社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2000-34 (2000-07)

音声の変調スペクトル中に含まれる情報の調査 — 音声認識情報と話者識別情報との比較 —

金寺 登[1] 荒井 隆行[2] 高橋 真保呂[2] 船田 哲男[3]

[1] 石川高専 〒929-0392 石川県河北郡津幡町北中条 kane@i.ishikawa-nct.ac.jp
[2] 上智大学 〒102-8554 東京都千代田区紀尾井町 7-1 {arai,mahoro}@splab.ee.sophia.ac.jp
[3] 金沢大学 〒920-8667 石川県金沢市小立野 2-40-20 funada@t.kanazawa-u.ac.jp

あらまし

対数スペクトルあるい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．本報告では，音声認識のために重要な変調周波数バンドと話者識別のために重要な変調周波数バンドを調査した結果を報告し，両者を比較・検討する．

連続音声の音節認識実験及び話者識別知覚実験による調査の結果，音声認識及び話者識別いずれの場合も2～8 Hz の変調周波数バンドが重要であることがわかった．また，音声認識情報は 16 Hz 以上にほとんど存在しないが，話者識別情報は 16 Hz 以上にも存在する可能性があることが実験結果により示された．

キーワード: 変調スペクトル，変調周波数，音声認識，話者識別

Investigations on information of speech recognition and speaker identification in modulation spectrum.

Noboru Kanedera[1] Takayuki Arai[2] Mahoro Takahashi[2] Tetsuo Funada[3]

[1]Ishikawa National College of Technology, Tsubata, Ishikawa 929-0392
[2]Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554
[3]Faculty of Engineering, Kanazawa University, Kanazawa, Ishikawa 920-8667

Abstract The Fourier transform of the time trajectories of a parameter such as logarithmic spectrum or cepstrum is called the modulation spectrum. In this paper we report on the important modulation-frequency bands for speech recognition and speaker identification. The results by continuous speech recognition experiments and perceptual speaker-identification experiments show that the modulation-frequency band between 2 and 8 Hz is important for both speech recognition and speaker identification. The results also suggest that the information for speaker identification lies in the range above 16 Hz, while information for the information for speech recognition does not lie in the range.

Keywords: modulation spectrum, modulation frequency, speech recognition, speaker identification

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おり，人間と機械による音声認識·話者識別の分野で有効な変調周波数成分が広く調査されている [1–10].

知覚実験 [1, 2] により，一部の変調スペクトル成分が他に比べて重要であることが知られている．この事実は日本語 [3] や英語 [4] においても確認されている．Drullman ら [1, 2] は，16 Hz 以下の低域通過フィルタリングや 4 Hz 以上の高域通過フィルタリングによって，音声の明瞭度が低下しないことを示している．荒井ら [3, 5] は，Drullman らの研究をケプストラムに対応する対数領域に拡張し，低域／高域通過フィルタばかりでなくバンドパスフィルタを適用した．この結果，明瞭度を保持するために必要なほとんどの情報が 1～16 Hz の変調周波数バンドに存在することが明らかとなった．

ASR (automatic speech recognition) に対して，金寺ら [6, 7, 8] は単語音声認識実験を行い，変調スペクトル成分の重要性を調査した．この結果，ASR にとって重要な情報のほとんどが 1～16 Hz の変調周波数バンドに存在し，その中でも音声の音節速度 (syllabic rate) に対応する 4 Hz 付近が最も重要であることがわかった．また雑音環境においては，2 Hz 以下や 16 Hz 以上の変調スペクトル成分が認識性能を劣化させることがあることがわかった．特に 1 Hz 以下の変調スペクトル成分は認識性能を著しく低下させる．

一方，話者情報を担う変調周波数バンドに関して，Vuuren ら [9] は，自動話者識別実験によって，0.1～10 Hz に重要な話者情報が含まれると報告している．

以上のように，音声認識のために重要な変調周波数バンドに関して，知覚実験，単語音声認識実験による調査が行われているが，連続音声認識実験による調査は行われていない．また，話者識別のために重要な変調周波数バンドに関しては，自動話者識別実験による調査が行われているが，知覚実験による調査は行われていない．

そこで本報告では，連続音声認識実験による重要な変調周波数バンドの調査と，知覚実験による話者識別のために重要な変調周波数バンドの調査を行った結果について報告する．また，これらの調査結果を基に音声認識のために重要な変調周波数バンドと話者識別のために重要な変調周波数バンドとを比較·検討する．

まず第 2 節では，各変調スペクトル成分の重要性を表す尺度として貢献度を定義する．第 3.節では，音声認識のために重要な変調周波数バンドを調査した結果を報告する．第 4 節では話者識別のために重要な変調周波数バンドを調査した結果を報告する．第 5 節では変調スペクトル中の音声認識情報と話者識別情報とを比較・検討する．

2 変調スペクトル成分の貢献度

本節では各変調スペクトル成分の重要性を表す尺度として貢献度を定義する．いくつかのバンドから得られた複数の認識率が与えられているとき，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が目的である．

まず，あらかじめ 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誤り率が得られているものとする．このとき認識誤り率 $q(f_L, f_U)$ 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 による認識誤り率をそれぞれ q_1, q_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誤り率 q_A は $q_A = q_1 q_2$ のようにそれぞれのバンドの誤り率の積になる．ここで A は，$A = \{1, 2\}$ のようにバンド番号の自然数を要素とする集合を表す．

一般に 任意のバンドの集合 A を用いた時の誤り率は，

$$q_A = \prod_{i \in A} q_i \tag{1}$$

となる．積が和になるように両辺を対数に変換すると，

$$Q_A = \sum_{i \in A} Q_i \tag{2}$$

となる．ここで $Q_i = \log q_i$ である．式 (2) の線形性により，式 (2) は次式のように変形できる．

$$Q_A = \sum_{\text{all } i} Q_i X_A(i) \tag{3}$$

ここで，$X_A(i)$ 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．

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

いくつかのバンドの集合 $A, B, \cdots$ から得られた認識誤り率の対数 $Q_A, Q_B, \cdots$ が与えられているときに，$A, B, \cdots$ のすべての場合に式 (3) をなるべく満たす（2乗誤差を最小にする）ような Q_i の推定量 $\hat{Q}_i$ を求めたい．式 (3) は直線回帰の形式であるため，一般的な回帰計算法により 回帰重み $\hat{Q}_i$ とその信頼区間を求めることができる．この $\hat{Q}_i$ はバンド i が認識性能にどの程度貢献するかを表している．よって各変調スペクトル成分の認識性能への貢献度 C_i を

$$C_i = \exp(-\hat{Q}_i) \tag{4}$$

と定義する．

以上をまとめると，まずいくつかのバンドの集合 $A, B, \cdots$ から得られた認識誤り率の対数 $Q_A, Q_B, \cdots$ より式 (3) の回帰重み $\hat{Q}_i$ を求める．次に 式 (4) により各変調スペクトル成分の認識性能への貢献度 C_i が求められる．

この貢献度 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 よりも大きければ大きいほど対応する変調周波数バンドは，有用な情報を含んでいることになる．

3 音声認識のために重要な変調周波数バンド

本節では，音声認識のために重要な変調周波数バンドを調査した結果を報告する．

3.1 連続音声認識実験による重要な変調周波数バンドの調査

これまでに知覚実験 [5]，単語音声認識実験 [8, 10] による重要な変調周波数バンドの調査が行われている．しかし，連続音声認識実験による調査は行われていない．そこで本節では，連続音声認識実験による重要な変調周波数バンドの調査を行った結果を報告する．

まず 8 次の PLP (perceptual linear predictive coding)[11] と対数パワーを求めた．次に これらの各特徴パラメータの時間軌跡について，64 フレームを切り出し，ハミング窓を適用後，64 点の DFT により変調スペクトルを計算した．この変調スペクトルは，各変調周波数バンド成分の集合と見なせる．よって各変調周波数バンドからなる任意の集合は，対応する変調スペクトル成分のみを選択することにより得られる．以上の処理により，時間軌跡切り出し時刻における，任意の変調周波数バンド集合に対応する特徴量が求められる．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，すべてのフレームにおいて対象とする変調周波数バンド集合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集合を様々に変化させ，音声認識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．

図 1は，連続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．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．(a) は音節認識率として正解率＝正解音節数／音節総数 を用いた場合で，(b) は 正解精度＝(音節総数－置換音節数－脱落音節数－付加音節数)／音節総数 を用いた場合に対応する．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．HMM には 各音節毎に 5 状態 3 出力分布 1 混合 HMM を使用した．各分布の共分散は対角とした．その他の条件は，表 1のとおりである．

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．この連続音声認識実験による貢献度の結果より，連続音声の認識にとって約 1 ～ 10 Hz の変調スペクトル成分が重要であることがわかった．一方，0 Hz 付近や 10 Hz より大きい成分の貢献度が低くなった．

表 1: 連続音声認識実験条件

Recognizer	HMM
Sampling frequency	16 kHz
Window length	25 ms
Frame period	12.5 ms

3.2 単語音声認識実験による重要な変調周波数バンドとの比較

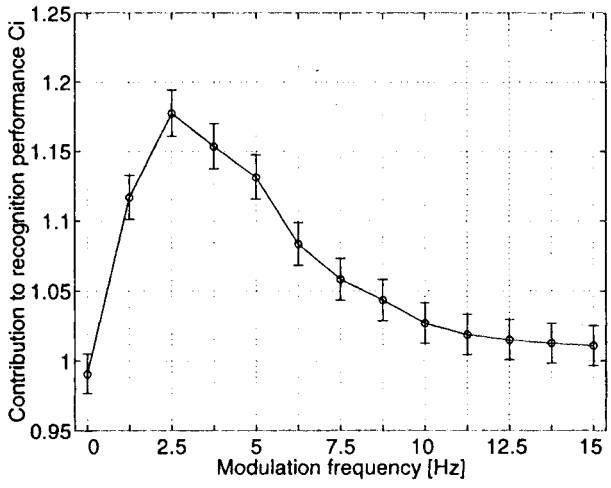
文献 [10] において行った単語音声認識実験結果より，2節の貢献度を求めた結果を図 2に示す．貢献度の計算においては，単語認識率を使用した．この実験には，13 単語の Bellcore digit database を使用した．HMM の学習には HMM toolkit (HTK) を用い，各単語毎に 8 状態 6 出力分布 2 混合 HMM を学習した．各混合の共分散は対角とした．図 2(a) は雑音が少ない環境での結果を示しているのに対し，図 2(b) は，評価データが 加法性雑音 (SNR 10 dB) と乗法性雑音 (HPF, 6 dB/oct) によって劣化させた場合の結果を示している．この加法性雑音は，コンピュータ室 (約 55 dB の背景雑音環境) においてダイナミックマイクロホンを使用し Sound Blaster 互換ボードを介して，直接パソコンに収録された．

図 2の単語音声に対する各変調スペクトルの貢献度は，図 1の連続音声に対する各変調スペクトルの貢献度の傾向と類似している．すなわち，約 1～ 10 Hz の変調スペクトル貢献度が大きい．一方，0 Hz 付近や 10 Hz より大きい成分の貢献度が低い．

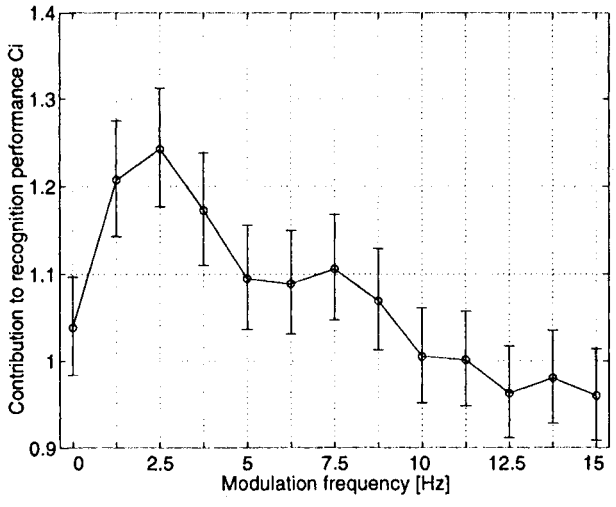
3.3 知覚実験による重要な変調周波数バンドとの比較

文献 [5] の知覚実験結果より，2節の貢献度を求めた結果を図 3に示す．この実験では，日本語音節の明瞭度が使用されているので，この明瞭度を使用して貢献度を計算した．

知覚実験による各変調スペクトルの貢献度は，図 1や図 2の ASR による各変調スペクトルの貢献度の傾向と類似している．しかし，0 Hz 付近では，知覚実験の貢献度が ASR による貢献度に比べて高くなっている．このことから，人間は 0 Hz 付近の変調スペクトル（静的な情報）についても，有効に利用していると考えられる．



(a) 正解率



(b) 正解精度

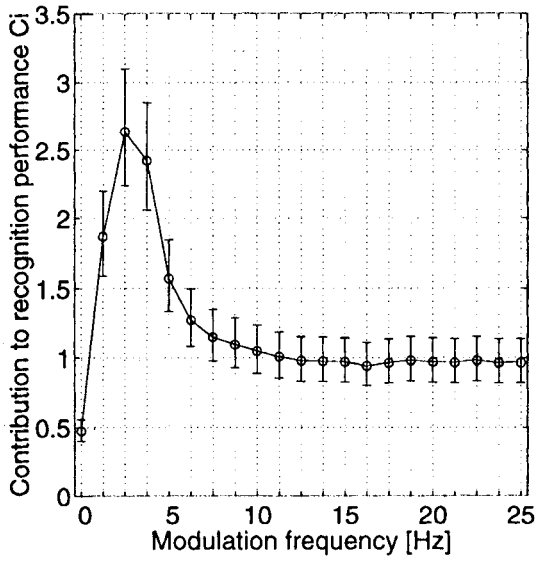
図 1: 連続音声認識に対する変調スペクトル貢献度

4 話者識別のために重要な変調周波数バンド

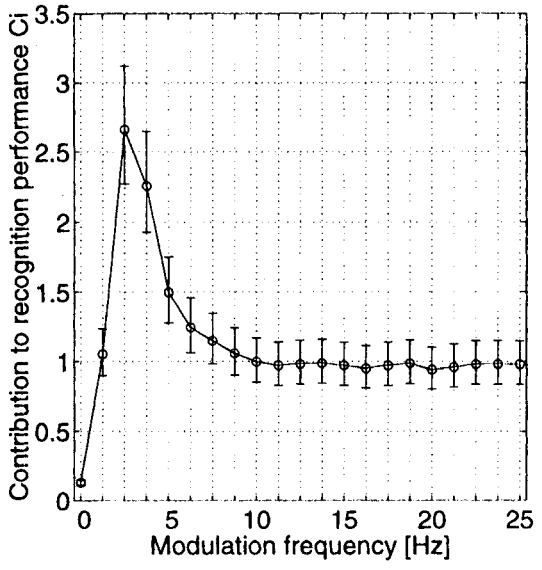
本節では話者識別のために重要な変調周波数バンドを知覚実験により調査した結果について報告する．

4.1 分析条件

様々な変調周波数成分を持つ音声を生成し，話者識別知覚実験を行うために，信号処理ツール [12] を用いて，図 4に示す分析合成を行った．まず原音声より，窓長 25ms のブラックマン窓を用いて 12 次



(a) Clean（雑音なし）



(b) Noisy（雑音あり）

図 2: 単語音声認識に対する変調スペクトル貢献度

の mel-frequency cepstrum coefficients (MFCC) を 5ms 毎に求めた．次に MFCC の時間軌跡を 1023 点の FIR フィルタによりフィルタリングし，ある範囲の変調変調周波数バンドの成分のみを抽出した．さらにフィルタリング後の MFCC とピッチを用いて音声を合成し，最後に文全体の音声の大きさを正規化した．

4.2 話者識別知覚実験

まず話者識別の対象音声に前節の分析合成を施し，特定の変調周波数バンドのみを含む提示音声を作成した．使用した変調周波数バンドは，0, 0.25, 0.5 1,

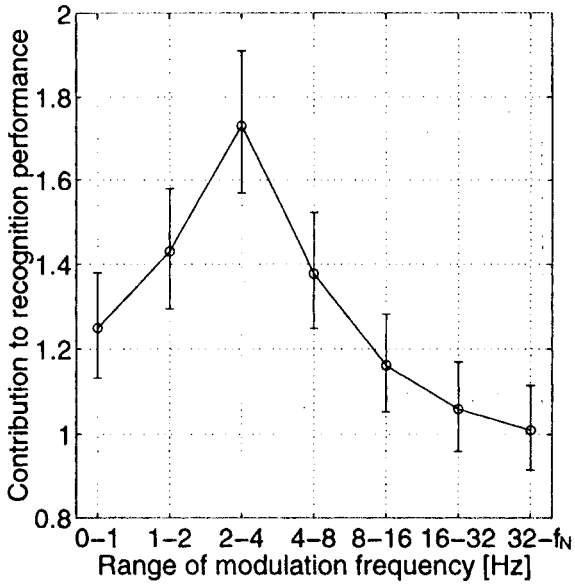


図 3: 音節認識知覚実験による変調スペクトル貢献度

2, 4, 8, 16 Hz, f_N(ナイキスト周波数) を遮断周波数とする 36 種類である．

識別対象話者は 石川高専 電子情報工学科の教官 5 名，被験者は 5 名の教官の声を日頃良く聞いている同学科の学生 16 名 (男女各 8 名) とした．提示文には「あらゆる現実をすべて自分のほうへねじまげたのだ.」を用いた．

実験の前に識別対象となる話者の音声を確認してもらった．確認用の文には提示文とは異なる「青い植木鉢」を用い，2 度被験者に提示した．次に，各被験者に 180 文 (36 種類× 5 名分) を提示した．提示

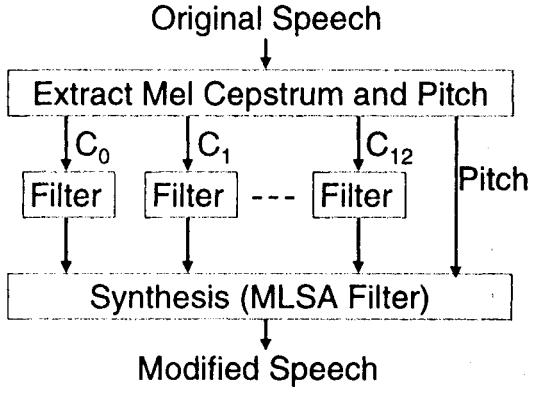


図 4: 分析合成方法

表 2: 話者識別率 [%]

f_L [Hz]	f_U [Hz]							
	0.25	0.5	1	2	4	8	16	f_N
0	1.3	2.5	2.5	20.0	45.0	60.0	76.3	82.5
0.25		5.0	5.0	18.8	36.3	58.8	63.8	72.5
0.5			12.5	16.3	36.3	55.0	58.8	63.8
1				13.8	22.5	41.3	56.3	67.5
2					22.5	36.3	38.8	46.3
4						16.3	13.8	32.5
8							16.3	3.8
16								6.3

順序は，判別が難しいと判断されるフィルタリング条件の順とし，各条件ごとに乱数で提示話者順を決定した．1文を聴取する度にいずれの話者であるかまたは わからないかを番号で答えるよう指示した．

4.3 実験結果

表2に，種々の変調周波数バンドに対する話者識別率を示す．表中の f_L は低域遮断変調周波数，f_H は高域遮断変調周波数を表している．これらの話者識別率を基に2節の方法で各変調周波数バンドの話者識別に対する貢献度を95%信頼区間付きで求めたものを図5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，話者識別誤り率が1/(貢献度) になることを表している．従って，貢献度が1より大きければ大きいほど対応する変調周波数バンドに多くの話者情報が含まれていることになる．

図5より，2 Hz ～ 8 Hz に多くの話者情報が含まれていることがわかった．この範囲の変調周波数バンドは，言語情報が多く含まれる範囲と一致している．文献 [9] の話者自動識別実験では，0.5 Hz ～ 8 Hz に多くの話者情報が含まれているのに対し，0.125 Hz 以下はかえって話者識別性能を低下させると報告している．文献 [9] と比較して，今回の知覚実験では，0 Hz ～ 0.25Hz の貢献度が高く，0.25 Hz ～ 1 Hz の貢献度が低い結果となった．

5 音声認識情報と話者識別情報との比較

本節では，3節及び4節の調査結果を基に音声認識のために重要な変調周波数バンドと話者識別のために重要な変調周波数バンドとを比較・検討する．

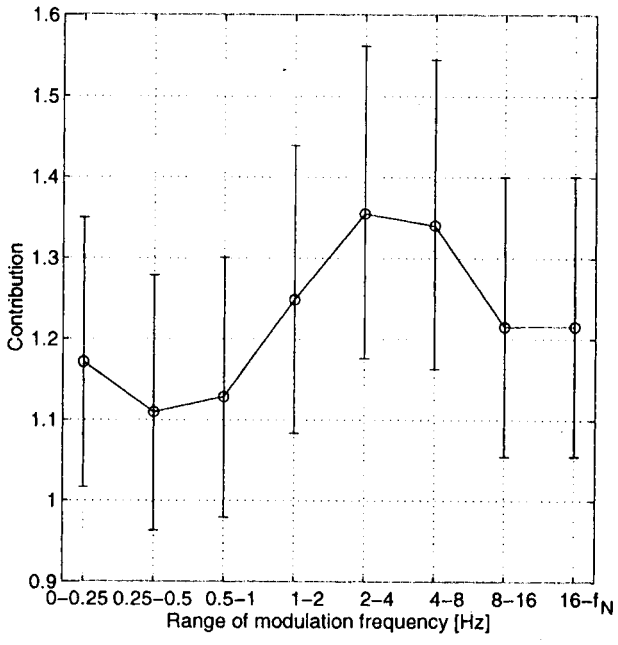


図 5: 話者識別知覚実験による変調スペクトル貢献度

3節では音声認識にとって重要な変調スペクトルを調査した．図1の連続音声認識実験による貢献度の結果より，連続音声の自動認識にとって約 1 ～ 10 Hz の変調スペクトル成分が重要であることがわかった．一方，0 Hz 付近や 10 Hz より大きい成分の貢献度が低くなった．図2の単語音声の自動認識に対する変調スペクトル貢献度の傾向は，連続音声の自動認識に対する変調スペクトル貢献度と同様であった．図3の知覚実験による各変調スペクトルの貢献度は，図1や図2の ASR による各変調スペクトルの貢献度の傾向と類似していたが，0 Hz 周辺では，知覚実験の貢献度が ASR による貢献度に比べて高くなった．

一方，4節では話者識別にとって重要な変調スペクトルを調査した．図5の知覚実験結果より，2 Hz ～ 8 Hz に多くの話者情報が含まれていることがわかった．また，文献 [9] の話者自動識別実験と比較して，話者識別知覚実験では，0 Hz ～ 0.25Hz や 8 Hz より大きい変調周波数バンドの貢献度が高い結果となった．

これらの結果より，音声認識，話者識別に対するすべての実験に対して，2 Hz ～ 8 Hz の変調周波数バンドが特に重要であった．このことは，平均音節速度が 4 Hz 周辺であること [13] と符合している．

実際，この範囲を含めた変調周波数バンドのみを通過させるバンドパスフィルタリングを行うことで，ASR の耐雑音性能が向上することを確認している [14]．話者自動識別においても同様なフィルタリングにより性能が向上することが確認されている [9]．

また，言語情報や話者情報は声道の変化などにコード化され，この変化が音声信号に反映される．雑音などの非言語的情報の変化の割合は，言語情報や話者情報を担う声道の変化の割合と異なることが多い [15]．このことから声道の変化の割合に対応する変調周波数バンドのみを使用することによって，耐雑音性能が向上することを説明できる．

音声認識の知覚実験では 16 Hz 以上の変調周波数バンドの貢献度が低いのに対し，話者識別の知覚実験では 16 Hz 以上の変調周波数バンドの貢献度も高かった．このことは，話者識別情報は 16 Hz 以上の変調周波数バンドにも存在する可能性を示している．ただし，Vuuren らの話者自動識別においては 16 Hz 以上の変調周波数バンドを含めても改善されていない [9]．よって，16 Hz 以上の変調周波数バンドを有効に活用するためには何らかの工夫が必要と考えられる．

音声認識の知覚実験や話者識別の知覚実験における 0 Hz 近くの貢献度に比べ，ASR や話者自動識別における 0 Hz 近くの貢献度が小さい．このことから，人間は 0 Hz 付近の変調スペクトル（静的な情報）についても，有効に利用していると考えられる．

以上のように音声認識情報と話者識別情報は密接に関係しているため，これらを統一的に扱う枠組みが重要となるであろう．

6 まとめ

音声認識情報と話者識別情報とを比較するため，連続音声認識実験及び話者識別知覚実験により音声の変調スペクトル中に含まれる情報の調査を行った．調査の結果，音声認識，話者識別いずれにおいても 2～8 Hz の変調周波数バンドが重要であることがわかった．また，話者識別情報は，16 Hz 以上にも存在する可能性が示された．さらに人間は 0 Hz 付近の変調スペクトル（静的な情報）についても，有効に利用している可能性があることがわかった．

参考文献

[1] R. Drullman, J. M. Festen, and R. Plomp (1994), "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Amer., Vol. 95, pp. 1053 – 1064.

[2] R. Drullman, J. M. Festen, and R. Plomp (1994), "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Amer., Vol. 95, pp. 2670 – 2680.

[3] T. Arai, M. Pavel, H. Hermansky and C. Avendano (1996),"Intelligibility of speech with filtered time trajectories of spectral envelopes," In Proc. of the ICSLP, Philadelphia, pp. 2490 – 2493.

[4] S. Greenberg (1996), "Understanding speech understanding — Towards a unified theory of speech perception," In Proc. of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, Keele, England, pp. 1 – 8.

[5] T. Arai, M. Pavel, H. Hermansky and C. Avendano (1999), "Syllable intelligibility for temporally filtered LPC cepstral trajectories," J. Acoust. Soc. Amer., Vol. 105, No. 5, pp. 2783 – 2791.

[6] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel (1997), "On the importance of various modulation frequencies for speech recognition," Proc. Eurospeech, Rhodes, Greece, pp. 1079 – 1082.

[7] N. Kanedera, H. Hermansky and T. Arai (1998), "On properties of modulation spectrum for robust automatic speech recognition," Proc. IEEE ICASSP, Seattle, WA, pp. II-613 – II-616.

[8] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel (1999), "On the relative importance of various components of the modulation spectrum for automatic speech recognition," Speech Communication, Vol.28, pp.43 – 55.

[9] S. van Vuuren and H. Hermansky (1998), "On the importance of components of the modulation spectrum for speaker verification," in ICSLP, Sydney Australia.

[10] 金寺 登, 荒井隆行, H. Hermansky, 船田哲男 (1997), "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," 電子情報通信学会 技術研究報告, SP97-70, pp.15-22.

[11] H. Hermansky (1990), "Perceptual linear predictive (PLP) analysis for speech," J. Acoust. Soc. Amer., Vol. 87, No. 4, pp. 1738 – 1752.

[12] 徳田恵一ほか, "音声信号処理ツールキット," http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/

[13] T. Houtgast and H. J. M. Steeneken (1985), "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in

auditoria," J. Acoust. Soc. Amer., Vol. 77, pp. 1069 – 1077.

[14] 金寺 登, 荒井隆行, 船田哲男 (1998), "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," 電子情報通信学会 技術研究報告, SP98-51, pp.45-52.

[15] H. Hermansky and N. Morgan (1994), "RASTA processing of speech," IEEE Trans. Speech and Audio Process., Vol. 2, No. 4, pp. 578-589.