社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

バイラテラルフィルタによる実雑音下音声認識のための音声特徴量抽出

山田馨士朗† 滝口 哲也†† 有木 康雄††

† 神戸大学大学院工学研究科 〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 神戸大学自然科学系先端融合研究環 〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †yamada@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし 本稿では，音声のスペクトログラムに対してバイラテラルフィルタを適用することにより，発話のフォルマント遷移情報を保存つつ音声特徴量抽出を行う手法を提案する．実雑音環境下での単語認識実験により，メルフィルタバンク出力だけでなく，MFCC の出力等，どの段階で バイラテラルフィルタを適用すればより効果的な改善をはかることが可能か，また，パワースペクトルやΔとの組み合わせの有効性に関して検証している．

キーワード バイラテラルフィルタ，特徴量抽出，雑音下音声認識

Feature Extraction with Bilateral Filter for Real Environment Speech Recognition

Yamada KEISHIRO†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of Engineering, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
†† Organization of Advanced Science and Technology, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
E-mail: †yamada@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract In this paper, we propose a method to reduce the noise component superimposed on speech with keeping the formants and their transition geometrically by applying the bilateral filter on mel-frequency spectrogram. From the word recognition experiments under the real noisy environment, we clarify that the significant improvement is achieved when the bilateral filter is applied on the mel-frequency spectrogram only once and resultant MFCC, energy and their Δ are utilized as the acoustic parameters.

Key words bilateral filter, feature extraction, noisy speech recognition

1. はじめに

MFCC(Mel-Frequency Cepstrum Coefficient) は，短時間のメル周波数対数スペクトルを直交変換したものであり， 音声認識において代表的な特徴量として用いられている．しかし，短時間スペクトルの特徴から，MFCC には時間方向の変動情報が欠落している．この時間方向の変動情報を特徴量に組み込んで認識率を改善するために，MFCC の線形回帰係数Δ，ΔΔがよく用いられる [1]．しかし，これらは MFCC の微分成分であることから，雑音環境下での音声認識においては効果を落としてしまう．

音声認識における重要な要素のひとつに，時間-周波数平面上のフォルマント遷移がある．この情報は発話内容の音素情報を有しているとされ，これを的確に捉えることによって音声認識率改善ができると考えられている．時間-周波数領域におけるこのような情報は，雑音により容易に歪んでしまうため，結果としてフォルマント遷移情報を捉えることが困難となり，音声認識率が低下してしまう．

これらの問題を解決するためには，MFCC を生成する前のスペクトル領域において，雑音を抑制する必要がある．図 1 に「democrats」という単語を発話したときの

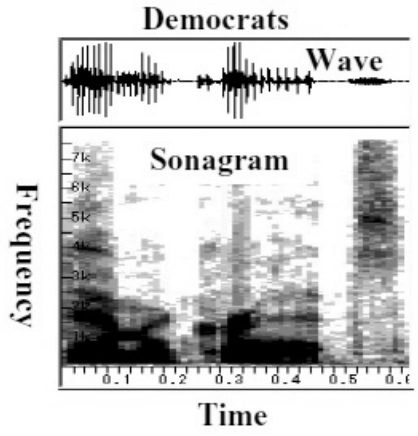

図 1　音声信号とスペクトルの例

時間波形とスペクトルを示す．低い周波数の領域においては複数のフォルマントが見られ，高い周波数の領域では破裂音のスペクトルが見られる．これより，振幅スペクトルが大きいフォルマント遷移や破裂成分は保ちつつ，その値が小さい雑音成分を抑制することが効果的な雑音除去手法となる [2].

雑音下音声認識における代表的な手法として，スペクトルサブトラクション法 [3]〜[7] がある．これは，雑音重畳音声から推定した雑音スペクトルを減算することで雑音を抑制する手法であり，白色雑音のような定常的な雑音では効果が高い．しかし，スペクトル減算によりフォルマント遷移などの重要な情報が欠落し，ミュージカルノイズを引き起こす．また，対数スペクトルの時系列に帯域通過フィルターをかけることにより雑音を除去し，スペクトルの重要な成分のみを選択的に抽出する変調スペクトル [8] なども提案されている．しかし，この方法は，フォルマントやフォルマント遷移のような認識にとって重要な情報を，幾何的に直接抽出する方法とはなっていない．

雑音を除去する代表的な手法にガウシアンフィルタがある．しかし，通常ガウシアンフィルタによって平滑化を行うと，雑音は除去できるが，同時に変化の大きい部分も平滑化してしまい，いわゆるぼやけた情報を生成してしまう．音声の時間-周波数平面上ではフォルマント遷移の情報を失ってしまうことにあたる．そこで，このような変化の大きいところは情報として保存し，変化の小さいところを雑音として除去するために，バイラテラルフィルタ [9]〜[13] が提案されている．

本稿では，音声のどの時点でバイラテラルフィルタを適用することにより，発話のフォルマント遷移情報を保存つつ，音声特徴量抽出を行う手法を提案する．雑音環境下での単語認識実験により，メルフィルタバンク出力だけでなく，MFCC の出力等，どの時点でバイラテラルフィルタを適用すればより効果的な改善をはかることが可能か，また，パワーや Δ との組み合わせの有効性に関して検証する．　以降の 2 章でバイラテラルフィルタについて述べ，3 章ではバイラテラルフィルタを用いた音声特徴量抽出の手法について述べる．4 章で，評価実験の条件とその結果を報告し，最後に 5 章で，結論と今後の課題について述べる．

2. バイラテラルフィルタ

バイラテラルフィルタは次の (1) 式で与えられる．

$$f_i = \frac{\sum_{j \in J_n} w(i,j) d_j}{\sum_{j \in J_n} w(i,j)}, \tag{1}$$

$$w(i,j) = w_x(x_i, x_j) w_d(d_i, d_j) \tag{2}$$

ここで，x_j は，スペクトル平面上のある点 j を表しており，フレーム番号，周波数番号を 2 次元の座標として要素にもつベクトルである．d_j は，スペクトル平面上のある点 j における，対数パワースペクトル値を表す．f_i はスペクトル平面上のある点 i におけるバイラテラルフィルタの出力値である．J_n は $n \in \mathbf{N}$, i について $\| x_i - x_j \|^2 \leqq n$ を満たすの点の集合である．(2) 式は重み関数であり，次の (3) 式，(4) 式の乗算の形で表される．

$$w_x(x_i, x_j) = \frac{1}{\sqrt{2\pi\sigma_x}} e^{-\frac{\|x_i - x_j\|^2}{2{\sigma_x}^2}} \tag{3}$$

$$w_d(d_i, d_j) = \frac{1}{\sqrt{2\pi\sigma_d}} e^{-\frac{\|d_i - d_j\|^2}{2{\sigma_d}^2}} \tag{4}$$

ここで，σ_x と σ_d は，重みに関するパラメータであり，これらは，適用するスペクトル平面の大きさにより次の式で求められる [14].

$$\sigma_x = \ min(\text{ 時間長, 周波数範囲 })/16 \tag{5}$$

$$\sigma_d = \ (\text{対数パワースペクトルの最大値} - \text{対数パワースペクトルの最小値})/10 \tag{6}$$

ガウシアンフィルタの場合は (2) 式で表されるような重み関数を，$w(i,j) = w_x(x_i, x_j)$ として，ベクトル間の幾何学的距離のみに基いて決定するため，エッジの有無に関わらず平滑化してしまう．一方，バイラテラルフィルタは幾何学的な距離の差だけでなく，2 つの点の対数パワースペクトル値の差を考慮に入れているため，エッジを保存しつつ細かいノイズを取り除くことができる．

例として，図 2 に示すスペクトル平面上でガウシアンフィルタを適用したものを図 3 に, バイラテラルフィルタを適用したものを図 4 に示す.

3. 提 案 手 法

図 5 に，フォルマント遷移を保持しつつ雑音を抑制する提案手法の流れを示す．まず，音声信号を短時間フーリエ変換 (STDFT) し，メル尺度の周波数軸上でフィル

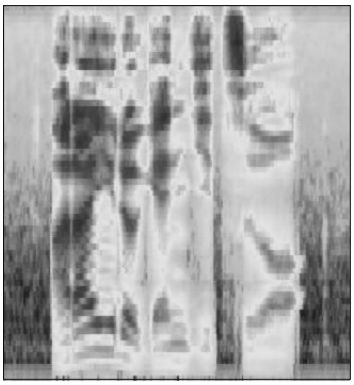

図 2　元スペクトル

図 3　ガウシアンフィルタ適用後

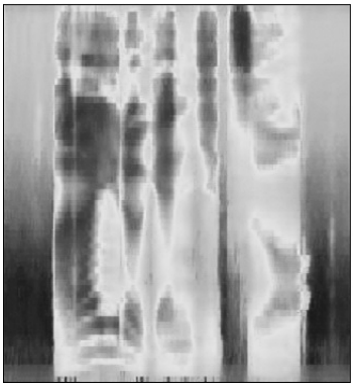

図 4　バイラテラルフィルタ適用後

タバンク分析を行うことにより，メルフィルタバンク出力を得る．このメルフィルタバンク出力にバイラテラルフィルタを適用する．さらに対数をとり，離散コサイン変換 (DCT) を行うことにより，バイラテラルフィルタを含めた MFCC 特徴量を得る．この手法を提案手法 1 とする．

提案手法 2，提案手法 3 は提案手法 1 のバリエーションであり，その処理の流れを図 6，図 7 に示す．提案手法 2 は，2 回のバイラテラルフィルタをメルフィルタバンク出力に適用することにより，より強い平滑化を行っているものである．提案手法 3 は，バイラテラルフィルタを MFCC の後に適用したものである．この手法はケ

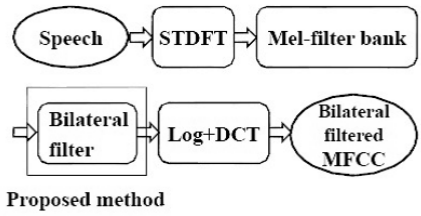

図 5　提案手法 1 の流れ

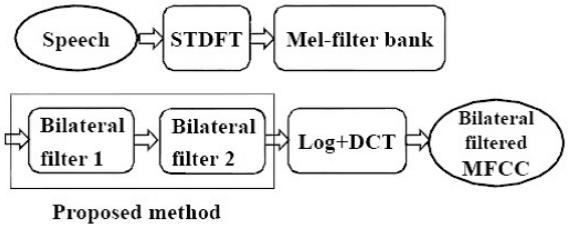

図 6　提案手法 2 の流れ

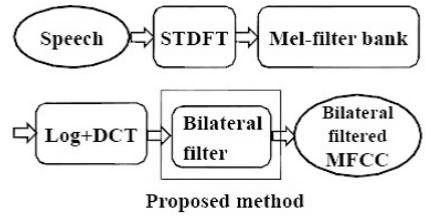

図 7　提案手法 3 の流れ

プストラムに対してバイラテラルフィルタを用いているため，スペクトル上のフォルマント遷移情報を保存することは保証していない．これらの手法の優劣については 4 章の評価実験で示す．

4. 評価実験

4.1 実験条件

提案手法の評価を行うために，単語音声認識実験を行った．男性 5 名，女性 5 名の計 10 名の話者が発声したラベル付き音声データベース (ATR 音素バランス文 A セット) に，CENSREC-1-C データベース [15] に収録されている雑音を重畳したものを音声データに用いた．ATR 音素バランス文 A セットに収録されている音声データの標本化周波数は 20kHz であるが，CENSREC-1-C は標本化周波数 8kHz で収録されているため，音声信号は 8 kHz にダウンサンプリングして使用した．各話者に対して，学習データとして 2,620 単語，評価データとして学習データに使用していないデータ 1,000 単語を用いた．

音声信号は，フレーム幅，シフト幅をそれぞれ，25ms, 10ms とした短時間フーリエ変換によりスペクトルに変換される．そこから 64 次元メルフィルタバンク分析を行い，その出力に対して提案手法を適用する．提案手法を実環境雑音で評価するために，CENSREC-1-C データベース [15] に収録されている食堂内 (restaurant) と高速

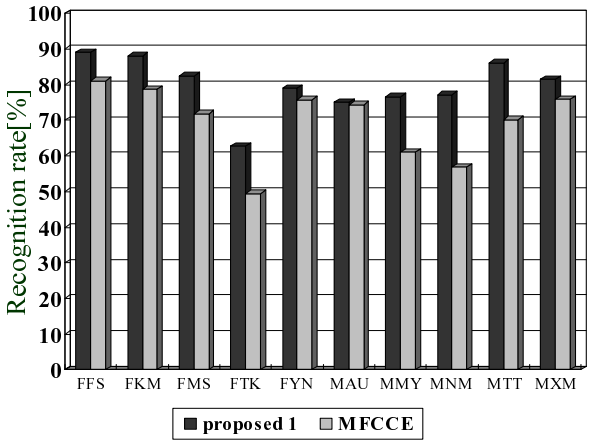

図 8　提案手法 1 による実験結果

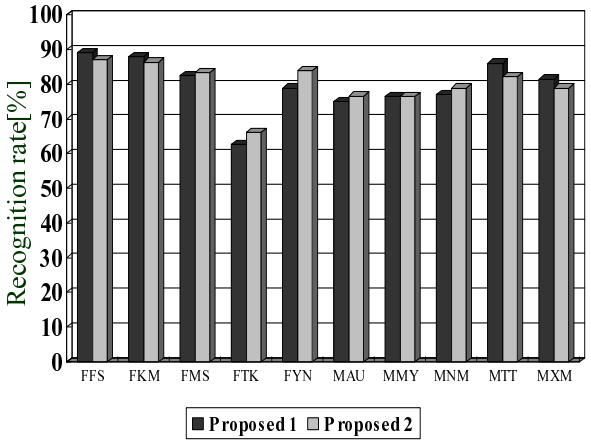

図 9　提案手法 2 による実験結果

道路付近 (street) の-5≦SNR≦10 dB である 2 種類の雑音を用いる．音素数は 54 音素，音響モデルは HMM(5 状態，8 混合) を用いた．

4.2 実　験　1

まずはじめに，どの段階でバイラテラルフィルタを適用すれば最も効果的であるかを調べるため，3 章で述べた提案手法 1，提案手法 2，提案手法 3 について実験を行った．結果をそれぞれ図 8，図 9，図 10 に示す．ここでは，レストランの実環境雑音を重畳した音声データを用いている．図 8 では，10 人の話者ごとに，MFCC12 次元+パワースペクトル (MFCCE と呼ぶ) の 13 次元の特徴量を用いた場合に，提案手法 1 と通常の MFCC を比較している．ここで，提案手法 1 を用いた場合，通常の MFCC でも高い認識率を示した FFS，FKM，MTT の話者だけではなく，FTK，MMY や MNM のような認識率が低かった話者も含めたすべての話者について，認識率の改善がみられた．

図 9 では，バイラテラルフィルタをメルフィルタバンク出力に 2 回適用した提案手法 2 と提案手法 1 の認識率の比較を行った．結果は FTK など認識率が低い話者で改善されているものも見られたが，全体的にほとんど同

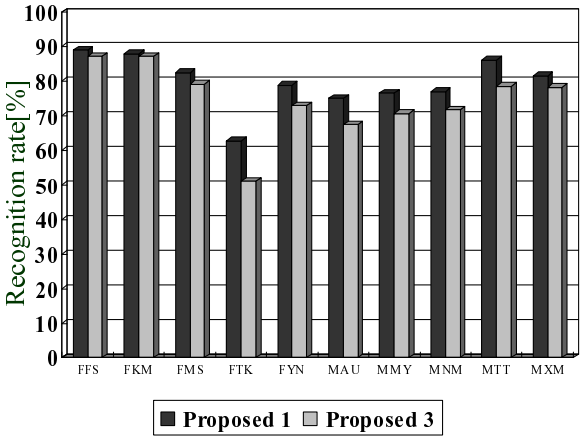

図 10　提案手法 3 による実験結果

表 1　話者 10 人の平均認識率 (%)

		MFCC+E (13dim)
Restaurant noise	Baseline	69.5
	Proposed 1	79.7
	Proposed 2	78.0
	Proposed 3	74.4
Street noise	Baseline	74.9
	Proposed 1	84.1
	Proposed 2	82.5
	Proposed 3	80.6

じであった．これより，2 回のバイラテラルフィルタの適用は，それほど効果的ではないといえる．

図 10 では，離散コサイン変換 (DCT) により生成された MFCC に対して，バイラテラルフィルタを適用した提案手法 3 と提案手法 1 の比較を行った．結果は，10 人全員の話者に対して，提案手法 3 の結果が，提案手法 1 の結果を下回った．これより，MFCC に対するバイラテラルフィルタの適用は物理的な根拠がなく，認識率の改善については提案手法 1 が有効であるといえる．

高速道路付近で収録された実環境雑音 (street) を重畳した音声データに関しても，ほとんど同じ傾向がみられた．レストランの雑音と高速道路付近の雑音を重畳した音声データについて，ベースラインである通常の 13 次元 MFCCE と，3 つの提案手法による行った認識実験について，10 人の話者の平均認識率を表 1 にまとめた．ベースラインに比べて，提案手法 1 が平均的に他の手法よりよい結果を示していることがわかる．ベースラインに比べて約 10 ポイントの改善が見られた．

4.3 実　験　2

次に，どのような特徴量に対して提案手法がより有効か調べる．ここでは次の 3 つの音声特徴量をベースラインとする．

- 12 次元 MFCC+パワースペクトルの 13 次元 (MFCCE)
- 13 次元 MFCCE とその Δ の 26 次元 (MFCCE+Δ)
- 13 次元 MFCCE とその Δ，ΔΔ39 次元

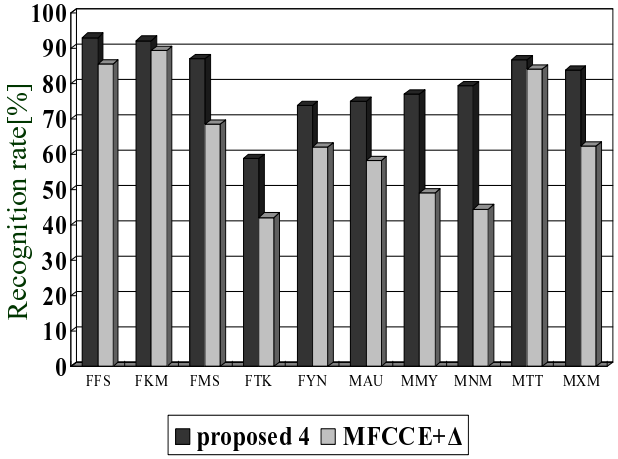

図 11　提案手法 4 による実験結果

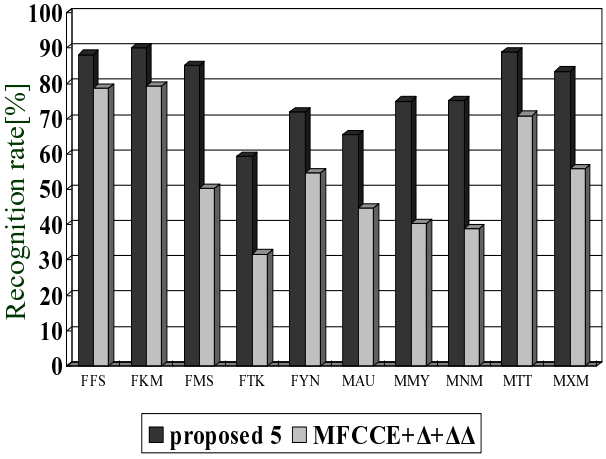

図 12　提案手法 5 による実験結果

(MFCCE+Δ+ΔΔ)

バイラテラルフィルタは実験 1 の結果より最も効果的であった提案手法 1 の位置で適用することにする．1 つ目の 13 次元 MFCCE に関する比較は前節の図 8 に示した通りである．図 11 に，メルフィルタバンク出力にバイラテラルフィルタを適用した 26 次元 MFCCE+Δ を提案手法 4 として，ベースラインとの比較を示す．結果は 10 人すべての話者について，ベースラインから認識率の改善が見られた．特に認識率の低かった FTK,MMY, MNM の話者については大きな改善が見られた．

図 12 に，メルフィルタバンク出力にバイラテラルフィルタを適用した 39 次元 MFCCE+Δ+ΔΔ を提案手法 5 として，ベースラインとの比較を示す．結果はベースラインの MFCC はレストランの雑音に対して大きく認識率を落としたが，バイラテラルフィルタを用いた提案手法 5 では，それを大きく改善することができた．

高速道路付近の雑音に対してもほぼ同様の傾向が見られた．表 2 に，レストランの雑音 (Restaurant noise) と高速道路の雑音 (Street noise) について，MFCC13 次元, MFCCE+Δ26 次元，MFCCE+Δ+ΔΔ39 次元の 3 つに関するベースラインと提案手法を適用した場合，そして，メルフィルタバンク出力に対して，バイラテラルフィルタではなくガウシアンフィルタを適用した場合についての認識率をまとめた．

結果を見ると，ベースライン，ガウシアンフィルタを用いて平滑化を行った手法よりも，バイラテラルフィルタを用いた提案手法において，いづれの特徴量においても高い認識率を見ることができる．レストランで収録した雑音を用いた実験で，提案手法 4 に関してはベースラインに対し，16 ポイント，ガウシアンフィルタを用いた手法に対しては 0.8 ポイントの改善，提案手法 5 に関しては，ベースラインに対し，24.1 ポイント，ガウシアンフィルタを用いた手法に対しては，0.3 ポイントの改善が見られた．

表 2　話者 10 人の平均認識率 (%)

		MFCCE (13dim)	MFCCE+Δ(26dim)	MFCCE+Δ++ΔΔ(39dim)
Restaurant noise	Baseline	69.5	64.6	54.5
	Proposed	79.7	**80.6**	78.6
	Gaussian	76.4	79.8	78.3
Street noise	Baseline	74.9	80.1	79.5
	Proposed	84.1	**86.9**	84.3
	Gaussian	80.75	83.7	83.1

4.4 考　　察

実験 1 に関して，3 つの提案手法の中では，提案手法 1 が最も効果的であるという結果が得られたが，いづれもベースラインである MFCC 単体よりは高い認識率を示した．これよりバイラテラルフィルタを時間-周波数スペクトル上に適用することによって，雑音環境下での音声認識率の改善に効果があることがわかった．特に提案手法 2 の 2 回バイラテラルフィルタをかける手法は，1 回では認識率が低かった話者に対して効果があることが分かった．これは雑音の平滑化が不十分だった音声データに対して，さらにもう一度平滑化を行うことで，十分な雑音抑制効果が得られ，認識率が向上したものと考えられる．このように十分な平滑化が一度で，もしくは複数回で行えるように，自動的にパラメータや回数を決定する仕組みが分かれば，さらなる改善につながるものと考えられる．

実験 2 について，特にレストランで収録された実環境雑音に対して，ベースラインでは著しい音声認識率の低下が見られた．人間の話声など不規則な雑音は，スペクトルサブトラクション法でも推定が困難なため，うまく効果が表れない．それに対し，提案手法は雑音の推定を行う手法ではないため，このような不規則な実環境雑音に対しても認識率の改善を図ることができた．ガウシアンフィルタと比較すると，バイラテラルフィルタを用いた提案手法との差は僅差である．今後両手法のパラメー

タの調整を行って，差違を検討する予定である．

5. おわりに

本稿では，バイラテラルフィルタをメルフィルタバンク出力に対して適用し，フォルマント遷移情報を保ちつつ，雑音を抑制して音声特徴量を求める手法を提案した．実雑音環境下での音声認識を効果的に行うためにバイラテラルフィルタの適用が有効であると考えられる．今後は，スペクトルサブトラクション等，他手法との比較や，組み合わせを通じて，バイラテラルフィルタが，実雑音環境下音声認識において，より効果的に作用するような手法を考えていく予定である．

文　献

[1] S.Furui,“ Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum ”, IEEE Trans. on ASSP, 34, pp.52-59, 1986.

[2] Y.Ariki, K.Kajimoto and T.Sakai, “ Acoustic Noise Reduction by Two Dimensional Spectral Smoothing and Spectral Amplitude Trasnformation ”, ICASSP ’86, pp.97-100, 1986.

[3] S.F.Boll,“ Suppression of acoustic noise in speech using spectral subtraction ”, IEEE Trans. ASSP, 29, pp.113-120, 1979.

[4] V. Stahl, A. Fischer, R. Bippus, “ Quantile based noise estimation for spectral subtraction and Wiener 6ltering ”, Proceedings of the 25th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-00, Istanbul, Turkey, pp. 1875-1878, 2000.

[5] I. Cohen,“ On speech enhancement under signal presence uncertainty, Proceedings of the 26th IEEE International Conference on Acoustics Speech ”, and Signal Processing, ICASSP-01, Salt Lake City, Utah, 2001.

[6] R. Martin,“ Spectral subtraction based on minimum statistics ”, Proceedings of the Seventh European Signal Processing Conference, EUSIPCO-94, Edinburgh, Scotland, 1994, pp. 1182-1185.

[7] S. Kamath, PC. Loizou,“ A multi-band spectral subtraction method for enhancing speech corrupted by colored noise ”. student research abstracts of Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. 2002.

[8] H. Hermansky and N. Morgan,“ RASTA Processing of Speech ”, IEEE Trans. of Speech and Audio Processing, 2(4), pp.578-589, 1994.

[9] C.Tomashi, R.Manduchi,“ Bilateral filtering for gray and color images ”, ICCV, pp.839-846, 1998.

[10] D. Barash,“ A fundamental relationship between bilateral Filtering, adaptive smoothing and nonlinear diffusion equation ”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24(6), pp. 844-850, 2002.

[11] F. Durand, and J. Dorsey,“ Fast bilateral Filtering for the display of high-dynamic range images ”, ACM Transactions on Graphics, special issue on Proc. of ACM SIGGRAPH 2002, San Antonio, Texas, vol. 21(3), pp. 249-256, 2002.

[12] M. Elad,“ On the bilateral Filter and ways to improve it ”, IEEE Transaction Image Processing, vol. 11(10), pp. 1141-1151, 2002.

[13] D. Barash, D. Comanichiu,“ A common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean shift ”, Image and Video Computing 22, 1, 73-81，2004.

[14] S.Paris, F.Durand,“ A Fast Approximation of the Bilateral Filter using a Signal Processing Approach ” ECCV 2006.

[15] N. Kitaoka, K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura,“ Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance ”, ASRU, pp.607-612, 2007.