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.
SP2002-64(2002-07)

変調スペクトルの貢献度に基づく連続音声認識

金寺 登† 荒井 隆行†† 岡田 賢治†† 百村 裕智††

† 石川高専 〒929-0392 石川県河北郡津幡町北中条
†† 上智大学 〒102-8554 東京都千代田区紀尾井町7-1
E-mail: †kane@i.ishikawa-nct.ac.jp, ††arai@sophia.ac.jp

あらまし 音声特徴量の時間軌跡をフーリエ変換したものは変調スペクトルと呼ばれ，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．本報告では，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に応じて変調スペクトルを強調した音声認識特徴量を提案する．自動音声認識実験の結果，提案した特徴量は，雑音環境下において音声認識性能が約5% 改善されることを確認した．

キーワード 音声認識特徴量，変調スペクトル，変調周波数，雑音

Continuous speech recognition based on the contribution of modulation frequency components

Noboru KANEDERA†, Takayuki ARAI††, Kenji OKADA††, and Yasunori MOMOMURA††

† Ishikawa National College of Technology, Tsubata, Ishikawa, 929-0392 Japan
†† Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan
E-mail: †kane@i.ishikawa-nct.ac.jp, ††arai@sophia.ac.jp

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 propose new feature for automatic speech recognition based on contribution of modulation frequency components. The contribution shows the importance of each modulation frequency component for speech recognition. Testing proposed feature on IPA98 task in noisy environments (SNR 10 dB) gave a relative improvement of 5 % in word accuracy over the MFCC with dynamic feature.

Key words feature for automatic speech recognition, modulation spectrum, modulation frequency, noise

1. はじめに

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]～[5],[12]．変調スペクトルとは，音声認識特徴量の時間軌跡をフーリエ変換したもので音声のダイナミクスを表している．

知覚実験[1]～[3]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[4]や英語[6]においても確認されている．Drullmanら[2],[3]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[4],[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[10]～[12]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．

一部の変調スペクトルを選択的に利用すると，雑音環境下での音声認識性能が向上することが，英語数字音声認識実験によって示された[17]．一部の変調スペクトルを選択的に取り出すためにバンド幅の異なる3つのバンドパスフィルタを使用する方法「Modulation FT」を用いた．さらに，百村らは異なるバンド幅を制御しやすい「Modulation Wavelet」を用いて，同様の実験を行い，良好な結果を得た[18]．

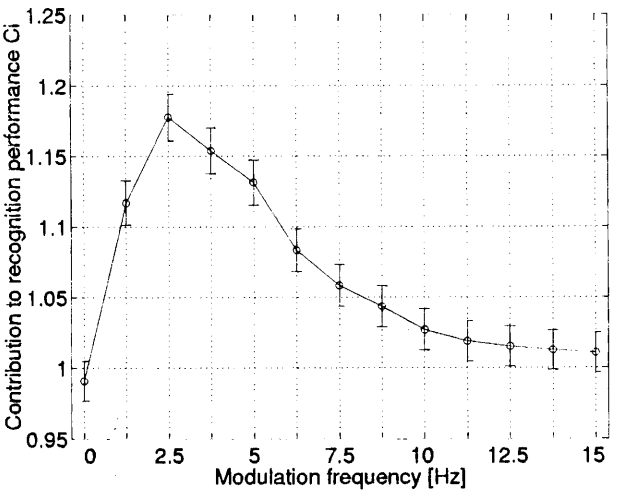

図 1　連続音声認識に対する変調スペクトル貢献度

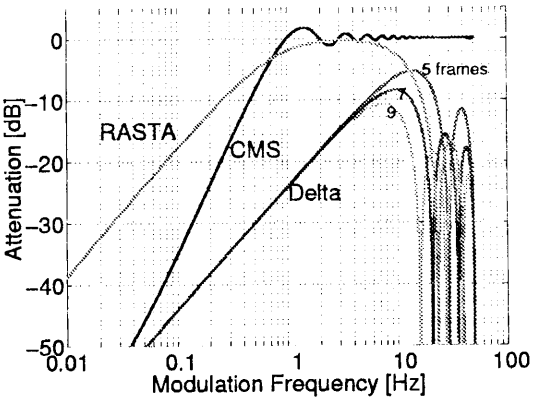

図 2　CMS, Delta, RASTA の変調周波数特性

Fig. 2　Modulation-frequency characteristics of CMS, Delta, and RASTA.

本報告では連続音声についても変調スペクトルの選択的利用が有効かどうかを調査する．また，各変調スペクトル成分の重要性を表す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を提案し，その有効性を示す．

まず第 2 節では，変調スペクトルを選択的に利用した音声認識特徴量について述べる．第 3 節では，word モデル及び tri-phone モデルによる単語音声認識実験結果を報告する．実験では，「Modulation FT」などに加えて，各変調スペクトル成分の重要性を表す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を用いた結果についても示す。第 4 節では連続音声認識実験結果を報告する．

2. 変調スペクトルを選択的に利用した音声認識特徴量

2.1 変調スペクトル成分の貢献度

文献 [12], [16] 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し，各変調スペクトル成分がどのくらい重要であるのかを定量的に調査した．この貢献度は，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．

文献 [16] 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．この結果より，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 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．

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(RelAtive SpecTrAl processing) [8] が知られている．JC-RASTA [9] 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図 2 の変調周波数特性を持つ RASTA フィルタを用いて，約 1〜12 Hz の成分のみを抽出する．この変調周波数バンドは，図 1 において調査した重要な変調周波数バンドを包含している。

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．

2.3 Modulation FT

2.2 節の RASTA は，図 1 で示される重要な（貢献度の高い）変調周波数バンドを 1 つのバンドパスフィルタで抽出する方法である．これに対し，さらに効率的に重要な変調スペクトル成分を表現するため，複数の変調スペクトル解像度を持つ変調周波数フィルタを用いる方法を提案した [17]．

図 3 の変調周波数特性をもつ複数の変調周波数フィルタを用いて英語単語音声認識実験を行ったところ，一般によく用いられている動的特徴量 [21] を含めた MFCC に比べ雑音環境下 (SNR 10dB) において，認識誤り率が平均 8% 改善された．図 3 では，特定の変調周波数バンドを抽出するフィルタとして DFT の一部の成分を利用した．図では 32 点 DFT の第 2，第 3 成分に加えて，低周波数成分を表現するため 64 点 DFT の第 2 成分を用いた．解像度の異なる複数の DFT 成分を利用することにより効率的に重要な変調スペクトル成分を表現している．

2.4 変調スペクトルの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 [22]．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．

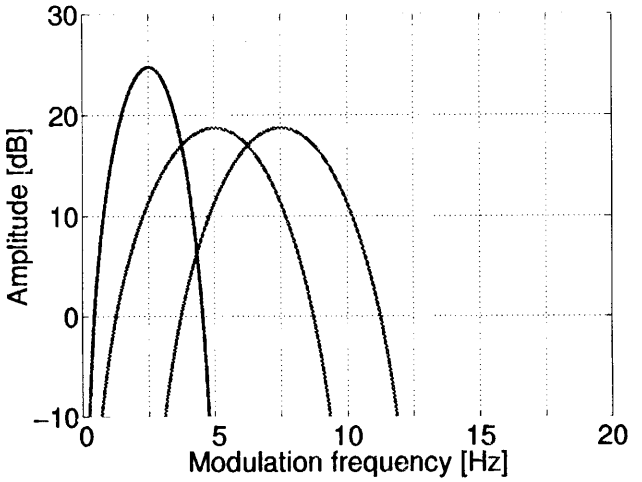

図 3 Modulation FT の変調周波数特性

3. 変調スペクトルを選択的に利用した単語音声認識

一部の変調スペクトルを選択的に利用すると，雑音環境下での音声認識性能が向上することが，英語数字音声認識実験によって示された [17]．本節では日本語の一般単語についても変調スペクトルの選択的利用が有効かどうかを調査するため，日本語 216 単語音声認識実験を行った結果について報告する．また連続音声にも適用しやすいように，HMM には単語単位のモデルだけでなく，tri-phone モデルを使用した場合の結果についても報告する．

3.1 日本語 216 単語音声認識実験条件

ATR 音声データベース [23] セット C の 216 単語，男女各 15 名分を学習に用い，各 5 名分を評価に用いた．雑音データには NOISEX-92 [15] 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，付加雑音を SNR 0 dB になるように音声データに波形レベルで加算したものを用いた．付加雑音には各雑音データの中よりランダムに切り出したものを用いた．これらの学習・評価データを Jack-knife 方式で 4 組用意した．HMM には単語単位のモデル (8 状態 6 出力分布，混合数 2) を用い，離散単語認識を行った．

3.2 Word モデルを用いた日本語 216 単語音声認識実験

3.1 節の実験条件を用い，種々の雑音環境下において，各特徴量による音声認識実験結果を表 1 に示す．表中の clean は評価データに雑音を付加しない場合である．その他は，雑音を付加した場合に対応している．mean は，雑音環境下での平均単語誤り率を示している．図 3 の変調スペクトルを用いる方法 (Modulation FT) は，広く用いられている MFCC と比較し，雑音環境下での認識性能が向上している．なお、MFCC と PLP [7] には CMS [20] を施し，動的特徴量 (Δ) を併用した．

Modulation FT は 2.3 節の特徴量を使用した場合で，「実部」はフーリエ変換の実部のみを用いた場合，「実部&虚部」はフーリエ変換の実部と虚部の両方を用いた場合である．特徴量の数が近い MFCC と Modulation FT (実部のみ) を比較すると 15 種類の雑音に対して，平均で 20%も単語誤り率が改善されている．日本語の一般単語についても変調スペクトルの選択的利用が有効であることが確認できた．

表 1 Word モデルによる単語音声認識実験結果 (単語誤り率 [%])

	特徴量	MFCC +Δ	PLP +Δ	PLP-RI +Δ+Δ^2	Modulation FT 実部	Modulation FT 実部&虚部
clean		9.8	8.1	4.4	2.5	1.8
	babble	42.6	44.9	41.5	16.8	16.4
	buccaneer1	31.8	29.7	29.9	18.7	12.6
付	buccaneer2	26.2	25.9	27.3	19.5	14.3
	destroyerengine	43.6	49.8	36.4	25.3	17.8
	destroyerops	34.2	30.9	36.1	11.0	8.4
加	f16	36.4	30.2	31.3	17.4	12.8
	factory1	33.9	30.8	32.8	16.1	11.3
	factory2	36.5	36.2	28.2	10.8	7.7
雑	hfchannel	60.1	49.7	26.4	27.3	20.4
	leopard	39.8	46.2	42.2	6.6	4.9
	m109	30.3	29.9	27.2	8.1	6.3
音	machinegun	54.9	56.7	64.1	32.8	30.1
	pink	26.7	24.5	23.6	15.6	11.4
	volvo	24.2	27.4	11.2	3.4	2.2
	white	34.1	40.0	21.0	26.4	20.0
mean		37.0	36.9	32.0	17.0	13.1
feature size		25	17	26	27	54

PLP-RI は 2.4 節の特徴量を使用した場合に対応する．Word モデルの場合には，PLP-RI は Modulation FT には及ばないが，MFCC と比較し，15 種類の雑音に対して，単語誤り率が平均 5%改善されている．

3.3 Tri-phone モデルを用いた日本語 216 単語音声認識実験

連続音声を扱うためには，学習に使用できる音声データの制約により，単語より小さい単位 (tri-phone モデルなど) を使用するのが一般的である．

3.1 節の実験条件を用い，種々の雑音環境下において，tri-phone モデルを用いた音声認識実験結果を表 2 に示す．3.1 節では音響モデルに単語単位の HMM モデルを用いたのに対し，ここでは 2000 状態 16 混合の tri-phone モデルを用いた．

表 1 の単語モデルによる結果と表 2 の tri-phone モデルによる結果を比較すると，MFCC, PLP, PLP-RI の認識性能が向上しているのに対して，Modulation FT の認識性能が劣化している．

MFCC, PLP, PLP-RI に使用した動的特徴量は図 2 に示すように主として 10 Hz 付近の変調スペクトルを強調している。一方 Modulation FT は図 3 に示すように 2.5 Hz, 5 Hz, 7.5 Hz 付近の変調スペクトルを強調しており動的特徴量に比べ注目している変調周波数が低い．一般的に低い変調周波数バンドをモデル化するには，認識単位が大きい必要がある．

したがって，単語モデルから tri-phone モデルに認識単位を小さくすることによって Modulation FT の認識性能が劣化したものと考えられる．これより，tri-phone モデルを用いた場合には 3.2 節のような大幅な認識性能向上が期待できず，Modulation FT の効果が生かせないことがわかった．

表 2 Tri-phone モデルによる単語音声認識実験結果 (単語誤り率 [%])

	特徴量	MFCC +Δ	PLP +Δ	PLP-RI +Δ+Δ^2	Modulation FT 実部	実部&虚部
clean		5.6	5.7	3.5	8.5	9.4
	babble	54.9	68.7	31.1	48.2	21.0
	buccaneer1	13.2	19.4	6.3	19.9	20.9
付	buccaneer2	14.7	23.8	7.4	20.1	22.1
	destroyerengine	14.9	34.9	8.7	27.0	24.5
	destroyerops	37.8	56.8	19.0	21.0	17.1
加	f16	12.7	21.6	7.4	18.3	20.1
	factory1	16.4	28.5	8.2	19.1	19.4
	factory2	13.0	23.8	5.4	14.2	16.0
雑	hfchannel	36.3	30.5	7.7	25.5	27.3
	leopard	42.9	48.6	11.2	16.6	14.2
	m109	18.8	38.0	6.5	14.9	14.7
音	machinegun	51.8	55.9	60.7	49.3	32.4
	pink	8.6	12.5	5.1	16.0	19.2
	volvo	8.1	8.9	4.0	9.2	10.8
	white	12.2	15.8	6.8	21.4	26.5
mean		23.7	32.5	13.0	22.7	20.4
feature size		25	17	26	27	54

表 3 連続音声認識結果 (単語正解精度 [%])

特徴量	特徴量数	clean	SNR 20dB	10dB	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+Δ+Δ^2	26	92.6	89.8	81.5	58.0
(7) PLP-RI+Δ+Δ^2	26	91.4	90.1	84.5	63.9
(8) MFCC+Δ+Δ^2	38	92.6	90.1	84.1	63.0

MFCC, PLP, PLP-RI の場合は，Modulation FT より注目する変調周波数が高い動的特徴量を用いているため，認識単位を小さくした悪影響が少なく，逆にモデル当りの学習データ量増加の影響で認識性能が向上したものと思われる．

4. 変調スペクトルの貢献度に基づく連続音声認識

4.1 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) [19] 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 [19] 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．

表 4 連続音声認識結果 1 (SNR 10 dB での単語正解精度 [%])

	特徴量	PLP +Δ	JC-RASTA +Δ	PLP-RI +Δ
clean		90.2	85.8	90.3
	babble	75.8	79.0	81.6
	buccaneer1	77.0	77.4	82.5
付	buccaneer2	77.6	77.7	85.0
	destroyerengine	68.7	73.4	75.0
	destroyerops	79.5	80.6	79.8
加	f16	76.0	78.0	82.6
	factory1	77.2	78.3	86.1
	factory2	82.1	81.2	86.3
雑	hfchannel	66.8	75.9	78.4
	leopard	76.6	82.2	82.0
	m109	85.3	82.0	86.3
音	machinegun	47.1	42.3	42.8
	pink	78.9	78.4	84.0
	volvo	87.2	85.2	89.0
	white	71.9	75.5	81.3
mean		75.2	76.5	80.2
feature size		17	17	17

表 5 連続音声認識結果 2 (SNR 10 dB での単語正解精度 [%])

	特徴量	MFCC +Δ	MFCC -RI+Δ	PLP +Δ+Δ^2	PLP-RI +Δ+Δ^2	MFCC +Δ+Δ^2
clean		92.1	91.4	92.6	91.4	92.6
	babble	81.8	81.3	84.0	86.6	87.3
	buccaneer1	80.1	80.9	82.6	85.9	83.8
付	buccaneer2	79.1	79.0	82.1	86.6	83.8
	destroyerengine	77.9	80.8	77.5	84.4	85.0
	destroyerops	84.1	85.9	85.8	87.2	89.2
加	f16	80.5	80.9	81.9	86.4	85.9
	factory1	80.3	81.3	81.8	87.0	85.3
	factory2	86.4	86.7	88.7	90.4	89.3
雑	hfchannel	69.3	71.2	77.5	85.7	79.6
	leopard	84.1	85.1	86.1	87.9	88.7
	m109	87.5	88.8	88.6	91.0	88.8
音	machinegun	56.3	57.0	51.7	44.7	58.1
	pink	81.5	81.3	82.4	86.3	85.1
	volvo	91.6	88.9	92.8	90.8	90.9
	white	73.7	74.9	78.7	86.7	80.7
mean		79.6	80.3	81.5	84.5	84.1
feature size		25	25	26	26	38

4.2 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 3,4,5 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 **2.4** 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 2.4 節の方法で変調フィルタリングを行った特徴量

(6) PLP+Δ+Δ^2 ··· (1) の特徴量に，Δ^2 特徴量を加えた特徴量

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

(8) MFCC+Δ+Δ^2 ··· (4) の特徴量に，Δ^2 特徴量を加えた特徴量

表 3 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 [4] 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)(7) 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 (7) 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 (7) では約 5% 認識率が改善されていることがわかった．

5. ま と め

雑音環境下での日本語 216 単語音声認識実験の結果，認識単位を単語とした場合，変調スペクトルを選択的利用する Modulation FT は，広く用いられている MFCC と比較し，平均 20%の単語誤り率が改善が認められた．これより日本語の一般単語についても変調スペクトルの選択的利用が有効であることが確認できた．

また、認識単位を tri-phone とした場合にはこのような大幅な認識性能向上が期待できず，Modulation FT の効果が生かせないことがわかった．

そこで認識単位を tri-phone とした場合でも，認識性能向上が期待できる方法として，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た．雑音環境下 (SNR 10 dB) での連続音声認識実験の結果，MFCC と比較し音声認識性能が約 5% 改善されることを確認した．

文 献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[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] 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.

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

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

[9] H. Hermansky, N. Morgan and H. Hirsch (1993), "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," Proc. IEEE ICASSP, Minneapolis, MN, pp. II-83 – II-86.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[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] A. Varga and H. J.M. Steeneken (1993), "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, Vol. 12, No. 3, pp. 247 - 251.

[16] 金寺 登, 荒井隆行, 船田哲男 (2000), "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 –," 電子情報通信学会 技術研究報告, SP2000-34, Vol.100, No.240, pp.15 – 22.

[17] 金寺 登, 荒井 隆行, 船田哲男 (2001), "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 DII, Vol.J84-D-II,No.7,pp.1261-1269.

[18] Y. Momomura, K. Okada, T. Arai, N. Kanedera, Y. Murahara (2001), "Using the Modulation Complex Wavelet Transform for Feature Extraction in Automatic Speech Recognition," Proceedings of European Conference On Speech Communication And Technology, Vol.4, pp.2639 - 2642.

[19] 河原 達也, 李 晃伸, 小林 哲則, 武田 一哉, 峯松 信明, 伊藤克亘, 山本 幹雄, 山田 篤, 宇津呂 武仁, 鹿野 清宏 (2000), "日本語ディクテーション基本ソフトウェア (98 年度版)," 音響学会誌, **56**, 4, pp. 255 – 259.

[20] B. S. Atal (1974), "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., Vol. 55, No. 6, pp. 1304 - 1312.

[21] S. Furui (1986), "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-34, No. 1, pp. 52 – 59.

[22] V. Johan, B. Louis (1998), "Channel normalization techniques for automatic speech recognition over thetelephone," Speech Communication, Vol.25, pp.149 - 164.

[23] 桑原 尚夫, 匂坂 芳典, 武田 一哉, 阿部 匡伸 (1989), "研究用 ATR 日本語音声データベースの作成," ATR Technical Report, TR-I-0086.