社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（47）
2005／12／22

MSD-HMMに基づく音声のスタイル識別

川島　啓吾†　　橘　　誠†　　山岸　順一†　　小林　隆夫†

† 東京工業大学 大学院総合理工学研究科 物理情報システム専攻
〒 226-8502 横浜市緑区長津田町 4259-G2-4
E-mail: †{keigo.kawashima,makoto.tachibana,junichi.yamagishi,takao.kobayashi}@ip.titech.ac.jp

あらまし　本論文では，多空間上の確率分布 (MSD) に基づく HMM を用いた音声の感情・発話様式の識別について検討している．MSD-HMM により音声のスペクトル情報と基本周波数 (F0) の同時モデル化を行い，複数の話者の平静調音声で学習されたユニバーサルバックグラウンドモデル (UBM) を目標話者・スタイルの少量の文章によりモデル適応し，話者及びスタイルの同時適応を行ったモデルを用いて識別を行っている．まず MSD-HMM を用いて特徴量に F0 を含めることで識別率が改善することを示し，次に，適応時の初期モデルとして UBM を用いる場合と，目標話者の読上げ音声から作成した話者依存モデルを使用する場合の比較を行い，UBM を用いて話者とスタイルの同時適応を行った場合においても，話者依存モデルと同等の性能が得られることを示す．最後に，ナレーション経験のない話者の音声を用いた評価実験を行った結果を示す．

キーワード　感情音声，発話様式，MSD-HMM，スタイル識別，話者適応，韻律的特徴

Style classification of speech based on MSD-HMM

Keigo KAWASHIMA†, Makoto TACHIBANA†, Junichi YAMAGISHI†, and Takao KOBAYASHI†

† Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology,
4259-G2-4, Nagatsuta-cho, Midori-ku, Yokohama-shi, 226-8502, Japan
E-mail: †{keigo.kawashima,makoto.tachibana,junichi.yamagishi,takao.kobayashi}@ip.titech.ac.jp

Abstract　This paper describes a classification technique of emotional expressions and speaking styles of speech based on multi-space probability distribution HMM (MSD-HMM). By using MSD-HMM, we model spectral and fundamental frequency (F0) features simultaneously. A universal background model (UBM) is trained by using neutral style speech data of multiple speakers and then adapted to the target speaker and style using a small amount of speech data. In this study, first, we investigate the effect of the use of F0 and show that including F0 in the feature vector improves the classification rate. Then, we compare the performance of speaker and style adapted UBM with that of speaker dependent model trained by target speaker's neutral style data and show that classification result of the adapted UBM are close to that of speaker dependent model. We also perform classification experiments using recorded speech by unprofessional speakers.

Key words　emotional speech, speaking style, MSD-HMM, style classification, speaker adaptation, prosodic features

1. まえがき

近年，コールセンター，コミュニケーションロボット，カーナビゲーションシステム／運転支援システムなどへの応用を目的として，音声に含まれる感情・発話様式を識別・検出するための研究が行われている [1]〜[4]．例えばコールセンターへの応用では，音声から話し手の「苛立ち」や「焦り」などの感情を検出し，表出度合を推定することで，顧客の対応優先度の決定などに役立てることが可能となる．

音声に含まれる感情表現や発話様式は，それぞれが単独で現れる場合だけでなく，複合的に現れる場合や個人によってもその定義が異なることから，ここではそれらが単独あるいは複合的に含まれた一つの話し方のことを音声のスタイルと呼ぶことにする．

音声のスタイルの特徴は，発話内容などの言語情報や声質情報であるスペクトルパラメータだけでなく，基本周波数 (F0) パターンやテンポなどの韻律情報にも大きく現れることが知られている [1]．そのため，スタイルの識別には，これらの韻律情報，

特に F0 の利用が不可欠であるが，F0 は有声音区間では 1 次元の連続値となるものの，無声音区間では値を持たないため，通常の GMM や HMM で直接モデル化できない．そこで本研究では，多空間上の確率分布に基づいた HMM(MSD-HMM) [5] を用いることで，音声の各スタイルのスペクトルと F0 の特徴を同時にモデル化する．さらに，最尤線形回帰 (Maximum Likelihood Linear Regression: MLLR) [6] を MSD-HMM に拡張した MSD-MLLR [7] を用いて，話者とスタイルを同時に適応することで識別用モデルを作成し，スタイル識別への応用について検討する．

2. MSD-HMM による音声のスタイル識別

2.1 スペクトルと基本周波数のモデル化

感情や発話様式などを伴う音声の特徴は，平均 F0 が変わる，F0 の変動幅が広がるといった全体的な特徴の他に，F0 の局所的な変動が激しくなる，あるいは緩やかになるなどといった時系列上の情報の特徴としても現れる．従って，スタイル識別のためには，対象とする韻律情報の時系列変動を適切に表現可能なモデルを導入することが重要であると考えられる．時間方向の変動のモデル化には HMM が適していると考えられるが，F0 は有声音区間では 1 次元の連続値を持ち，無声音区間では無声という離散シンボルを持つ，連続値と離散シンボルの混合系列であり，通常の HMM では直接モデル化できない．このような次元が異なる数種の特徴が混合した系列に対しても，HMM の枠組みを適用可能とするためにこれを拡張したものが，多空間上の確率分布に基づいた HMM(MSD-HMM) [5] である．MSD-HMM を用いると一つの状態に離散分布と連続分布を混在させることが可能であり，連続値と離散シンボルの混合系列である F0 を直接モデル化することができる．

本論文では，F0 を有声区間に対応する 1 次元空間と無声区間に対応する零次元空間の 2 つの空間から出力される観測事象と考え，MSD-HMM でモデル化している．さらに，スペクトル情報も同時に利用するため，スペクトルと F0 を独立のストリームでモデル化したマルチストリーム MSD-HMM を識別に利用している．

2.2 MSD-HMM に基づくスタイル識別システム

MSD-HMM に基づくスタイル識別システム [8] を図 1 に示す．本システムの構成は，話者認識において一般的に使用されている HMM に基づくシステムと同様の構成である．

システムの学習部では，複数の話者の「平静 (読上げ)」調の音声で構成されるデータベースを用いて，ユニバーサルバックグラウンドモデル (UBM) [9] をマルチストリーム MSD-HMM により音素ごとに作成し初期モデルとする．

識別用スタイル登録部では，識別したいスタイルのモデルを作成する．各スタイルのモデルは，UBM に対し目標とする話者・スタイルの少量の音声データを用いて話者とスタイルとを同時にモデル適応を行うことで作成する．モデル適応には，MLLR [6] を MSD-HMM に拡張した MSD-MLLR [7] を用いスペクトル・F0 の同時適応を行っている．

識別部では，モデル適応によって作成された各識別用モデルを用いて入力スタイルの尤度を計算し，全モデルの中で最も高い尤度のモデルのスタイルを，入力音声のスタイルであると判定する．

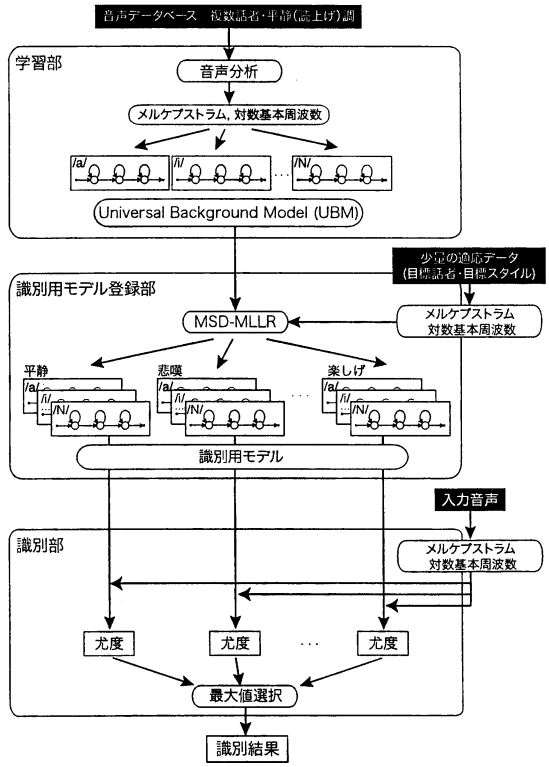

図 1　スタイル識別システム

3. 実　　験

3.1 実験条件

識別に用いる音声データは，男性ナレータ 2 名，女性ナレータ 1 名が「平静 (読上げ)」「悲嘆」「楽しげ」「ぞんざい」の各スタイルで発声した ATR 音韻バランス文 503 文章を用いた．なお，このうち男性ナレータ 1 名，女性ナレータ 1 名のデータは文献 [10] と同様のデータである．

サンプリングレート 16kHz の音声信号を，フレーム長 25ms，フレーム周期 5ms のブラックマン窓を用いてメルケプストラム分析し，0 次から 24 次のメルケプストラムを求め，これに対数基本周波数 (対数 F0) を加えたパラメータを特徴パラメータとした．特徴パラメータに動的特徴は含まず，メルケプストラムは，各次元の平均除去を文章単位で行っている．HMM は 3 状態の left-to-right monophone でモデル化した．音素は無音を含んだ 42 種類である．

HMM の学習データは目標話者を含まない各話者 50 文章ずつ，100 文章の「平静 (読上げ)」調音声を基に UBM を作成し，初期モデルとした．識別用スタイルのモデルは，最大で 50 文章の目標話者・スタイルの適応データを用いて各スタイル毎に作成した．スタイル識別は，学習・適応データに用いなかった文章の中から 400 文章を選んで評価した．この操作を 5-fold

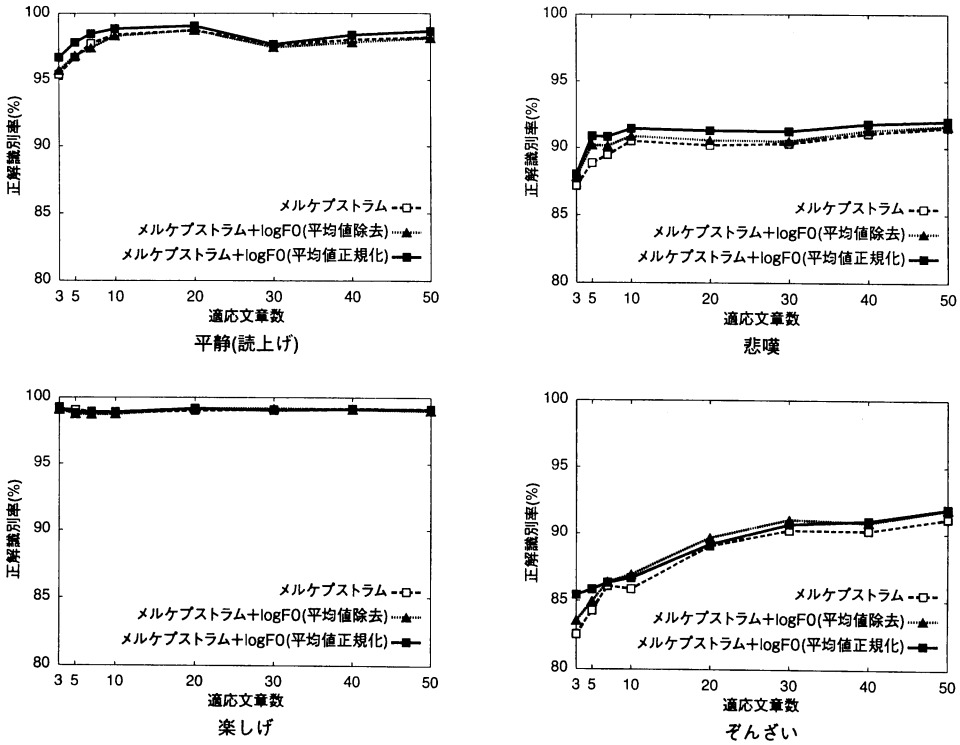

図 2　特徴量による識別率の比較

cross validation で全ての音声データに対して行った．

3.2　特徴量の比較

まず HMM の学習及び適応データの特徴量としてメルケプストラムと対数 F0 を用いた場合と，メルケプストラムのみを用いた場合について識別率の比較を行った．また対数 F0 の平均値の影響を調べるために，平均値除去を文章単位で行った場合と，各スタイルごとの学習データ全体の平均値に文章単位で正規化を行った場合について識別率の比較を行った．なお対数 F0 を学習データの平均値に正規化したモデルに対する評価では，評価データの対数 F0 の値は正規化せず，抽出した値をそのまま用いている．本実験では，識別時に入力音声の書き起こしテキストラベルを与えた．

図 2 に適応文章数に対する識別結果を，評価スタイルごとに示す．適応文章数は 3 文章から 50 文章である．図中のメルケプストラム＋ logF0(平均値除去) は特徴量としてメルケプストラムに対数 F0 の平均値除去を文章単位で行ったもの，またメルケプストラム＋ logF0(平均値正規化) は各スタイルごとの学習データ全体の平均値に文章単位で正規化を行ったものを加えて識別を行った結果である．これらの図より，全体的な特徴として「平静」,「楽しげ」に比べると「悲嘆」,「ぞんざい」の識別率が下がる傾向にあることがわかるが，いずれのスタイルでも特徴量としてメルケプストラムのみを用いた場合においても高い識別率が得られていることがわかる．これは実験に用いた音声データがプロのナレーターによって上手に演技された感情表現・発話様式の音声であったためと考えられる．また，適応文章数が 10 文章から 20 文章程度で概ね安定した識別率が得られていることがわかる．対数 F0 の効果に関しては,「楽しげ」に関しては識別率が高いため差が見られないが，他のスタイルに関しては特徴量に対数 F0 を含めることで識別率が若干改善していることがわかる．また対数 F0 の平均値の効果については,「楽しげ」に関しては同様に差が見られないが，他のスタイルにおいては特に適応文章数が少ない場合において識別率が改善する傾向にあることがわかる．これらの結果より以降の実験では学習データにおける対数 F0 の特徴量として，各スタイルごとの学習データ全体の平均値に文章単位で正規化を行ったものを使用する．

3.3　初期モデルの比較

次に識別用スタイルモデルを作成する際，特定話者のスタイルのみを適応する場合と，話者とスタイルを同時に適応する場合で，どの程度識別に影響するかを調査べるために，適応時の初期モデルとして，目標話者を含まない複数話者の「平静 (読上げ)」調音声を基に作成された UBM を使用する場合と，目標話者の「平静 (読上げ)」調音声を基に作成された話者依存 (Speaker Dependent: SD) モデルを使用する場合について識別率の比較を行った．

話者依存モデルを用いた識別用スタイルモデルは，目標話者の 100 文章の「平静 (読上げ)」調音声を基に作成されたモデルを初期モデルとし，最大で 50 文章の目標話者・スタイルの適応データを用いて各スタイル毎に作成した．また，識別時には入力音声の書き起こしテキストラベルを与えた．

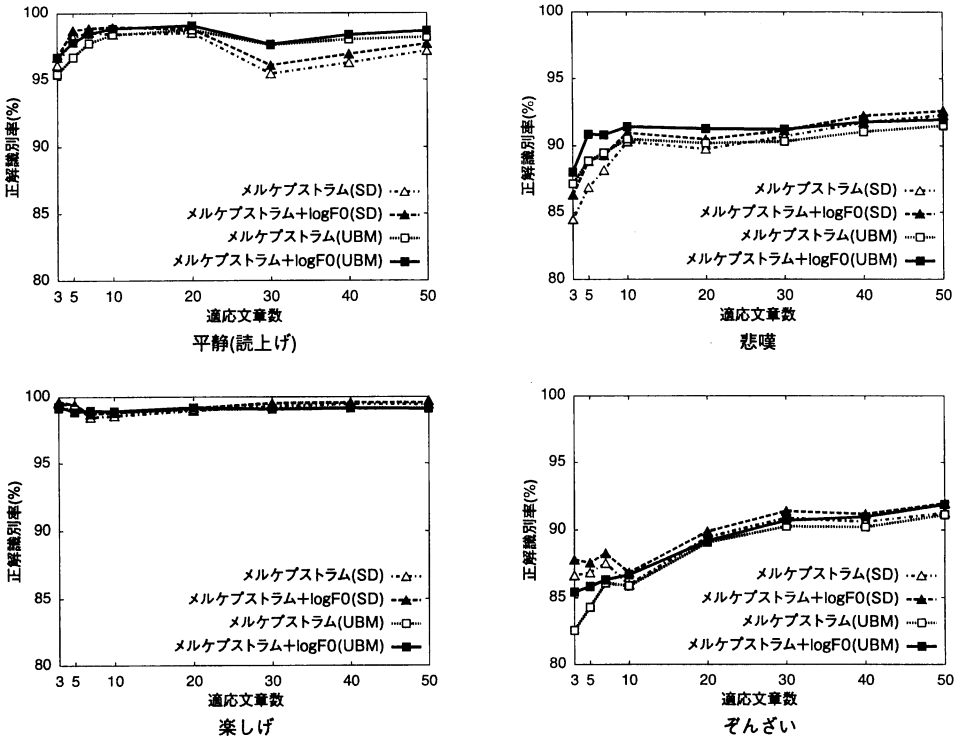

図 3　初期モデルによる比較

図 3 に適応文章数に対する各初期モデルによる識別結果を示す．これらの図より，特定話者のスタイルのみを適応する場合と話者とスタイルを同時に適応する場合において，識別率の上では大きな違いがないことがわかった．

3.4　書き起こしテキストの影響

次にスタイルの識別時に入力音声の書き起こしテキストラベルを使用する場合と，書き起こしテキストラベルを使用しない場合について識別率の比較を行った．この実験では，実際にスタイル識別を行う際には，入力音声の正確な書き起こしテキストが使用できないことが一般的と考えられることから，入力音声の書き起こしテキストの使用がどの程度識別に影響するかを調べている．

図 4 に適応文章数に対する書き起こしテキストを使用した場合と使用しない場合の識別結果を示す．この図において，「書き起しあり」は入力音声の書き起こしテキストを使用しビタビアルゴリズムを行った結果で，「書き起しなし」は入力音声の書き起こしテキストを使用せず，簡単な音素ネットワークのみでビタビアルゴリズムを行った結果を示している．これらの図より，「平静」，「楽しげ」に関しては識別率が高いために二つの手法に大きな差はみられないが，全体的に書き起こしテキストを使用しない方の識別結果が若干下がる傾向であることがわかる．しかし，いずれのスタイルにおいても書き起こしテキストを使用しない場合においても，大幅な識別率の低下はみられなかった．

4.　ナレーション経験のない話者の音声を用いた識別実験

4.1　音声の収録

最後にナレーション経験のない話者の音声を用い，プロのナレータ音声との識別結果の違いを比較した．これは 3. の実験で使用した音声が，感情や発話様式を演技することに長けているナレータによる発話音声であったのに対し，感情や発話様式の演技経験のほとんどない話者による音声を用いることで，より現実に近い状況での評価を行うためである．自然発声で生じた各スタイルの音声を用いることがより適切であると考えられるが，このような音声の収録は容易ではないため，各自の日常の感情・発話様式により指定した発話をしてもらった．

識別対象のスタイルとして「平静 (読上げ)」「悲嘆」「楽しげ」「苛立ち」「怒り」の 5 種類を設定した．ここでは「ぞんざい」スタイルに替えて，発話者がより日常的に発声すると考えられる類似スタイルの「苛立ち」，さらに発話者にとって比較的発声が容易だと考えられるスタイルの「怒り」を設定した．

男性話者 3 名，女性話者 2 名の話者に ATR 音韻バランス文 50 文章の各スタイルと，それぞれのスタイルが発声しやすいと考えられる評価用 10 文章をそのスタイルで発声してもらい防音室で収録を行った．付録 1. に収録した評価用文章を示す．

なお，音声のスタイルは，各話者の主観に基づいて発声してもらったものである．

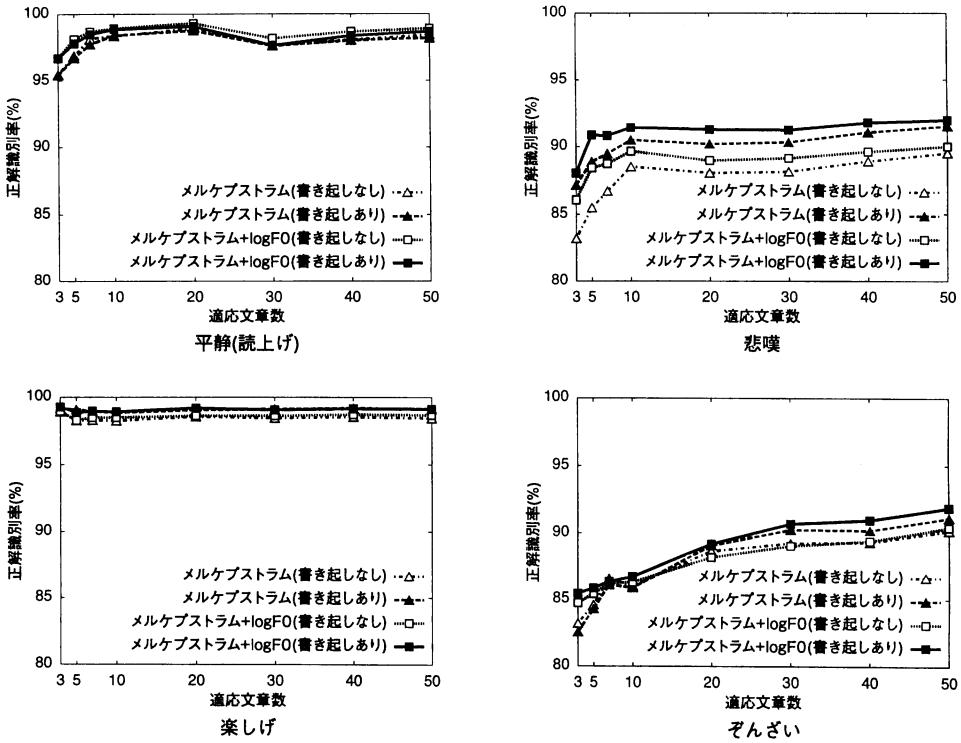

図 4　入力音声の書き起こしテキストの影響の比較

表 1　収録音声の主観評価結果

	Classification (%)				
	平静	悲嘆	楽しげ	苛立ち	怒り
平静	91.3	6.0	2.0	0.0	0.7
悲嘆	0.7	97.3	0.0	2.0	0.0
楽しい	2.7	0.7	95.3	1.3	0.0
苛立ち	8.0	5.3	1.3	59.3	26.0
怒り	8.0	5.3	0.0	17.3	69.3

表 2　提案手法による収録音声の評価結果

	Classification (%)				
	平静	悲嘆	楽しげ	苛立ち	怒り
平静	82.0	3.6	12.4	1.2	0.8
悲嘆	3.2	93.2	2.0	0.8	0.8
楽しい	0.0	2.0	96.8	0.0	1.2
苛立ち	0.0	2.0	0.0	69.6	28.4
怒り	0.0	1.2	10.8	14.8	73.2

4.2　収録音声の評価

まず収録した音声について，カテゴリーテストによる主観評価試験を行った．収録話者と直接面識のない第三者である 3 名の被験者に，各音声のスタイルが “平静”，“楽しい”，“悲嘆”，“苛立ち”，“怒り” のどれに聞こえるかを判定してもらった．テストデータは評価用に収録した 5 名の話者の各スタイル 10 文章とし，被験者毎にランダムな順序で提示し評価を行った．

表 1 に主観評価結果を示す．表中の数値は各テスト音声がどのスタイルに認識されたかを割合で示す．この結果より，主観評価試験においては「平静」,「楽しげ」,「悲嘆」については，ほとんどが収録時に設定したスタイルに判断されていることがわかる．一方,「苛立ち」,「怒り」のスタイルでは一部が混同されて識別されており，収録時のスタイルに判断された割合が低くなっている．これは，設定した「苛立ち」,「怒り」のスタイルが類似しており，発話者・評価者ともに判別が困難であったためと考えられる．

4.3　収録音声の識別実験

次にスタイル識別システムを用いて，評価音声の識別実験を行った．初期モデルとなる UBM は，ナレータ 3 名の各話者 400 文章，計 1200 文章の「平静 (読上げ)」調音声を基に作成し，モデルの適応・作成には収録した ATR 音韻バランス文 50 文章の中から 20 文章を選び用いている．適応文章のばらつきを抑えるために 20 文章の組合せを 5 通り用意し，識別用モデルをそれぞれ作成して評価を行った．なお本実験においては，特徴量としてメルケプストラムと対数 F0 (平均値正規化) を用い，識別時には入力音声の書き起こしテキストラベルを与えた．

表 2 に識別結果を示す．この結果より，本実験においても「楽しげ」,「悲嘆」については，ほとんどが意図したスタイルに識別されており,「苛立ち」,「怒り」のスタイルでも一部が混同されているものの，7 割程度は正しく識別されていることがわかる．また，表 1 の主観評価結果と比較すると,「平静」の一部が「楽しげ」に識別されて正解率が低くなっているが，その他のスタイルに関しては，概ね主観評価結果と同程度の識別結果

を得ることができた．

5. むすび

本研究では，MSD-HMM を用いた音声の感情・発話様式を伴う音声の識別について検討を行った．まず識別における基本周波数の影響を調査し，MSD-HMM を用いて特徴量に基本周波数を考慮することでメルケプストラムだけを用いた場合に比べ，識別率が改善することを示した．また，UBM に対して話者・スタイルの MSD-MLLR による同時適応を行った場合において，話者依存モデルと同等の性能が得られることを示した．

今後の課題として，様々な話者による評価，感情の表出度合の推定，および話速 (テンポ) を考慮したスタイル識別などが挙げられる．

文　献

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, 18(1), pp.32–80, Jan. 2001.

[2] R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," Speech Communication, vol.40, nos.1–2, pp.5–32, April 2003.

[3] L. Bosch, "Emotions, speech and the ASR framework," Speech Communication, vol.40, nos.1–2, pp.213–225, April 2003.

[4] T.L. Nwe, S.W. Foo, and L.C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol.41, no.4, pp.603–623, Nov. 2003.

[5] 徳田 恵一, 益子 貴史, 宮崎 昇, 小林 隆夫, "多空間上の確率分布に基づいた HMM," 信学論, vol.J83-D-II, no.7, pp.1579–1589, July 2000.

[6] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, no.2, pp.171–185, 1995.

[7] 田村 正統, 益子 貴史, 徳田 恵一, 小林 隆夫, "HMM に基づく音声合成におけるピッチ・スペクトルの話者適応," 信学論, vol.J85-D-II, no.4, pp.545–553, April 2002.

[8] 川島啓吾, 山岸順一, 小林 隆夫, "MSD-HMM を用いた音声のスタイル識別の検討," 日本音響学会 2005 年秋期研究発表会講演論文集, I, 1-P-24, pp.199–200, Sept. 2005.

[9] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol.17, pp.91–108, Aug. 1994.

[10] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Information and Systems, vol.E88-D, no.3, pp.502–509, March 2005.

付　録

1. 評価に用いた文章

平静

- 日米の首脳会談が、東京で開かれることになった。
- 環境に優しい低公害の列車を開発した。
- 自己責任は個人の責任の問題と位置づける。
- フランス産のワインを積んだ貨物便がパリから到着した。
- シンガポールのブロードバンド加入者数は横ばいだ。
- 中間発表を踏まえ、今後の取り組みに関する基本方針を決定した。
- ポルトガル料理は、格別旨くもないがまずくもない。
- 日本各地で紅葉がはじまり、見物客でにぎわっています。
- 太平洋側も朝まで雨の所がありますが、日中は各地で曇るでしょう。
- 大臣は生活保護費の国負担を現状維持とする考えを示した。

悲嘆

- もうあの人に会えないなんて悲しい。
- もうすでに夜中ですが、まだたくさん仕事が残っています。
- 今年のクリスマスイブも一人きりでケーキを食べた。
- 年賀状を出したのに、一通も返って来なかった。
- 私は顔をあげずにうなずき、ひたすら机の拭き掃除をした。
- 家に帰ってみると、みんな寝静まり夕食が用意されていなかった。
- せっかく駅まで走ったのに、ちょうど電車がいってしまった。
- 研究室には泊まらないと決めたが、結局泊まることになってしまった。
- ふと気が付くと目の前で犬が骨折していた。
- また試験を受けさせられることになった。

楽しげ

- 今年のゼミ旅行はとても楽しかった。
- でっぷりと太った男が、キレの良いダンスを踊っている。
- 微笑んだ彼女の口元には、ケチャップがついていた。
- 長い間入院していたあの子が、ようやく大学へもどってきた。
- この時間に帰れば、あのドラマが見られる。
- 頑張って何通も応募したら懸賞に当たった。
- 万引きを取り抑えて、警察から感謝状をもらった。
- 勤め始めて 3 ヵ月でバイトの時給がアップした。
- 勉強をほとんどしなかったのにテストで 100 点が取れた。
- 誕生日に欲しかったバッグをプレゼントされた。

苛立ち

- 校長先生の長話にイライラする。
- 今年の梅雨はいつあけるのだろうか。
- 何度も起こしているのに、彼はまったく起きようとしない。
- この古いカメラは、すぐにバッテリーが切れてしまう。
- 他の人は当たるのに、俺の宝くじはどうしていつもはずれるのか。
- 信号機故障で、電車に閉じ込められて、会社に遅れそうだ。
- 熱帯夜が続いて、今日もまともに寝つけない。
- 渋滞からなかなか抜け出せない。
- サポートセンターに電話したのに話し中でつながらない。
- 霜焼けが治ったと思ったら今度は花粉症だ。

怒り

- 後輩は気が利かなくて腹立たしい。
- こんなに迷惑な話があって、たまるものか。
- 赤ん坊だってそのぐらいのことは分かるのに、どうして分からないのか。
- あの人は、いつも何か一言多い。
- 実家に帰ると、いつまでたっても子供扱いされる。
- あいつは、いつになったら金を返してくれるのだろうか。
- 姪は少しも努力せず、いつも泣きごとばかり言って困る。
- いくら掃除しても、すぐ汚されてしまう。
- お釣りを誤魔化そうとしたくせに、謝りもしない。
- 説明書通りに組み立てたのに、危うく感電しそうになった。