1—1—2

音声中の言語情報を担う変調スペクトル特性の検討 *

○金寺 登　荒井 隆行　船田 哲男
(石川高専)　(上智大・理工)　(金沢大・工)

1. はじめに

対数スペクトルの時間軌跡またはケプストラムの時間軌跡のフーリエ変換は変調スペクトルと呼ばれている。単語音声に対して変調スペクトル間の相対的な重要性を以前に調査した結果、以下のことが明らかになった[1, 2]。(1) 言語情報のほとんどが 1~16 Hz (特に 2~10Hz) の変調周波数バンドに存在し、その中でも 4 Hz 付近が最も重要である。(2) 変調スペクトルにおいては位相情報も重要である。(3) 4Hz 付近の変調周波数を含む特徴量を用いることで動的特徴量と同等以上の結果が得られる。(4) 適切な中心周波数とバンド幅をもつ複数のバンドを変調周波数上で用いることにより、認識性能がさらに向上する。

本報告では、変調スペクトル間の相対的な重要性の新しい尺度を定義し、単語音声及び連続音声に対する変調スペクトル間の相対的な重要性の調査結果を報告する。

2. 変調スペクトル間の相対的な重要性

本節では変調スペクトル間の相対的な重要性の新しい尺度を定義する。ケプストラム等の時間軌跡に種々のバンドパスフィルタを適用して得られたパラメータによる認識率が得られているものとする。このとき認識率 $p(f_L, f_U)$ は 時間軌跡に対するバンドパスフィルタの低域遮断周波数 f_L と高域遮断周波数 f_U の関数である。オーバラップしない 2 つのバンド 1,2 による認識率をそれぞれ p_1, p_2 とする。ここで、オーバラップしない バンドは独立に認識結果に貢献すると仮定する。このとき、バンド 1,2 を両方用いた時の認識率 $p_{1,2}$ は $p_{1,2} = 1-(1-p_1)(1-p_2)$ となる。言い替えれば、誤り率 $q_{1,2} = 1 - p_{1,2}$ は $q_{1,2} = q_1 q_2$ のようにそれぞれのバンドの誤り率の積になる。

一般に k 個のバンドを用いた時の誤り率は、$q_{1,2,\cdots,k} = \prod_{i=1}^{k} q_i$ となる。積が和になるように両辺を対数に変換すると、

$$Q_{1,2,\cdots,k} = \sum_{i=1}^{k} Q_i \qquad (1)$$

となる。ここで $Q_i = \log q_i$ である。

我々の目的は、いくつかのバンドから得られた複数の認識率より、個々の狭いバンドが認識性能にどの程度貢献するかを推定することである。すなわち、いくつかの $Q_A, Q_B, \cdots$ が 得られている時に Q_i を推定することである。ここで、$A, B, \cdots$は、$A = \{1,2,3,4\}$ のように、バンド番号の自然数を要素とする集合を表す。式 (1) の線形性により、この推定は直線回帰に帰着できる。例えば、あるバンドの集合 A による誤り率の推定量$\widehat{Q}_A$は、次式により与えられる。

$$\widehat{Q}_A = \sum_{i=1}^{k} w_i X_A(i) \qquad (2)$$

ここで、$w_i = \widehat{Q}_i$ であり、X_A は、バンド i が A に含まれるかどうかを示す関数で次式で定義される。

$$X_A(i) = \begin{cases} 1 & i \in A \\ 0 & i \notin A \end{cases}$$

以上をまとめると、まずいくつかのバンドから得られた複数の認識率より式 (2) の回帰重み w_i を求める。次に各変調スペクトルの認識性能への貢献度 C_i は、$C_i = \exp(-w_i)$ と定義する。標準的な回帰計算法により、回帰重みとその信頼区間を求めることができる。

3. 変調スペクトルの貢献度

変調スペクトル間の相対的な重要性を調査するため、以下の音声認識実験を行った。まず 8 次の PLP と対数パワーを求め、これらの各時間軌跡について、64 フレームを切り出し、ハミング窓を適用後、64 点の DFT を計算した。次に 対象とする変調周波数バンドに対応する成分のみを抽出し、その時刻における特徴量とした。さらに 時間軌跡切り出し位置を 1 フレームずつシフトすることにより、すべての時刻において対象とする変調周波数バンドに対応する特徴量を抽出した。対象とする変調周波数バンドを様々に変化させ、対応するシステムの認識率を求めれば、2.節の方法により、各変調周波数が認識性能に寄与する貢献度 C_i を求めることができる。

---

* On the properties of modulation spectrum for continuous speech recognition.
By Noboru Kanedera (Ishikawa National College of Technology), Takayuki Arai (Sophia University) and Tetsuo Funada (Kanazawa University).

図 1は、単語音声に対する各変調スペクトルの貢献度を 95%信頼区間付きでを示している。横軸は各 DFT フィルタの中心変調周波数を表している。この実験には、Bellcore digit database を使用した。図 1(a) は 雑音が少ない環境での結果を示しているのに対し、図 1(b) においては、評価データが 加法性雑音 (SNR 10 dB) と乗法性雑音 (HPF, 6 dB/oct) によって劣化された場合の結果を示している。その他の詳細な条件は、文献[2] 表 3 と同様である。

図中の貢献度 C_i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。図 1より、2 Hz ～ 10 Hz は クリーンな環境と雑音環境の両方で重要であった。また雑音環境では 2 Hz 以下の変調周波数成分の重要性は低くなった。特に 1 Hz 以下の変調周波数成分は著しく認識率を劣化させることがわかった。

連続音声に対して 同様に導出した貢献度を図 2に示す。ただし、貢献度を求める際に使用する認識率には音節認識率を用いた。また、学習データには ATR 音声データベース セット C 連続音声 150 文, 男女各 10 名分を用い、評価データは 学習データとは異なる男女各 10 名分とした。

図 2の変調周波数間の相対的な重要性の傾向は、図 1の傾向と類似していることがわかった。

4. まとめ

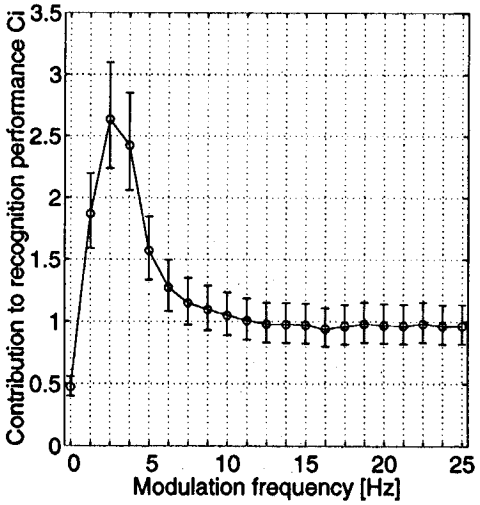
変調スペクトル間の相対的な重要性の新しい尺度を定義し、単語音声及び連続音声に対する変調スペクトル間の相対的な重要性の調査した。その結果、2 Hz ～ 10 Hz の変調周波数バンドが、単語音声についても連続音声についても重要であることがわかった。

参考文献

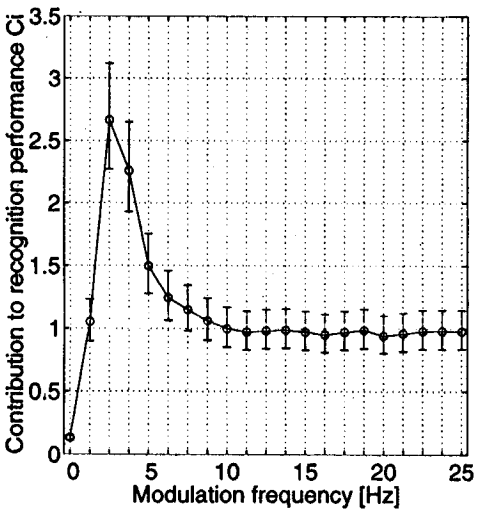
[1] N. Kanedera, H. Hermansky and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," ICASSP98, pp. II-613 – II-616 (1998).

[2] 金寺 登, 荒井隆行, H. Hermansky, 船田哲男, "ロバストな音声認識実現を目的とした変調スペクトル特性の検討," SP97-70 pp.15-22 (1997.12).

[3] 金寺 登, 荒井隆行, 船田哲男, "複数の変調スペクトル解像度を用いた音声認識の耐雑音性," SP98-51 pp.45-52 (1998.07)



(a) Clean



(b) Noisy

図 1. Contribution to recognition performance for word speech.

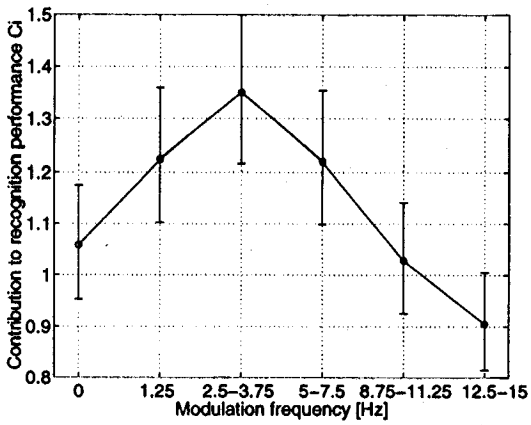


図 2. Contribution to recognition performance for continuous speech.