バイラテラルフィルタによる 雑音重畳音声の認識効果に関する検討*

☆山田馨士朗，滝口哲也，有木康雄 (神戸大)

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，短時間分析を行うため，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．

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 [2] 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 [3] 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 [4] が提案されている．

本稿では，音声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ことにより，発話のフォルマント遷移情報を保存しつつ，音声特徴量抽出を行う手法を提案する．また雑音環境下での単語認識実験において、スペクトルサブトラクション，ガウシアンフィルタ等の手法との比較や組み合わせから提案手法の効果を検討する．

2 バイラテラルフィルタ

バイラテラルフィルタを次の (1) 式に示す．

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

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

f_i は，時間-メル周波数領域上のある点 i に対するバイラテラルフィルタ出力である．また x_j は時間-メル周波数領域上のある点 j の (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．点 d_j はある点 j の対数パワースペクトル値を表す．J_n は $n \in \mathbf{N}$ について，点 i を中心とする，$(2n-1) \times (2n-1)$ の正方形に含まれる点の集合である．(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．

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

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

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が，$w(i,j) = w_x(x_i, x_j)$ と，ベクトル間の幾何学的距離のみに基いて決定されるため、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重みの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，雑音が重畳された平面にバイラテラルフィルタを適用したとき，変化の差が大きいところは保存したまま細かな雑音のみが平滑化される様子を Fig.1 に示す．

3 提案手法

本論文の提案手法は，前章で紹介したバイラテラルフィルタを，時間-周波数平面における 64 次元メルフィルタバンクに適用する．その時間-周波数平面上でフォルマン遷移を保存しつつ，雑音除去を行えば，スペクトルが平滑化され雑音が抑圧された特徴量に対して認識を行うことができると考えられる．提案手法の流れを FIg.2 に示す．

*Effectiveness of bilateral filter for noisy speech recognition. by YAMADA, Keishiro, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe University)

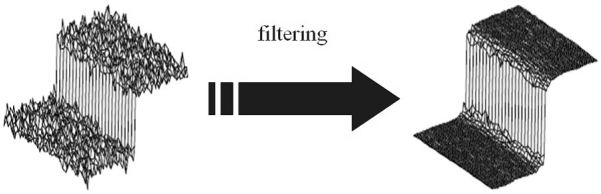

Fig. 1　バイラテラルフィルタによる平滑化

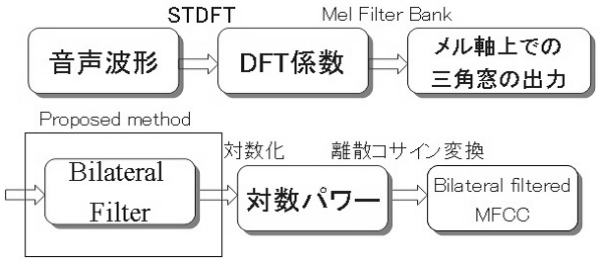

Fig. 2　提案手法の流れ

4　評価実験

提案手法の効果を評価するために単語認識実験を行った．音声データは男女 10 名の話者が発声したラベルつき音声データベース（ATR 音素バランス文 A セット）を用いた．データ数は各話者，学習データに 2,620 単語，評価データに学習に使用していないデータ 1,000 単語，音素数は 54 音素，音響モデルは HMM(5 状態，8 混合) を用いた．雑音環境として CENSREC-1-C データベースの食堂内，高速道路付近で収録された音声の無音声部分の雑音を使用し音声データに重畳した．SNR は 10〜20dB である．音声特徴量は MFCC12 次元にパワーを加えた MFCCE(13 次元) の $\Delta, \Delta\Delta$ をとった，MFCCE+Δ + $\Delta\Delta$(39 次元) を用いた．バイラテラルフィルタを用いないものをベースライン (Baseline) とし，提案手法 (Proposed) と比較を行う．またスペクトルサブトラクション法 (SS)，バイラテラルフィルタでなくガウシアンフィルタを用いた場合 (Gaussian) と比較を行っている．結果は 10 人の話者の平均認識率を Table.1 に示す．レストラン雑音，高速道路雑音ともに，最も高い認識率を示した手法は提案手法を用いたものであった．

5　おわりに

本論文ではバイラテラルフィルタによる雑音除去の音声認識に関する効果について，他手法との比較を行った．今後の課題として，より雑音に対し頑健な音声認識を実現するようフィルタの改良，パラメータの調整，そして雑音を平滑化するだけでなく，劣化したフォルマントを際立たせるような仕組みを検討していくことがあげられる．

Table 1　話者 10 人の平均認識率 (%)

		MFCCE+Δ+ΔΔ
Restaurant noise	Baseline	54.5
	Proposed	**78.6**
	SS	68.6
	SS+Proposed	78.3
	Gaussian	78.3
Street noise	Baseline	79.5
	Proposed	**84.3**
	SS	81.1
	SS+Proposed	81.9
	Gaussian	83.1

参考文献

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

[2] 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, 1979.

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

[4] C. Tomashi, R. Manduchi,“ Bilateral filtering for gray and color images, ”In Proceedings of the International Conference on Computer Vision, pp.839-846, 1998.

[5] S. Paris and F. Durand, “ A fast approximation of the bilateral filter using a signal processing approach, ” In ECCV, volume 4, pp.568-580, 2006.

[6] 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.