1—5—2　変調スペクトルの貢献度に基づく音声認識特徴量 *

○金寺 登　釜井政義 (石川高専)　荒井隆行　岡田賢治　百村裕智 (上智大・理工)

1. はじめに

音声特徴量の時間軌跡をフーリェ変換したものは，変調スペクトルと呼ばれている．音声の認識には特定の変調スペクトルが重要であることが知られている[1, 2]．文献[3, 4, 5]では音声認識にとって変調スペクトルの各成分がどの程度重要であるかを示す貢献度を定量的に調査した．

本報告では，この貢献度に応じて変調スペクトルを強調した音声認識特徴量の自動音声認識実験結果について報告する．

2. 変調スペクトルの貢献度に基づく音声認識特徴量

2.1 変調スペクトル成分の貢献度

文献[5]において調査した連続音声に対する各変調スペクトルの貢献度とその95%信頼区間を図1に示す．図中の貢献度は，対応する変調周波数バンドを含めることで，誤り率が1/(貢献度)になることを表している．従って，貢献度が1より大きければシステム性能が向上し，1未満であればシステム性能が低下することを意味する．

2.2 JC-RASTA

重要な変調周波数バンドを通過させ認識性能を向上させる方法としてRASTA[6]が知られている．JC-RASTA[7]では，まずスペクトル x を

$$y = \log(1 + Jx) \tag{1}$$

によって，非線形変換する．ここで J は正定数である．振幅変換関数 (1) 式は $J \ll 1$ のとき線形的であり，$J \gg 1$ のとき対数的である．ここでは $J = 10^{-6}$ としたため，(1) 式は線形的である．次に (1) 式の時間軌跡に対してRASTAフィルタを用いて，約1~12 Hzの成分のみを抽出する．さらに

$$x' = \frac{1}{J} e^{y} \tag{2}$$

によって元のスペクトル次元に戻す．

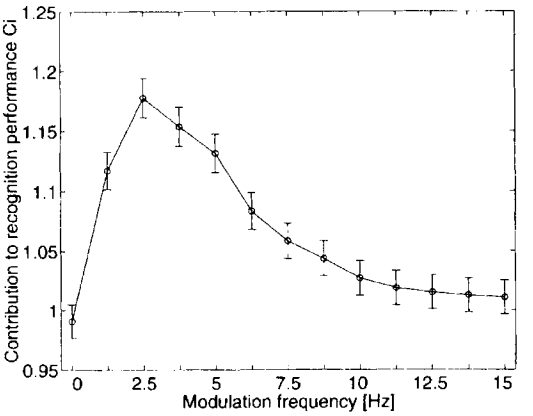

図 1. 連続音声認識に対する変調スペクトル貢献度

2.3 貢献度に基づく音声認識特徴量

RASTAフィルタはIIRフィルタであるため位相歪が原因で認識性能が劣化することがある．そこで，2.2節のRASTAフィルタの代わりに図1の貢献度に比例した変調周波数特性をもつ直線位相FIRフィルタを用いる方法を提案する．

各変調周波数バンドの重要性（貢献度）に応じて重み付けした抽出フィルタを用いることにより，音声の変調周波数成分が強調されるのに対し，雑音の変調周波数成分の多くが軽減され，雑音環境下での音声認識性能が向上することが期待できる．

3. 実験結果

3.1 分析条件

日本語ディクテーション基本ソフトウェア（98年度版）[8]を用いて，各特徴量による音響モデルの学習・評価を行った．音響モデルは2000状態16混合のtri-phoneとした．各種学習・評価条件は文献[8]と同様である。ただし，学習・評価データには男性話者のみを用いた．

学習データには雑音を付加しないクリーンなデータを使用した．一方評価データには，SNRが20, 10, 0 dBになるように付加雑音を音声データに波形レベル

* A speech recognition feature based on the contribution of modulation frequency components.

By Noboru Kanedera, Masayoshi Kamai (Ishikawa National College of Technology), Takayuki Arai, Kenji Okada, and Yasunori Momomura (Sophia University)

表 1. 評価試験結果 (単語認識率 [%])

特徴量	特徴量数	clean	SNR 20dB	SNR 10dB	SNR 0dB
(1) PLP+Δ	17	90.2	85.5	75.2	51.8
(2) JC-RASTA+Δ	17	85.8	84.1	76.5	52.9
(3) PLP-RI+Δ	17	90.3	87.9	80.2	56.0
(4) MFCC+Δ	25	92.1	88.3	79.6	57.8
(5) MFCC-RI+Δ	25	91.4	88.4	80.3	57.9
(6) PLP-RI+Δ+Δ^2	26	91.4	90.1	84.5	63.9

で加算したものを用いた．付加雑音にはNOISEX-92データベースより15種類の雑音データを用いた．

3.2 実験結果

以下の特徴量について，雑音環境下における評価試験を行った結果を表1に示す．

(1) PLP+Δ ··· 8次のPLPケプストラムとその動的特徴量 Δ，対数パワーの動的特徴量

(2) JC-RASTA+Δ ··· **2.2**節の特徴量

(3) PLP-RI+Δ ··· PLPに**2.3**節の方法で変調フィルタリングを行った特徴量

(4) MFCC+Δ ··· 12次のメルケプストラムとその動的特徴量 Δ，対数パワーの動的特徴量

(5) MFCC-RI+Δ ··· MFCCに**2.3**節の方法で変調フィルタリングを行った特徴量

(6) PLP-RI+Δ+Δ^2 ··· (3)の特徴量に，Δ^2 特徴量を加えた特徴量

表1において，cleanは雑音データがない場合の単語認識率を示している．また，20dB, 10dB, 0dBはSNRを表しており，15種類の雑音をそれぞれのSNRで混入した場合の平均単語認識率を示している．

まず，(1)のPLP分析を用いた特徴量とPLP分析後に約1～12 Hzの変調周波数バンドを抽出する(2)のJC-RASTAを比較する．雑音の少ないclean, 20dBにおいては(1)が多少良い結果が得られているのに対し，10dB以下の雑音環境においては(2)がわずかに優れている．この結果より，特定の変調周波数バンドを抽出する方法は，雑音環境下において効果があることがわかる．これは，人間の聴覚特性の調査[2]によって明らかになったように音声を認識するために重要な変調周波数バンドが約1～16Hzであるのに対し，雑音の変調周波数バンドが広範囲の変調周波数バンドに分布しているためと考えられる．すなわち音声認識にとって重要な変調周波数バンドのみを抽出することによって，音声の変調周波数成分が保持されるのに対し，雑音の変調周波数成分の多くが軽減されるため，雑音環境下での結果が良くなったと考えられる．

次に，(2)のJC-RASTAと提案法(3)を比較する．提案法(3)では，重要な変調周波数バンドを抽出する際に，各変調周波数バンドの重要性（貢献度）に応じて重み付けした抽出フィルタを用いることにより性能向上を目指している．特徴量数17の条件では，提案法(3)が最も良い結果となった．

現在，最も広く用いられている従来法(4)と提案法(5)(6)を比較する．従来法(4)に貢献度に応じた変調フィルタリングを施した提案法(5)は，雑音環境下においてわずかながら効果がある．また，従来法(4)と特徴量数が近い提案法(6)を比較すると，雑音環境下においてさらに効果があることが確認できた．SNRが10dBの場合について従来法(4)に比べ，提案法(6)では約5％認識率が改善されていることがわかった．

4. まとめ

各変調周波数バンドがどの程度重要であるかを示す貢献度に基づく音声認識特徴量を提案し，雑音環境下において音声認識性能が改善されることを確認した．

参考文献

[1] R. Drullman, J. M. Festen, and R. Plomp, J. Acoust. Soc. Amer., **95**, pp. 2670 – 2680, (1994).

[2] T. Arai, M. Pavel, H. Hermansky and C. Avendano, J. Acoust. Soc. Amer., **105**, 5, pp. 2783 – 2791, (1999.5).

[3] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, Speech Communication, **28**, pp.43 – 55, (1999.5).

[4] 金寺, 荒井, 船田, SP98-51, pp.45 – 52, (1998.7).

[5] 金寺, 荒井, 高橋, 船田, SP2000-34, pp.15 – 22, (2000.7).

[6] H. Hermansky and N. Morgan, IEEE Trans. Speech and Audio Process., **2**, 4, pp. 578 – 589, (1994).

[7] H. Hermansky, N. Morgan and H. Hirsch, Proc. IEEE ICASSP, Minneapolis, MN, pp. II-83 – II-86, (1993).

[8] 河原, 李, 音響学会誌, **56**, 4, pp. 255 – 259, (2000).