

# 雑音環境下音声認識のための バイラテラルフィルタを用いた音声特徴量抽出\*

◎山田馨士朗, 滝口哲也, 有木康雄 (神戸大)

## 1 はじめに

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というものがある。この情報は発話内容の音素や発音の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の改善が望まれると考えられている。雑音環境下での音声認識における問題として、時間-周波数領域のこのような情報は、雑音により容易に歪まされてしまうということがある。結果としてうまくフォルマント遷移情報を捉えることが困難になり、音声認識率は下がる。

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 [1], 変調スペクトルの重要な成分のみを選択的に用いた手法 [2] などが提案されているが、本論文では平滑化フィルタを用いた手法を提案する。平滑化フィルタの代表的なものにガウシアンフィルタがあるが、通常ガウシアンフィルタによって平滑化を行うと、細かな雑音は除去できるが、同時に輪郭のようなエッジ部分も平滑化してしまい、いわゆるぼやけた画像を生成してしまう。これは、音声の時間-周波数平面上だとフォルマント遷移の情報を失ってしまうことにあたる。そこでこのようなエッジ情報の劣化を防ぐため、C.Tomasi らによりバイラテラルフィルタ [3] というものが提案されている。バイラテラルフィルタはエッジ情報を保存しつつ平滑化を行うことができるので、これを時間-周波数領域に適用し、フォルマント遷移を失わないように雑音を除去して音声特徴量抽出を行う。最後に数値実験により、本手法の雑音下音声認識への有効性を示す。

## 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。

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

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

ここで、 $f_i$  はあるフレーム番号、周波数番号をもつスペクトル上の点  $i$  を入力とする出力値、 $x_j$  はある点  $j$  のフレーム番号、周波数番号を 2 次元の座標と

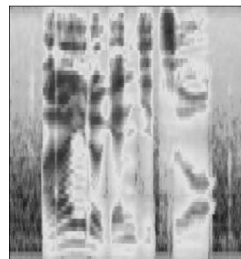


Fig. 1 Original image



Fig. 2 Gaussian filter

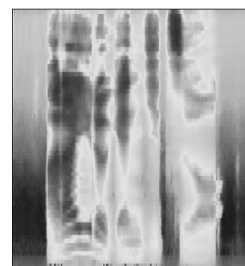


Fig. 3 Bilateral filter

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  $n \in \mathbf{N}, i$  について  $\|x_i - x_j\|^2 \leq n$  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。

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

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

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、Fig. 1 にガウシアンフィルタを適用したものを Fig. 2、バイラテラルフィルタを適用したものを Fig. 3 に示す。

\* Feature extraction using bilateral filter for noisy environment speech recognition. by YAMADA, Keishiro, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe University)

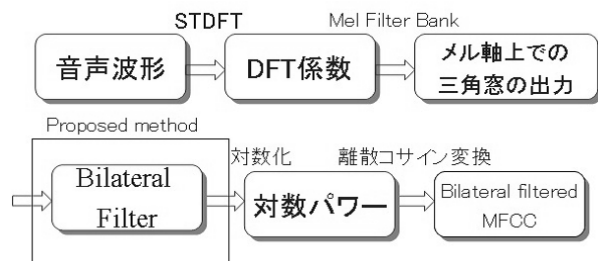


Fig. 4 提案手法の流れ

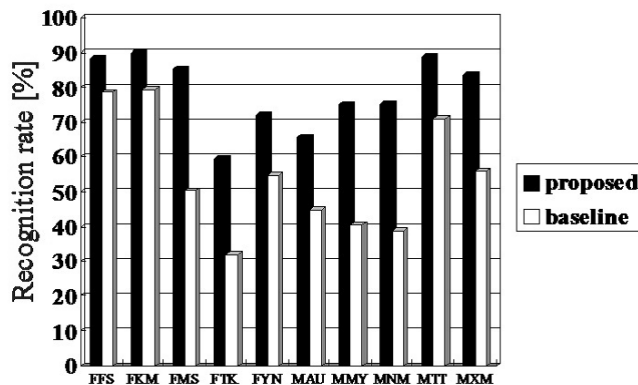


Fig. 5 各話者の認識率

### 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波形で認識を行うことができると考える。提案手法の流れをFig.4に示す。

### 4 評価実験

#### 4.1 実験概要

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。音声特徴量はベースラインとして、MFCC+ $\Delta$ + $\Delta\Delta$ (39次元)と提案手法で比較を行う。結果はそれぞれの話者ごとの結果をFig.5に示し、10名の話者の平均認識率をTable.1に示す。

Table 1 話者10名の平均認識率

|            |          | MFCC+ $\Delta$ + $\Delta\Delta$ 39dim |
|------------|----------|---------------------------------------|
| restaurant | proposed | 78.6%                                 |
| noise      | baseline | 54.5%                                 |
| street     | proposed | 84.3%                                 |
| noise      | baseline | 79.5%                                 |

#### 4.2 結果・考察

Fig.5でレストランの雑音の場合のみしか示していないが、高速道路付近の雑音に対しても同様に、すべての話者に対して提案手法を適用したほうに認識率の向上がみられた。元々認識率の高いFFSやFKMという話者に対しては10ポイントほどの改善が見られたが、FTK, MMY, MNMという認識率の低い話者に対しては2倍近くにもなる大幅な改善を見ることができた。これより、バイラテラルフィルタが認識率の低下を引き起こす雑音成分をうまく取り除くよう作用し、認識率の向上に繋がったと考えられる。

### 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が音声認識についても有効であるということについて示した。今後の課題として、スペクトルサブトラクションや以前我々が発表した雑音に頑健な音声特徴量[4]のような、他の雑音下音声認識に関する手法との比較、また、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を検討していくことがあげられる。

### 参考文献

- [1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-27, no.2, 113-120, 1979.
- [2] 金寺 登 他, "変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識", 電子情報通信学会論文誌 D-II Vol. J84-D-II, no.7, 1261-169, 2001.
- [3] C. Tomashi, R. Manduchi, "Bilateral filtering for gray and color images", In Proceedings of the International Conference on Computer Vision, 839-846, 1998.
- [4] Takashi Muroi, Tetsuya Takiguchi, Yasuo Ariki, "Speaker Independent Phoneme Recognition Based on Fisher Weight Map", The 2nd International Conference on Multimedia and Ubiquitous Engineering (MUE2008), 253-257, 2008.