電子情報通信学会論文誌 '99/11 Vol. J82–D–II No. 11

スペクトル傾斜に着目した音声認識のための特徴抽出

船田　哲男†（正員）　　続木　貴史†*（准員）

Feature extraction based on spectral slope for speech recognition

Tetsuo FUNADA†, *Member* and

Takashi TSUZUKI†*, *Associate Member*

† 金沢大学工学部，金沢市

Faculty of Engineering, Kanazawa University, Kanazawa-shi, 920–8667 Japan

* 現在，松下電器産業株式会社

あらまし　耐雑音性をもった特徴を音声から抽出することを目的とし，パワースペクトルの周波数軸に関する傾斜に着目した特徴量を提案した．この特徴抽出過程で行うスペクトル傾斜に対するしきい値操作が，耐雑音性に効果をもつことを，単語音声認識実験で確認した．

キーワード　スペクトル傾斜，FTTSS，耐雑音性，しきい値処理

1. まえがき

音声の認識に用いられる代表的な特徴量として，メル LPC ケプストラムや FFT ケプストラムなどが挙げられるが，雑音がある環境下でこれらの特徴を利用して音声認識すると，認識率が著しく低下する．本論文では，耐雑音性をもった特徴を抽出するため，帯域フィルタ対（BPFP: Band Pass Filter-Pair）バンクを利用した方法を提案する．この方法は，音声認識のための特徴を，周波数帯域の複数個の周波数点上でパワースペクトルの傾きを表現する量から抽出しようという考えに基づいている [5], [6].

スペクトル傾斜に基づいた特徴と音韻性との関連を調べ，その有効性を調べた実験例が報告されているが [1]，本研究では，スペクトル傾斜にしきい値を導入して，雑音に対する頑健性をもたせることを特徴としている．その有効性を調べるため，HMM を用いた不特定話者単語音声認識実験を行った．また，耐雑音性を評価するため，メル LPC ケプストラム [2], [3] を特徴として用いた場合との比較実験も行った．

2. 特徴量 FTTSS

使用する帯域フィルタ対 BPFP は図 1 に示すように，一対の帯域フィルタ（BPF）$H_c^+(z)$, $H_c^-(z)$ で構成される．

各フィルタの伝達関数 $H_c^{\pm}(z)$ は

$$H_c^{\pm}(z) = \frac{1-e^{-2\pi bT}\cos(2\pi(f_c \pm \Delta f)T)z^{-1}}{1-2e^{-2\pi bT}\cos(2\pi(f_c \pm \Delta f)T)z^{-1}+e^{-4\pi bT}z^{-2}} \quad (1)$$

で表される．ここで，T(0.1 ms) はサンプリング間隔，$f_c \pm \Delta f$ は各フィルタの中心周波数，$2\Delta f(\Delta f = 15\,\mathrm{Hz})$ はフィルタ対の中心周波数間隔，b は BPF の帯域幅である．

音声信号はそれぞれ式 (1) で表される BPF に入力され，各 BPF の出力の絶対値の差がしきい値 +SH より大きければ +1，−SH より小さければ −1 が出力される．この処理により音声スペクトルの周波数 f_c における傾斜方向を抽出することができる．スペクトル値ではなく，その傾斜方向を用いる理由はホルマント帯域幅の影響を受けにくい頑健性をもつことが期待できるからである．また，しきい値による非線形処理を用いることにより，雑音による多少の変動が BPFP の出力に影響を与えないように，傾斜の正負が明確な部分のみを抽出できる．

図 2 に示すように，BPFP の出力 $e_c(n)$ を 30 ms で平均し，複数チャネル用意したものを BPFP バンクと呼んでいる．

本研究では，対象とする周波数範囲を 100〜4,926 Hz とし，64 チャネルの BPFP バンクを構成した．このようにして抽出された 64 次の特徴量をチャネル方向（周波数軸方向）でフーリエ変換したものを FTTSS（Fourier

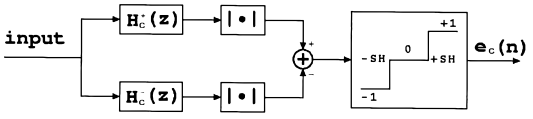

図 1　BPFP のブロック図
Fig. 1　Block diagram of BPFP.

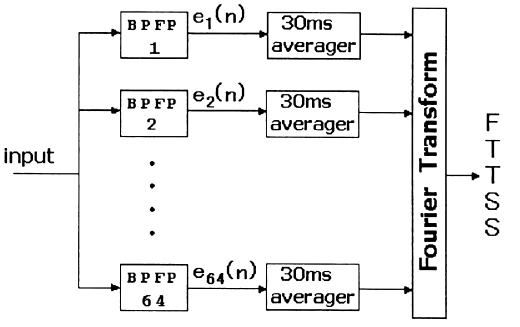

図 2　BPFP バンクのブロック図
Fig. 2　Block diagram of BPFP bank.

Transform of Three-valued Spectral Slope）と呼ぶことにする．また，BPFP の中心周波数をメルスケール上で等間隔にとってフーリエ変換したものをメル FTTSS と呼ぶことにする．周波数軸上のフーリエ変換であるためケプストラムと同様に，FTTSS あるいはメル FTTSS の低次のフーリエ係数が，認識に有効な特徴として利用できる．

3. 実験方法

提案した特徴を用いて HMM による単語音声認識実験を行った．HMM の学習には，認識単語音声とは発声内容の異なる，男女各 20 名ずつの音声（ATR 研究用日本語音声データベース（C2-M01，C2-F01））をテキストデータに従って分けた音節波形を用いた．1 音節当りの HMM 状態数を 7 として各音節ごとのモデルを学習した後，認識対象となる単語の音節系列に合わせて音節モデルを連結して単語モデルを作成した．なお，単語モデルの連結学習は行っていない [4]．

また，認識実験には，学習データの発声話者と異なる男性話者 10 名が無雑音下で発声した都道府県名など 50 種類の単語を用いた．雑音入り音声データの作成には，上に述べた雑音無し単語音声に 3 種類の雑音（走行中の車内雑音，ワークステーションノイズ，幹線道路雑音（（社）日本電子工業振興協会騒音データベース，No.2，No.14，No.9））を，それぞれ $S/N = 0,\ 10,\ 20,\ 30$ dB で付加した．

提案した特徴を比較評価するため，同一データに対し，既存の代表的特徴であるメル LPC ケプストラム分析と FTTSS の両方を用いてそれぞれ認識実験を行った．それぞれの分析におけるパラメータ値を，無雑音環境下での予備実験 [6] により以下のように定めた．まず，メル LPC ケプストラム分析（フレーム長：30 ms，フレーム周期：10 ms）で最も良い認識率を示す分析次数とケプストラム次数を調べた．9〜12 次までの分析によれば，分析次数が 11 次で，ケプストラム次数を 11 次まで用いたときが最も高い認識率を示した．以下ではこの次数による認識結果を示す．

次に，FTTSS を用いた予備実験でも，図 2 における BPFP のしきい値（SH）を 0，出力 $(e_c(n))$ を 30 ms 間で平均する処理を 10 ms ごとに行い，FTTSS の最高次数を 9〜14 次の間で実験したところ，10 次までの FTTSS を用いたときの認識率が最もよくなった．以下では 10 次までのメル FTTSS を用いた実験を行う．

4. 実験結果

4.1 メル LPC ケプストラム

メル LPC ケプストラムを用いて各雑音環境下で認識実験を行った結果の平均認識率を図 8〜図 10 に白抜きの棒グラフに示す．

4.2 零しきい値メル FTTSS

予備実験の結果 [6] より．メル FTTSS を算出するしきい値（SH）を 0 としたとき，すべての BPF で，バンド幅 (b) を 200 Hz に設定すると認識率が最もよくなることがわかった．そこで，SH=0，バンド幅を 200 Hz として，各雑音環境下で認識実験を行った結果の平均認識率をメル LPC ケプストラムと比較して，図 8〜図 10 に凡例一番上の棒グラフで示す．これらの図より，幹線道路雑音環境下の S/N の低いところを除いては，メル LPC ケプストラムの方がメル FTTSS よりもよくなることがわかる．メル FTTSS による認識結果が悪くなる理由を調べるために，一例として男性話者が発声した音節/ア/の BPFP メルスペクトル（バンド幅 200 Hz）とメル LPC スペクトルを図 3 と図 4 に示す．ここで，BPFP メルスペクトルとは，

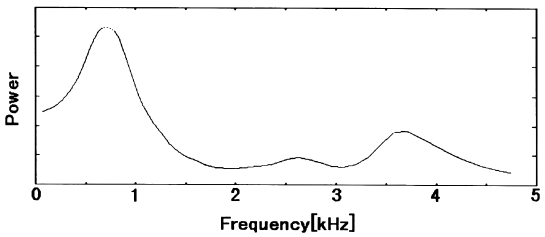

図 3　音節/ア/の BPFP メルスペクトルの一例
Fig. 3　An example of BPFP mel spectrum of syllable /a/.

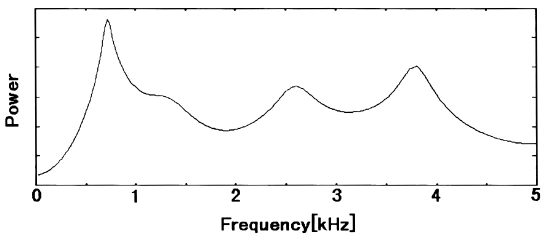

図 4　音節/ア/のメル LPC スペクトルの一例
Fig. 4　An example of mel LPC spectrum of syllable /a/.

（注 1）：近似的に周波数 f_C でのスペクトルに対応した値が得られるので，ここでは BPFP メルスペクトルと呼ぶ．

図 1 において，$H_c^{\pm}(z)$ の出力の絶対値の和をとり，30 ms で平均したものを指す(注1)．

図 3 と図 4 を比較すると，BPFP メルスペクトルにおいては，バンド幅が 200 Hz と広いために第 1 ホルマントと第 2 ホルマントが分離できていないことがわかる．このことがメル FTTSS による認識率が低下する理由と考えられる．そこで，このホルマントが分離できるような周波数分解能に対応するバンド幅を求めるため，バンド幅を 50 Hz まで狭めると図 5 に示すように第 2 ホルマントが分離できる．しかし，図 5 でも若干現れているように，バンド幅を狭くすると高調波（ピッチ）に由来した周波数微細構造による変動が現れる．この変動は認識に悪影響をもたらすのでバンド幅をこれ以上狭くすることはできない．したがって，バンド幅は 50 Hz が適当と思われる．この周波数微細構造による変動と雑音付加による変動を軽減するため，次に非零のしきい値を設定して認識実験を行う．

4.3 非零しきい値メル FTTSS

まず，しきい値の効果を探るために，バンド幅を 200 Hz とし，しきい値を平均振幅（各単語波形の絶対値を単語区間長で平均した値）の 1/40 から 1/2000 の値に設定して無雑音環境下で認識実験を行った．その結果を図 6 に示す．図 6 より，しきい値を適当な値に設定すると 0 のときよりも認識率が改善されることがわかる．しきい値が 0 のときは BPFP の出力値は絶対値によらず，正負のみに応じて +1，−1 に変換するが，非零のしきい値を設けることで BPFP の出力値がある程度以上大きい安定したスペクトルの傾きのみを選択して変換していることになる．したがって，音声本来のパワースペクトルの傾きに由来した特徴が強調され，認識率が改善できたと考えられる．

図 6 より，しきい値が平均振幅の 1/1000 で最も良い認識率を与えていることがわかる．そこで，しきい値 1/1000 を用いて各雑音環境下で認識実験を行った結果を図 8〜図 10 で凡例の 2 番目の棒グラフで示す．これらの図より，ほとんどの雑音環境下において非零しきい値を用いた方が認識率が改善されていることがわかる．これらの結果，無雑音環境下，雑音環境下どちらにおいても，しきい値を設けることが有効であるといえる．

次に，**4.2** で述べたように BPF のバンド幅を 50 Hz とし，しきい値は平均振幅の 1/5 から 1/120 の値を用いて無雑音環境下において実験を行った．結果を図 7 に示す．

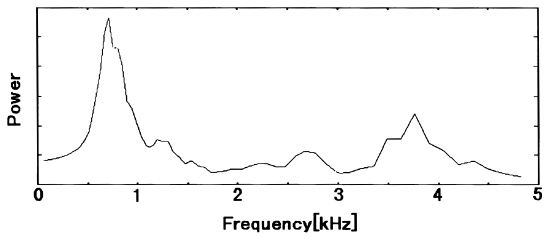

図 5 バンド幅 50 Hz で計算した音節/a/の BPFP メルスペクトルの一例
Fig. 5 An example of BPFP mel spectrum calculated by bandwidth 50 Hz

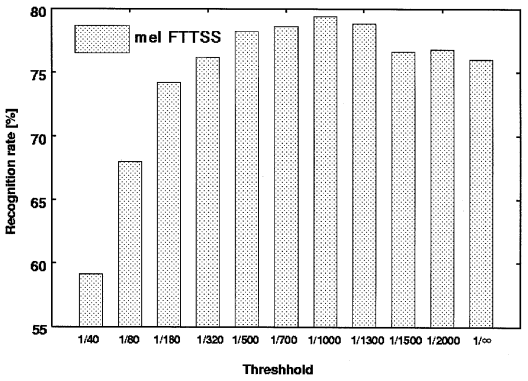

図 6 バンド幅 200 Hz におけるしきい値による認識率の違い
Fig. 6 Dependency of recognition rates on threshhold at 200 Hz bandwidth.

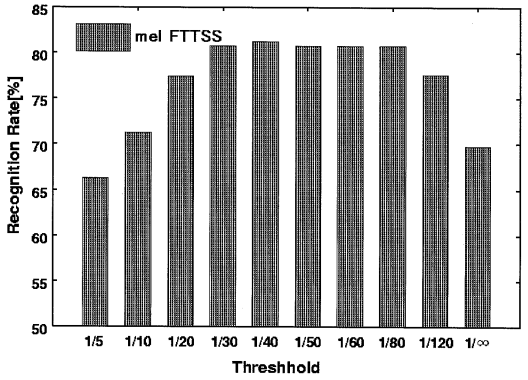

図 7 バンド幅 50 Hz におけるしきい値による認識率の違い
Fig. 7 Dependency of recognition rates on threshhold at 50 Hz bandwidth.

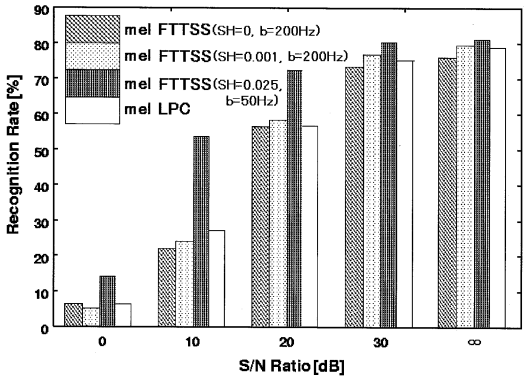

図 8　車内雑音環境下での認識率
Fig. 8　Recognition rates under car noise

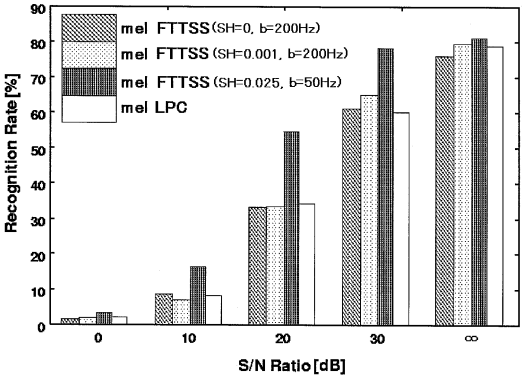

図 10　幹線道路雑音環境下での認識率
Fig. 10　Recognition rates under trunk road noise.

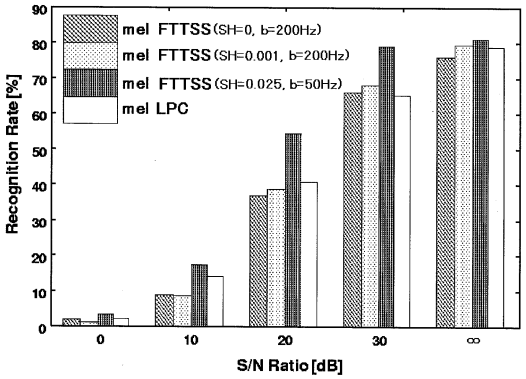

図 9　ワークステーションノイズ環境下での認識率
Fig. 9　Recognition rates under workstation noise.

図 7 より，平均振幅の 1/40 のしきい値が最も良い認識率を与えていることがわかる．バンド幅を 200 Hz としたときよりも最適なしきい値が大きくなっている．この理由は，**4. 2** で述べたように，バンド幅を狭くしたために高調波に由来する周波数微細構造が現れてきているので，その影響を抑制するためにより大きめのしきい値になったと考えられる．バンド幅を 50 Hz，しきい値を平均振幅の 1/40 とおいて各雑音環境下で認識実験を行った結果を，図 8〜図 10 で凡例の 3 番目の棒グラフで示す．

これらの図より，すべての雑音環境下においてメル FTTSS の方がメル LPC ケプストラムよりも認識率がよくなっていることがわかる．

5. むすび

雑音が重畳した音声の認識率向上を目指し，スペクトル傾斜に基づいた特徴（FTTSS）を提案し，特に有声音の高調波に由来する周波数微細構造による影響や，雑音による影響を軽減するため，スペクトル傾斜にしきい値操作を導入する有効性を，単語認識実験を通して確認した．なお，今回の実験では S/N に対する認識率の相対的な変化を調べることを目的としたため，例えば，動的な特徴量 [7] の導入，単語モデルの連結学習等，認識率を上げるための手法を適用していないので，無雑音環境下でも約 80%と低い単語認識率にとどまったが，これらの手法を適用することにより，認識率の向上が期待できると思われる．

分析に用いた帯域フィルタ対バンク（BPF）の周波数間隔 (Δf)，バンド幅 (b)，中心周波数 (f_c)，しきい値（SH）などの各設定値について，今後より適切な値を探索することが課題である．

文　献

[1] D.H. Klatt, “Prediction of perceived phonetic distance from critical-band spectra,” Proc. ICASSP ’82, pp.1278–1281, 1982.
[2] H.W. Strube, “Linear prediction on a warped frequency scale,” Jour. Acoust. Soc. Am., vol.68, no.4, pp.1071–1076, 1980.
[3] 中藤良久，松本　弘，“音声認識におけるメル線形予測分析法の評価,” 信学技報，SP98-22, 1998.
[4] 南　泰浩，松岡達雄，鹿野清宏，“不特定話者連続音声データベースによる連結学習 HMM の評価,” 信学技報，SP91-113, 1992.
[5] 八木裕之，船田哲男，“非線形処理による耐雑音性をもった音声の特徴抽出,” 第 11 回 DSP シンポジウム，B6-1, pp.497–502, 1996.
[6] 続木貴史，船田哲男，“BPFP メルケプストラムの音声認識における耐雑音性の検討,” 信学技報，SP98-50, 1998.
[7] 古井貞熙（監訳），音声認識の基礎（上），NTT アドバンステクノロジ，東京，1995.

（平成 11 年 5 月 7 日受付，6 月 8 日再受付）