雑音混入音声スペクトルのエントロピーと分散によるVADと雑音スペクトルサプレッション法への応用

A VAD Based on Entropy and Variance of Noisy Speech Spectrum and Its Application to Noise Spectral Suppression

山下 新司　　中山 謙二　　平野 晃宏
金沢大学大学院 自然科学研究科 電子情報工学専攻

Shinji YAMASHITA　　Kenji NAKAYAMA　　Akihiro HIRANO
Division of Electrical and Computer Engineering,
Graduate School of Natural Science and Technology, Kanazawa Univ.
E-mail : yamasita@leo.ec.t.kanazawa-u.ac.jp
nakayama@t.kanazawa-u.ac.jp

あらまし

携帯電話で用いられるノイズキャンセラとしてスペクトルサプレッション（SS）法が検討されている．SS 法においては雑音スペクトルの推定が非常に重要となっている．このために，雑音混入音声のフーリエ変換を Voice Activity Detector (VAD) により音声区間と無音区間に分けて雑音スペクトルを推定する方法が提案されている．従来の VAD では雑音混入音声スペクトルのエントロピーが用いられている．本稿ではさらに，分散を導入し，エントロピーと分散で形成される 2 次元平面上で傾斜を有する直線を境界線として音声区間／準音声区間／無音区間の識別を行う．境界線を自動的に制御する方法を提案する．シミュレーションにおいて，白色，バブル，車の 3 種類の雑音を用いて SNR 等を評価し，従来法に比べて特性が改善されていることを確認した．

ABSTRACT

Several approaches based on a spectral suppression (SS) method have been proposed for noise cancellers used in mobile phones. In the SS method, noise spectral estimation is very important. For this purpose, Fourier transform of noisy speech is divided into the speech frame and the non-speech frame by using Voice Activity Detector (VAD). The noise spectrum is estimated in the speech and the non-speech frames in different ways. The conventional VAD employs the entropy of the noisy speech spectrum. In this paper, we introduce the variance of the noisy speech spectrum. The noisy speech spectrum is classified into the speech, the quasi-speech and the non-speech frames on the 2-dimensional space spanned by the entropy and the variance. Straight lines with some angles are used for the decision boundaries. A method, automatically control the straight lines, is proposed. Simulation results, using three kinds of noises, white, bubble and car, demonstrate the speech, the quasi-speech and the non-speech frames are well descriminated and the segmental SNR is also improved .

1 はじめに

現在，携帯電話などの移動通信が普及し，街頭や車内など背景雑音が多い場所で携帯電話が使用される場合も多い．このような使用環境では雑音を除去するためのノイズキャンセラが必要となる．携帯電話ではマイクが単一である場合が多いので，これに適したスペクトルサプレッション（SS）法によるノイズキャンセラが開発されている．この方式では，雑音混入音声のスペクトルと雑音スペクトルの推定値からスペクトルゲインを計算し，雑音混入音声に乗じて雑音成分を抑制する [1]-[5]．従って，SS 法では雑音スペクトルの推定が非常に重要である．雑音スペクトルが過小推定されると，スペクトルゲインを乗じた後でも雑音が大きく残る．また，雑音スペクトルが過大推定になると，雑音抑圧後に音声が大きく歪み，音質が劣化する．[6]-[8]

従来法における雑音スペクトル推定の一つのアプローチとして，Voice Activity Ditector（VAD）により，雑音混入音声のフーリエ変換をそのエントロピーにより音声区間と無音区間に分けて，各々の区間で異なる方法で雑音スペクトルを推定する方法が提案されている [9],[10]．さらに，音声区間／準音声区間／無音区間に分ける方法も

提案されている [11],[12]．また，分散を導入し，エントロピーと分散で構成される 2 次平面において，T 字型の閾値を設け区間判別を行う方法も提案されている [13]．

本稿では，この 2 次平面上で傾きを有する直線を境界線として音声／準音声／無音区間を判別する方法を検討する．さらに，雑音の変化に対応するために，この直線（境界線）の分散方向への平行移動を自動制御する方法を検討する．シミュレーションでは，白色，バブル，車の 3 種類の雑音を用いてセグメンタル SNR 等の性能を評価する．

2　スペクトルサプレッション法

2.1　スペクトルサプレッション法の構成

スペクトルサプレッション法のブロック図を図 1 に示す [1]-[3]．

音声と雑音はともにスペクトル成分において統計的独立

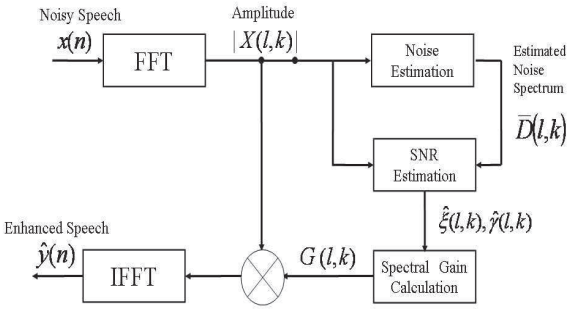

図 1: スペクトルサプレッション法のブロック図

であるとする．時間領域でのクリア音声を $x(n)$，雑音を $d(n)$ とおくと，雑音混入音声 $x(n)$ は，

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

と表せる．音声信号は一般に非定常でありその音響的特徴は変動している．そのため，音声のスペクトル分析では十分に短い時間の区間において音声は定常状態であるという仮定の基で，少しずつ時間区間をシフトさせながら窓関数を用いて切り出したフレームの波形のデータに対して順次 FFT 演算を実行し，スペクトル時系列を得ている．よって雑音混入音声は M サンプルのフレームに分けられていて，2M サンプルの窓関数を用いて 50% オーバーラップさせることにより，n 番目のフレームおける切り出された雑音混入音声 $\hat{x}_n(n)$ は次の式のように表せる．

$$\hat{x}_n(n) = \begin{cases} h(n)x_{n-1}(n) & ,1 \le n \le M \\ h(n)x_n(n-M) & ,M \le n \le 2M \end{cases} \tag{2}$$

この信号の l 番目のフレームにおける k 番目の周波数領域での表示を次のように表す．

$$X(l,k) = S(l,k) + D(l,k) \tag{3}$$

事前 SNR(クリーン音声対雑音比)，事後 SNR(雑音混入音声対雑音比) はそれぞれ次の式で表せる．

$$\xi(l,k) = \frac{E\{|S(l,k)|^2\}}{E\{|D(l,k)|^2\}} \tag{4}$$

$$\gamma(l,k) = \frac{|X(l,k)|^2}{E\{|D(l,k)|^2\}} \tag{5}$$

実際に利用可能なものは，雑音混入音声のみで，事前 SNR，事後 SNR は推定しなくてはいけない．事前 SNR は decision-directed 方式で以下のように推定できる [1]．

$$\hat{\xi}(l,k) = \alpha\gamma(l-1,k)G^2(l-1,k) + (1-\alpha)P[\gamma(l,k)-1] \tag{6}$$

ただし，α は $0 < \alpha < 1$，$P[x]$ は次の式を満たす．

$$P[x] = \begin{cases} x & (x > 0) \\ 0 & (otherwise) \end{cases} \tag{7}$$

事後 SNR は推定した雑音スペクトル $\hat{D}(l,k)$ を用いて，次のように推定する．

$$\hat{\gamma}(l,k) = \frac{|X(l,k)|^2}{\hat{D}(l-1,k)} \tag{8}$$

以上のように推定した事前 SNR，事後 SNR によりスペクトルゲイン $G(l,k)$ を求め，雑音混入音声に乗じることで雑音を抑える．

2.2　Joint MAP 法

Joint MAP 法は，雑音はガウス分布，音声をスーパーガウス分布という仮定のもとでスペクトルゲインを計算する方法である [2]．

Joint MAP 法におけるスペクトルゲインは，

$$G(l,k) = u(l,k) + \sqrt{u^2(l,k) + \frac{\tau}{2\hat{\gamma}(l,k)}} \tag{9}$$

$$u(l,k) = \frac{1}{2} - \frac{\mu}{4\sqrt{\hat{\gamma}(l,k)\hat{\xi}(l,k)}} \tag{10}$$

と求められる．

3　エントロピーを閾値とする VAD による雑音スペクトル推定法

本節では，従来の VAD を用いた雑音スペクトル推定法 [10] とその改良法 [11],[12] について述べる．

3.1　エントロピーの計算

VAD とは，入力信号のスペクトルエントロピー $H(l)$ を用いた音声区間検出である [7]．無音区間では，スペクトルエントロピーは音声フレームに比べて大きくなる．そこで，入力信号の最初の区間を無音区間と仮定し，最初の数フレーム分のスペクトルエントロピーの平均値に定数 $c1$，$c2$ を掛けたものを閾値 $\sigma1$，$\sigma2$ とし，その後のフレームでは，スペクトルエントロピーが閾値 $\sigma1$ よりも小

さい場合は音声区間，$\sigma 1$ よりは大きく，$\sigma 2$ よりは小さい場合は準音声区間，閾値 $\sigma 2$ よりも大きい場合は無音区間とする．スペクトルエントロピー $H(l)$ は次のように求められる．

$$H(l) = -\sum_{k=1}^{2M} P_r(l,k) \cdot \log(P_r(l,k)) \tag{11}$$

$$P_r(l,k) = \frac{|X(l,k)|^2 + C}{\sum_{k=1}^{2M} |X(l,k)|^2 + C} \tag{12}$$

ただし，式中の $2M$ は周波数のデータ数である．また，音声スペクトルのほとんどが周波数帯域 $250Hz$ 以上，$4000Hz$ 以下に存在するので，次のように定める．

$$|X(l,k)|^2 = 0,\ k \leq 250Hz\ or\ k \geq 4000Hz \tag{13}$$

これによって判断された各フレームは，それぞれに適した雑音推定アルゴリズムを適用することで，急激に雑音環境が変化した場合でも，高速かつ正確に雑音スペクトルを推定する．以下に各フレームで用いるアルゴリズムについて説明する．

3.2 雑音スペクトルの推定

3.2.1 無音フレーム

無音フレームでは雑音スペクトルを次のように推定する．

$$\bar{D}(l,k) = |X(l,k)|^2 \tag{14}$$

3.2.2 音声フレーム，準音声フレーム

準音声フレームと音声フレームは，重み付き雑音推定法を用いて雑音を推定している．具体的に，事後 SNR の推定値に基づき重み係数を決め，それをかけることで $z(l,k)$ を計算する．それらを複数フレームに亘って平均することによって，雑音スペクトルを推定している．

$$z(l,k) = W(l,k) \cdot |X(l,k)|^2 \tag{15}$$

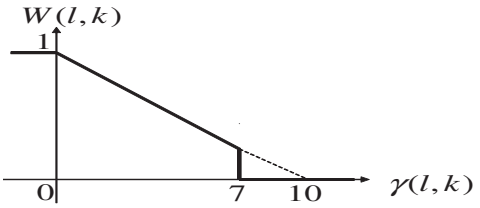

図 2: 音声フレームにおける重み係数

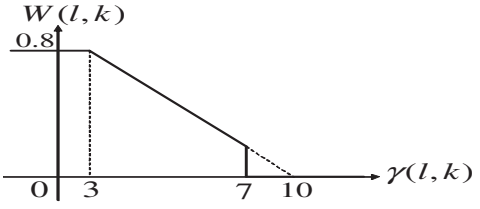

図 3: 準音声フレームにおける重み係数

4 エントロピーと分散による 2 次元閾値を用いた VAD による雑音スペクトル推定法

雑音混入音声スペクトルのエントロピーと分散による T 字型の 2 次元閾値を用いる VAD が提案されている [13]．これにより，区間判別の精度が向上されている．エントロピーと分散に対する閾値 T_h，T_v を以下に示す．

$$T_h(l) = c_1 E[H(l)] \tag{16}$$

$$T_v(l) = c_2 \frac{\max(V(l)) + median(V(l))}{2} \tag{17}$$

$$V(l) = \log |VAR(X(l,k))| \tag{18}$$

$E[H(l)]$ は無音区間と判定された最近 5 フレームのエントロピーの平均値を示し，$\max(V(l))$ と $median(V(l))$ は準音声または無音区間と判定された最近 5 フレームの最大値と中間値を表す．$T_h(l)$ と $T_v(l)$ の初期値は，無音区間とみなす最初の 5 フレームの平均値を用いることにより決定する．$H(l) > T_h(l-1)$ の場合は，$T_h(l)$ は式 1 6 によって更新され，$V(l) < c_3 T_v(l-1)$ または $V(l) > median(V(l))$ の場合，$T_v(l)$ は式 1 7 によって更新される．c_1，c_2，c_3 は経験則によりそれぞれ 0.98，1.1，1.02 とする．

この方式では，図 4 のように T 字型の分類法によってスペクトルエントロピーと分散の 2 次平面から音声・準音声・無音の 3 つの区間を判定する．$H(l) > T_h$ の場合，そのフレームは無音区間と判定され，$H(l) \leq T_h(l)$ かつ $V(l) < T_v(l)$ の場合，そのフレームは準音声区間と判定される．そして，$H(l) \leq T_h(l)$ かつ $V(l) \geq T_v(l)$ の場合，そのフレームは音声区間と判定される．

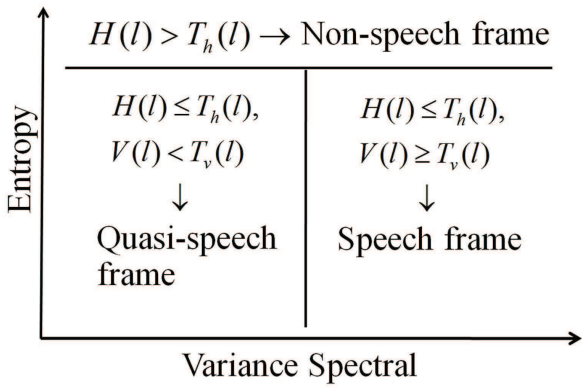

図 4: T 字型分類法

雑音推定アルゴリズムについては，3.2 で述べた方法と同じく，無音区間ではそのまま雑音として推定し，音声・準音声区間では重み付き雑音推定法によって雑音を推定している．

5　エントロピーと分散による 2 次元閾値の改良と雑音スペクトル推定

本節では，上記で述べたエントロピーと分散による 2 次平面を用いる方式において，VAD 法を改良した方式について述べる．

5.1　傾きを有する直線形閾値による VAD

従来の 2 次元閾値を用いる VAD は T 字型の分類法で区間判別を行うという方法であったが，本研究では，それを傾きを持った直線を用いて判別することにより，より区間判別の精度を高め，ノイズキャンセラとしての性能向上を図っている．各フレームのエントロピーと分散の値をもとに，一定の傾きをもった直線の切片を求め，それによって区間判別を行う方法を提案する．具体的に，エントロピー $H(l)$ を y 軸，分散 $V(l)$ を x 軸とみなし，傾き 5.7 の直線の y 軸との切片の値を $I(l)$ とする．ここでの傾きは経験則により決定している．2 つの閾値 T_1，T_2 を定めることにより，$I(l) \geqq T_1$ の場合は無音区間，$T_1 > I(l) > T_2$ の場合は準音声区間，$T_2 \geqq I(l)$ の場合は音声区間とする．以下に図で示す．

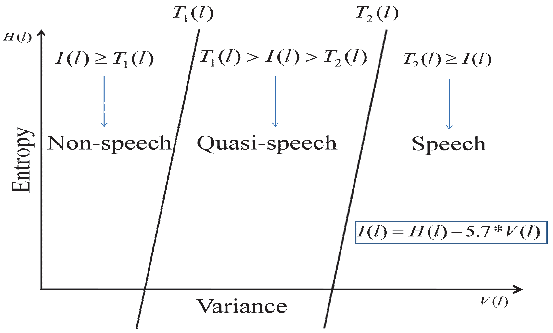

図 5: 傾きを持った直線による分類法

5.2　傾きを有する直線形閾値の制御方法

急激な雑音変化に対応するために，提案した VAD を制御する必要がある．提案法では，最初の 5 フレームを無音区間として保存し，その中で最小の $I(l)$ を t とする．6 フレーム目で t から c_1，c_2 を引くことにより，閾値 T_1，T_2 を作る．これによって $I(l)$ の値から，区間判別を行う．また，無音区間と判別された場合は，無音区間の最近 5 フレームの中で最小値を t とし，そこから d_1，d_2 を引くことにより，雑音変化時の閾値の調整を行う．c_1，c_2，d_1，d_2 については，それぞれ 3，6，1.545，11 とし，これは経験則により求められている．制御方法のフローチャートを図 6 に示す．

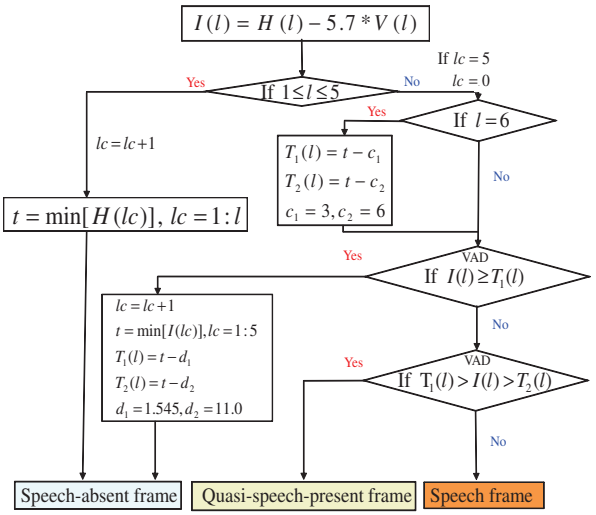

図 6: 制御方法のフローチャート

6　シミュレーション

6.1　雑音の種類

入力信号として，8kHz で標本化された男性及び女性の音声を用いた．雑音としては，雑音変化の無い場合は 3dB，9dB の White，Babble，Car Noise を付加し，雑音変化がある場合は，10000 サンプルまでは非定常な Babble Noise を付加し，10001〜30000 サンプルでは前半の Babble Noise より強さが大きい White，Babble，Car Noise を付加した．

6.2　雑音スペクトル推定及びノイズキャンセラとしての性能評価

6.2.1　雑音スペクトルの正規化推定誤差

雑音スペクトル推定精度の評価として，フレームごとに次の式 (19) で与えられる正規化推定誤差 $\varepsilon(l)$ を用いて評価した [3]．

$$\varepsilon(l) = 10\log_{10}\left(\frac{\sum_{k=0}^{M}\left||D(l,k)|^2 - |\hat{D}(l,k)|^2\right|}{\sum_{k=0}^{M}|D(l,k)|^2}\right) \quad (19)$$

$$\bar{\varepsilon} = \frac{1}{L}\sum_{l=1}^{L}\varepsilon(l) \quad (20)$$

ただし，L は全フレーム数である．上式の ε は，値が小さいほど雑音スペクトル推定が正確であるということを表している．また，$\bar{\varepsilon}$ は全フレームの正規化推定誤差 $\varepsilon(l)$ の平均値を表している．

6.2.2　時間領域における音声推定の評価

出力では，信号を 12ms の区間に分割し，各区間の SNR の平均を求めるセグメンタル SNR で評価を行う．SNR_{seg}

は各区間の SNR の平均を求める評価法である．音声信号は時々刻々と変化しているので，細かい時間間隔で SNR を求め，その平均値である SNR_{seg} は，雑音が低エネルギーで広域に分布している場合，雑音除去性能を正しく評価を行なうことができる．セグメンタル SNR は次式で定義される．

$$SNR_{seg} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{n=N_l}^{N_l+N-1}s^2(n)}{\sum_{n=N_l}^{N_l+N-1}(\hat{s}(n)-s(n))^2} \quad (21)$$

ただし，N は分析フレームの長さである．

6.2.3 音声スペクトル歪みの評価

音質の評価方法として，Log-Spectral Distortion を行なう．LSD は周波数領域においてクリーン音声の振幅値 $|S(l,k)|$ と，雑音抑圧音声の振幅値 $|\hat{S}(l,k)|$ の各分析フレームおにおける比率の平均値を求めていて，次式で表される．

$$LSD = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{2M}\sum_{k=1}^{2M}\left(\log\frac{|S(l,k)|+\delta}{|\hat{S}(l,k)|+\delta}\right)^2\right)^{\frac{1}{2}} \quad (22)$$

δ は微小値，2M は分析フレーム長である．LSD は常に正の値をとり，また値が小さいほど音質が良いことを表している．人間の聴覚に対して位相情報はあまり影響を与えないので，音質の評価方法としては位相情報も含まれる SNR_{seg} よりも，式 (22) で示される LSD の方が有効な評価方法と言える．

6.3 シミュレーション結果と考察

表 1，表 2，表 3 はそれぞれ雑音変化が無い場合，無音区間で雑音が変化した場合，音声区間で雑音が変化した場合のシミュレーション結果である．各表にて，上記が従来法，下記が提案法の値となっている．雑音が変化する場合の入力 SNR は変化前が 6dB で変化後が 2dB となっている．

SNR_{seg} について見てみると，表 1 では Babble（3dB），Car（3,9dB），表 2 では Babble → Car，表 3 では Babble → Babble，Babble → Car の場合について改善が見られた．LSD に関しては，表 2 及び表 3 の Babble → Car において改善が見られた．$\bar{\varepsilon}$ に関してはあまり改善が見られなかった．音声波形の復元という意味では SNR_{seg} が重要な指標であるので，雑音の種類によっては提案法によりノイズキャンセラとしての性能が改善されたと言える．

表 1: 雑音スペクトル推定と音声品質評価：雑音変化無し（上記が従来法 [13]，下記が提案法）

	$\bar{\varepsilon}$		SNR_{seg}		LSD	
入力 SNR	3	9	3	9	3	9
White	-3.84	-2.69	7.36	11.9	0.422	0.355
Babble	-2.17	-1.67	6.59	12.2	0.313	0.224
Car	-3.25	-1.84	9.71	14.1	0.271	0.208
White	-3.87	-2.95	7.32	11.9	0.431	0.359
Babble	-2.59	-1.52	7.19	12.2	0.315	0.244
Car	-2.82	-2.11	11.12	14.56	0.237	0.201

表 2: 雑音スペクトル推定と音声品質評価：無音区間で雑音変化　（上記が従来法 [13]，下記が提案法）

	$\bar{\varepsilon}$	SNR_{seg}	LSD
Babble → White	-2.97	9.01	0.341
Babble → Babble	-1.98	8.47	0.297
Babble → Car	-2.43	9.56	0.280
Babble → White	-2.96	8.97	0.346
Babble → Babble	-1.73	8.55	0.289
Babble → Car	-1.83	9.89	0.266

表 3: 雑音スペクトル推定と音声品質評価：音声区間で雑音変化　（上記が従来法 [13]，下記が提案法）

	$\bar{\varepsilon}$	SNR_{seg}	LSD
Babble → White	-2.83	9.01	0.336
Babble → Babble	-1.84	8.33	0.291
Babble → Car	-2.46	9.48	0.281
Babble → White	-2.91	8.99	0.343
Babble → Babble	-2.38	8.78	0.287
Babble → Car	-1.94	9.77	0.267

7 まとめ

本稿では，雑音混入音声スペクトルのエントロピーと分散による 2 次平面において傾きのある直線を境界線として用いた VAD を提案し，スペクトルサプレッション法によるノイズキャンセラに適用してその有効性を検討した．その結果，従来の 2 次元平面においてＴ字型の閾値を用いる方法よりも音声／準音声／無音区間の識別性能が向上し，雑音スペクトル推定，及び，音声品質の点でも性能が向上した．しかし，雑音の種類や入力 SNR によって改善されなかった部分もある．今後の課題としては，他の種類の雑音での有効性の検証と，より汎用的な直線のパラメータ（傾きと平行移動）の初期設定と自動制御方法の検討が上げられる．

参考文献

[1] Y.Ephraim and D.Malah, “Speech enhancement using minimum mean-square error short-time spectral amplitude estimator”, IEEE Trans Acoust., Speech, Signal Processing, ASSP-32, 6, pp.1109-1121, Dec.1984.

[2] T.Lotter and P.Vary, “Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-gaussian speech modeling”, Proc. EUSIPCO-04(Vienna，Austria), pp.1447-60, Sep.2004.

[3] M.Katou, A.Sugiyama and M.Serizawa, “Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA”, IEICE Trans.Fundamental, vol.E85-A, no.7, pp.1710-1718, Jul.2002.

[4] R.Martin, D.Malah, V.Cox and J.Accardi , “A noise reduction preprocessor for mobile voice communication”, EURASIP Journal on Applied Signal Processing, pp.1046-1058, Aug.2004.

[5] 大和一洋, 杉山昭彦, 加藤正徳, “Post-processing noise suppresor with adaptive gain-flooring suitable for distorted speech”, 電子情報通信学会 2006 年ソサイエティ大会, 金沢, A-4-20，pp.87, Sep.2006.

[6] I.Cohen and B.Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement”, IEEE Signal Process. Lett. 9(1), 12-15, 2002.

[7] 鈴木大和, 中山謙二, 平野晃宏, “スペクトルサプレッション法における無音区間の検出と雑音スペクトル推定の改善”, 第 21 回信号処理シンポジウム（京都）, C3-2, 2006.11.

[8] K. Nakayama, H. Suzuki and A. Hirano, ”Improved methods for noise spectral estimation and adaptive spectral gain control in noise spectral suppressor,” Proc. IEEE, ISPACS2007, Xiamen, China, pp.97-100, Dec. 2007.

[9] C.Jia and B.Xu, “An improved entropy-based endpoint detection algorithm”, Proc. Int. Sympo. Chinese Spoken Language Processing, pp.1399-1402, Aug. 2002.

[10] B.F.Wu, K.C.Wang, and L.Y.Kuo, “A noise estimator with rapid adaptation in variable-level noisy environments”, Proceeding ROCLING XVI, Taipei, sep.2004.

[11] 東尚哉，中山謙二，平野晃宏，“非定常雑音環境下における VAD を用いた高速雑音推定法，” 第 23 回信号処理シンポジウム（金沢）, P-16, pp170-175, 2008.11.

[12] K. Nakayama, S. Higashi and A. Hirano, ”A noise estimation method based on improved VAD used in noise spectral suppression under highly non-stationary noise environments,” EUSIPCO 2009, Glasgow, Scotland, pp.2494-2498, Aug. 2009.

[13] Sarayut Tungpontawee, K. Nakayama, A. HIrano, ”A noise spectral estimation method based on 2-dimension dynamically VAD used in noise spectral suppression,” 25th Signal Processing Symposium, Nara, P1-12, pp.256-261, Nov. 2010.