

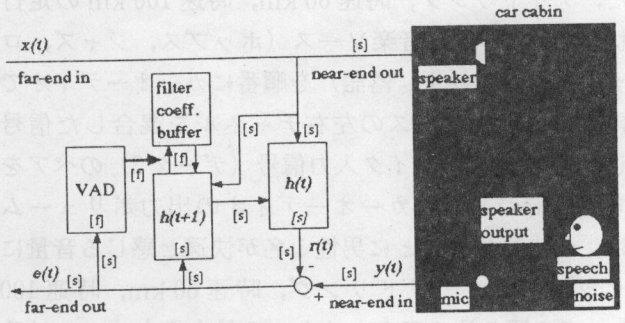


図 6 NLMS-VAD のブロック図
Fig. 6 Block diagram of NLMS-VAD.

回り込み音（図 3(b)の円で囲まれた部分）もキャンセルされている．

但し，図 3(e)で時刻 0.1 秒前後の回り込み音（図 3(e)の円で囲まれた部分）がキャンセルされていないことがわかる．発声ごとに推定されたフィルタ係数および VAD に用いられるパラメータを保存しておき，次の発声時にそれらを初期値として用いれば，フィルタ係数の推定速度は速まると考えられる．図 3(g)にその例を示す．時刻 0.0 秒直後の回り込み音は若干残存しているが，それ以後の回り込み音（図 3(e)の円で囲まれた部分）はほぼキャンセルされていることがわかる．また，ERLE を図 3(h)に示す．ERLE の最大値，平均値はそれぞれ 9.29 dB，4.50 dB である．

5. 音源既知，音源未知の加法性雑音と乗法性ひずみが存在する環境でのロバスト音声認識

筆者らは先に，音源未知の加法性雑音および乗法性ひずみが存在する環境におけるロバストな音声認識方法として，連続スペクトル減算 CSS と話者依存の音声・非音声分離型ケプストラム平均正規化 E-CMN を組み合わせた E-CMN/CSS 法を提案した [6]．CSS 法は，音声フレームと非音声フレームを区別せず，連続的にスペクトルの移動平均を求め，これを雑音スペクトルの推定値とみなして，入力スペクトルから減算する方法である．雑音スペクトルの推定値に音声スペクトルの影響が含まれるため，エネルギーの弱い音声スペクトルがマスクされてしまい，ひずみが生じるという問題点がある．一方，CSS 法は，過去のある一定時間長の区間に対して，相対的に大きなエネルギーをもつ周波数成分を残し，エネルギーの微弱な周波数成分を雑音，音声を問わず，マスクするという働きをもつ．このため，クリーンな音声に CSS 法を施した後に得られる特徴パラメータと加法性雑音が重畳した音声に CSS 法を施した後に得られる特徴パラメータの間の変動が，通常のスペクトル減算法や最小平均 2 乗誤差推定法に比べて小さい．この特長は，低い SNR での音声認識にとって有効である．また，E-CMN 法は，式(4)の乗法性ひずみを 10 単語程度の少量の音声から，音声区間のケプストラム平均として推定し，それを音声区間の入力ケプストラムから引くという方法である．不特定話者音素モデルを観測されたスペクトルから求められたケプストラムではなく，E-CMN 法により正規化されたケプストラムを用いて作成することにより，乗法性ひずみの正確な補正が可能である．

そこで，音源既知および音源未知の加法性雑音，乗法性ひずみが存在する実環境におけるロバストな音声認識手法として，既提案の E-CMN/CSS 法に NLMS-VAD 法を組み合わせる手法を提案する．本手法の概略図を図 7 に示す．なお，E-CMN 法で必要な音声/非音声フレームの区別は，NLMS-VAD 法に内蔵された VAD の結果をそのまま用いている．

図 8(a)に，停止中（エンジンオフ）の自動車内で女性が発声した音声（「明るい」，図 3(a)に示した音

声を計算機上で加算して作成した際に用いた音声と同一）にCSS法を施した後のスペクトログラムを，(b)に同一音声に時速60 km走行時の音源未知の加法性雑音と回り込み音が重畳した雑音データを計算機上で加算した後（図3(a)），NLMS-VAD法で回り込み音をキャンセルし（図3(g)），CSS法を施して得られるスペクトログラムを示す．図3(g)と図8(b)を比較すると，時刻0.9秒近辺の周波数1 kHz前後の回り込み音の残存成分（図3(g)の円で囲まれた部分）がCSS法により除去されていることがわかる．CSS法は，定常的な加法性雑音だけでなく，NLMS-VAD法でキャンセルできなかった回り込み音を抑圧する効果ももっている．式(11)において，回り込み音キャンセル信号$e(t)$にFFTを施して得られたスペクトルに対してCSS法を施した後のスペクトルを逆FFTにより時間領域に戻して得られる波形信号を，$e(t)$の代わりに用いた場合のERLEの平均値は13.60 dBであった．これに対し，NLMS-VAD法による回り込み音キャンセルを行わず，CSS法による加法性雑音のキャンセルのみを行って同様に求めたERLEの平均値は9.87 dBであった．CSS法のみでは，約3.7 dBの音源既知の加法性雑音をキャンセルできなかったとみなすことができる．

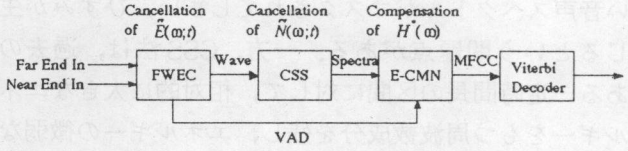
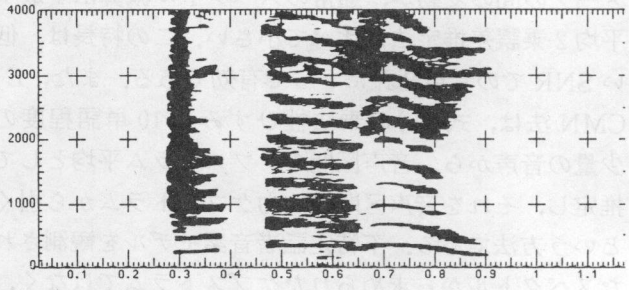
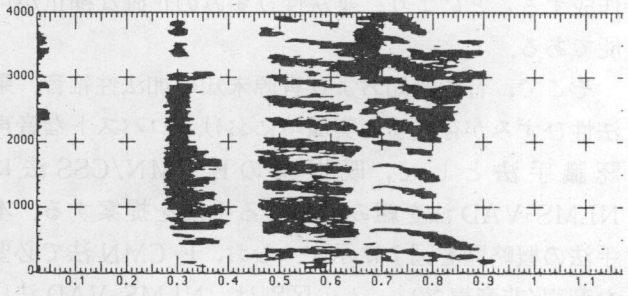


図7 NLMS-VADとE-CMN/CSSの組合せ手法のブロック図
Fig. 7 Block diagram of NLMS-VAD with E-CMN/CSS.



(a) Spectrogram after CSS for clean speech.



(b) Spectogram after CSS for Fig. 3 (g).

図8 CSSの効果（横軸：s）
Fig. 8 Effect of CSS (Horizontal Axis: s).

図8(a)と(b)を比較すると二つのスペクトログラムが極めて類似していることがわかる．NLMS-VAD法とCSS法の組合せにより，音源既知と音源未知の加法性雑音に対して，ロバストな特徴パラメータを抽出できることが示唆されている．

6. 性能評価

（実験）単一指向性マイクを2000 ccの乗用車の運転席サンバイザに設置し，男性2名女性2名が各々好みの位置にセットした運転席に座って発声した520単語（ATR音声データベースCセット）の音声（データ1）を収録した．音声区間の前後に250 msずつの無音区間が付属するように手動で切出しを行った．また，アイドリング，時速60 km，時速100 kmの走行状態で，5種類の音楽ソース（ポップス，ジャズ，ロック，クラシック，落語）を順番にカーオーディオで再生し，音楽ソースの左右チャネルを混合した信号（データ2）と，マイク入力信号（データ3）のペアを同時に録音した．カーオーディオの出力ボリュームは，各走行状態ごとに男性1名が快適と感じる音量にセットした．アイドリング，時速60 km，時速100 kmでの回り込み音のマイクへの最大入力レベルはそれぞれ，60.7 dBA，65.9 dBA，70.6 dBAであった．データ1とデータ3を計算機上で加算して評価データを作成した．データ2は，NLMS-VAD法の遠端入力として使用した．認識には，環境独立な54音素の不特定話者用Tied-MixtureHMM（ATRの音声データベースAセット4名，Cセット36名から作成）を用いた．分析条件は8 kHzサンプリング，フレーム長32 ms，フレームシフト10 msで，特徴パラメータは，10次MFCC，10次ΔMFCC，Δエネルギーであり，HMMが共有する正規分布の数は，それぞれ256，256，64である．不特定話者，520単語の認識タスクで，アイドリング，時速60 km，時速100 kmの走行状態で，回り込み音が存在しない場合（w/o Speaker Out），回り込み音が存在するが，NLMS-

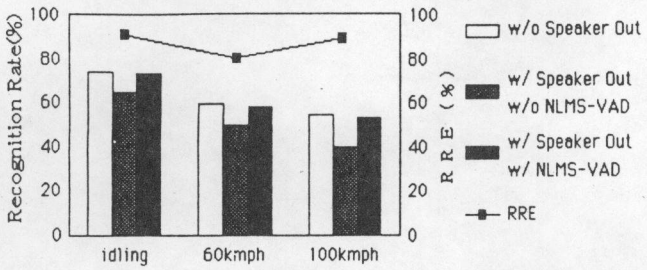


図 9 NLMS-VAD と E-CMN/CSS の組合せ手法の性能
Fig. 9 Performance of NLMS-VAD with E-CMN/CSS.

表 2 音楽ソースごとの認識性能 (%)
Table 2 Performance for each music source(%).

	idling	60kmph	100kmph
pops	72.4	57.1	53.2
rock	73.2	59.2	49.3
jazz	73.9	55.8	50.8
classic	72.9	57.1	54.2
rakugo	73.7	58.5	54.1
average	73.2	57.6	52.3

VAD 法を行わない場合 (w/Speaker Out w/o NLMS-VAD)，回り込み音が存在し，NLMS-VAD 法を行う場合 (w/Speaker Out w/ NLMS-VAD) の認識性能 (5 種類の音楽ソースの平均) を図 9 に示す．また，回り込み音の存在に起因して増加する誤認識のうち，NLMS-VAD 法により回り込み音をキャンセルすることにより正認識に転じる (回復できる) 割合を RRE (Recovery Rate of Error) と呼ぶことにする．アイドリング，時速 60 km，時速 100 km の走行状態での RRE の値も図 9 に示す．いずれの走行状態でも 80%以上の RRE が得られた．また，NLMS-VAD 法でも回復できない誤認識率は，アイドリング，時速 60 km，時速 100 km でそれぞれ 0.7%，2.1%，1.8%とわずかであり，NLMS-VAD 法の有効性が確認できた．表 2 に，NLMS-VAD 法を行う場合の各音楽ソースごとの認識性能を示す．音楽ソースによる認識性能の差が，走行時の音源未知の加法性雑音のレベルの差によるのか，音楽ソースの種類に対する NLMS-VAD 法の性能の差によるのかは今後の解析を待つ必要がある．

7. むすび

フレームワイズの音声/非音声の検出を利用した，音源既知の加法性雑音のキャンセルアルゴリズム NLMS-VAD を提案した．更に，音源既知および音源未知の加法性雑音と乗法性ひずみが存在する実環境におけるロバストな音声認識アルゴリズムとして，既提案の E-CMN/CSS 法の前処理として NLMS-VAD 法を組み合わせる方法を提案した．自動車内での不特定話者単語認識のタスクにより提案法を評価した結果，スピーカからの回り込み音に起因する誤認識はたかだか 2%程度であることを明らかにした．回り込み音に起因する誤認識を更に削減するには，NLMS-VAD 法の性能向上を図ることが重要である．しかし，回り込み音の完全な除去は困難であると思われるので，加法性雑音に対する音素モデルのロバスト性を向上させることも必要となろう．

謝辞 日ごろから熱心に御討論頂く奈良先端科学技術大学院大学音情報処理学講座および旭化成工業(株) LSI・情報技術研究所の諸氏に感謝します．

文 献

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. ASSP, vol. ASSP-27, no. 2, pp. 113-120, 1979.
[2] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," Computer Speech and Language, vol. 3, no. 2, pp. 151-167, 1989.
[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," IEEE Trans. ASSP, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude resonator," IEEE Trans. ASSP, vol. ASSP-33, no. 2, pp. 443-445, 1985.
[5] J. A. N. Flores and S. J. Young, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation," Proc. ICASSP, pp. I-409-412, Adelaide, 1994.
[6] 庄境 誠, 中村 哲, 鹿野清宏, "音声認識における音声強調手法及びモデル適応化手法の検討," 信学技報, SP96-19, 1996.
[7] 武田一哉, 黒岩眞吾, 井ノ上直己, 野垣内出, 山本誠一, 庄境誠, 尾和邦彦, 高橋正彦, 松本龍二, "連続音声認識に基づく内線番号案内システムの試作," 音講論集, 3-4-15, pp. 79-80, March 1993.
[8] 高橋 敏, 嵯峨山茂樹, "NOVO 合成法を用いた Barge-in 音声の認識," 音講論集, 2-5-1, pp. 59-60, March 1996.
[9] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Kluwer Academic Publishers, 1992.
[10] J. H. L. Hansen, B. D. Womack, and L. M. Arslan, "A source generator based production model for environmental robustness in speech recognition," Proc. ICSLP 94, pp. 1003-1006, Yokohama, 1994.
[11] S. Haykin, "Adaptive Filter Theory," Prentice-Hall,

Englewood Cliffs, NJ, 1991.
[12] 北脇信彦編著, "音のコミュニケーション工学ーマルチメディア時代の音声・音響技術ー," コロナ社, 1996.
[13] Recommendation GSM 06. 32.

（平成9年7月7日受付，12月8日再受付）

庄境　誠（正員）

昭56京大・工・数理卒．昭58同大大学院修士課程了．同年旭化成工業(株)入社．以来，音声音響情報処理，主として音声認識および音声符号化の研究に従事．昭60～62ヘルシンキ工科大客員研究員．現在，同社LSI・情報技術研究所勤務．平10奈良先端科学技術大学院大学情報科学研究科後期博士課程了．博士（工学）．IEEE，日本音響学会，人工知能学会各会員．

中村　哲（正員）

昭56京都工繊大・工・電子卒．平4京都大学博士（工学）．昭56～平6シャープ(株)中央研究所および情報技術研究所に勤務．この間，昭61～平1ATR自動翻訳電話研究所に出向．平6年4月より奈良先端科学技術大学院大学情報科学研究科勤務．助教授．音声情報処理，主として音声認識の研究に従事．平4日本音響学会粟屋学術奨励賞受賞．IEEE，日本音響学会，人工知能学会各会員．

鹿野　清宏（正員）

昭45名大・工・電気卒．昭47同大大学院修士課程了．同年電電公社武蔵野電気通信研究所入所．昭59～61カーネギーメロン大客員研究員．昭61～平2ATR自動翻訳電話研究所音声情報処理研究室長．平4NTTヒューマンインタフェース研究所主席研究員．平6より奈良先端科学技術大学院大学情報科学研究科教授．音情報処理学講座を担当．工博．主として音声・音情報処理の研究および研究指導に従事．昭50本会米沢賞．平3IEEE SP 1990 Senior Award，平6日本音響学会技術開発賞．IEEE，日本音響学会，情報処理学会各会員．