時間構造分割特徴量に基づく感情発声の自動分類
Automatic Emotional Speech Classification using Acoustic Features of Automatic Temporal Segmentation

原　雄太郎
Yutaro Hara
法政大学大学院情報科学研究科情報科学専攻
E-mail:09t0012@stu.hosei.ac.jp

Abstract

This paper describes emotional speech classification in anime films. In emotional speech analysis, F0 and power have been widely used as acoustic features. And, temporal structure of an emotional speech has some important characteristics. In a previous study, Attack and Keep and Decay were adopted as parameters to describe temporal characteristic based on a power transition. This paper proposed an improved method of A-K-D unit estimation, and evaluated it. Moreover, in recent studies, support vector machine has been effective for emotional classification. In this study, DAGSVM adopted feature selection is used for a classifier. An emotional speech corpus was constructed by using data collected over 8 h. The corpus consists of emotional speech material of a total of 408 utterances. Four emotions, namely, joy, anger, sadness, and the neutral case, were labeled.

As a result, a recognition rate of acoustic features using A-K-D unit estimation which is proposed by this paper was 58.3%, and was higher than a previous method. Therefore, a proposed method is more effective than the previous method. And, the highest recognition rate is 80.1% when DAGSVM using Split Feature Selection.

1　まえがき

近年，感情音声認識に関する多くの研究が進められてきた．文献 [1] では，感情そのものに対する定義づけや人間の感情判断のプロセス，機械学習による感情認識の判断ロジックなどが考察されている．また，ドライバーの精神状態の情報を用いて安全性を提供するために，自動車運転環境における感情認識を扱っている研究や [2]，コールセンターシステムに感情音声認識を適用するために平静と怒りの 2 感情を区別する研究 [3] のように，感情音声認識は様々なアプリケーションへの応用を目的としている．

しかし，感情音声認識に効果的な特徴量が明確でないことや，分類が単語の長さや話者に依存することが，感情音声認識を複雑にしている [4]．さらに，感情や心というものが科学的手法によって定義しにくく，認知影響や個人差によって人の主観がゆらぎやすいことも一因である．文献 [1] では，自然感情発話における人間同士の主観一致率は 60% 程度と述べられている．また文献 [4][5] における機械学習法による認識率は 45〜60% 程度であり，文献 [6] では，意図的発話における 4 感情分類（喜び，怒り，悲しみ，平静）において全体で約 60% の認識率を得ている．このように発話の言語列を認識する音声認識と比較すると，感情音声認識は高い認識率を得にくい．

本研究では，アニメーションから抽出した声優の感情発声の認識を行う．日本のアニメーションにおける声優の演技の特徴として，キャラクター独自の語り口，高めの声，万人に分かりやすい感情表現などがある．本研究と同様にアニメーション声優の感情発話による認識を行っている文献 [7] では，アニメーション映画である “The Incredibles” [8] を音声素材とした感情音声認識を行い，分散分析によってピッチや声の大きさが分類に有効であると述べられている．

本研究の応用として，医療やコミュニケーションツールへの適用があげられる．心や感情は健康や人間同士の会話に深く影響する．感性情報に対する客観的な分析手法が整備されれば，医療分野への応用や人間同士のコミュニケーションの円滑化などに役立ち，エンターテイメントツールの開発やコミュニケーションロボットの開発などにも貢献できる．また，声優向けの感情表現の演技練習の支援も可能になり，アニメーション制作の可能性を広げることもできる．このように感情を客観的に判断できる技術が持つ意義は大きい．

本論では感情音声の時間構造に含まれる特徴量に注目する．感情音声の時間構造には，感情を判断する情報が存在することが知られている．文献 [6] では，A-K-D 区間と呼ばれるパワー変化に基づく時間構造分割手法が提案されていた．本研究では A-K-D 区間推定法を新たに提案する．また，決定木を基にしたクラス出力法である DAGSVM[10] の各 SVM モジュールに特徴選択 [9] を行う分類器を提案する．

2　感情発声の音響特徴分析

2.1　音声素材

文献 [6] では，自然発話と意図的発話を対象とした感情音声認識を行っている．自然発話は，発話の瞬間の発話者の感情を主観的に特定することが難しく，精神状態や環境にも左右されやすい．逆に演技感情発声のような意図的発話は，声の大きさや声色による演出が目立つが，観測者にとって感情が主観的に判断しやすい．さらに，声優のように声の演技に習熟していれば，抑揚だけでも主観評価が一致しやすい感情表現ができる．また，日本人は感情表現が乏しい傾向があるとも述べられているが [1]，声優の演技感情発声は環境や精神状態に対するゆらぎに強く，観測者による感情音声サンプルの主観分類が容易である．以上の観点から，本研究では声優の演技感情発声を用いる．

本研究では，日本アニメーションの「ハチミツとクローバー」から発話を抽出した．「ハチミツとクローバー」は，1 話につき 20 分程度の 24 話で構成され，日常の学園生活を題材にしている．男性と女性で主要なキャラクターが多く登場しており，様々な感情の演技発声を収集できる．このアニメーション作品の 24 話分 (約 8 時間分) のシーンから，声優 6 人（男 4，女 2）について，サンプリング周波数 48kHz，量子化ビット数 16 ビットでサンプルを抽出した．なお特徴量を求める際には，サンプリング周波数は 8kHz にダウンサンプリングする．本論で扱う感情は，感情の違いがわかりやすく，多くの関連研究 [2][3][6][7] で用いられている，“喜び”，“怒り”，“悲しみ”，“平静” の 4 感情とする．本研究で用いるサンプルは複数の発話者の発話の重なりを持たない．また 2 人の判定者で主観分類を行い，感情ラベルが一致したサンプルのみを使用している．各感情で 102 個，のべ 408 個のサンプルを用いる．

Supervisor: Prof. Katunobu Itou

2.2 感情発声分類に用いる音響特徴量

本研究では，多くの関連研究 [2][5][6][7][11] で用いられている F0 及びパワーに関する特徴量 (最大値, 最小値, 平均, 分散など) や 12 次の MFCC，304 次の TEO-CB-Auto-Env（Teager Energy operator）などの音響特徴量を用いる．また時間構造に分割した各区間の F0，パワーも用い，のべ 378 次元の音響特徴量を扱う．

F0 は声の高さを表す特徴量であり，STRAIGHT[12] を用いて求める．パワー (デシベル) は窓幅 70ms のパワースペクトルの総和の平均と無声部分 (バックグラウンドノイズ) の比で求める．MFCC は人間の声道の音響特性 (口腔の形) を表す特徴量である [4]．TEO-CB-AutoEnv は，発話の長さや話者に依存しない特徴量であり周波数帯域におけるサブバンドごとのピッチの変動をみる．この特徴量は，F0 推定の精度に影響を受けず，F0 に依存しない点で安定した特徴量である．人間の臨界帯域に基づくフィルタバンクを構成したのち，サブバンドごとのピッチの変動具合を感情ごとの違いとして認識することができる．

また話者による個人差を考慮して，得られた特徴量の正規化を行う．本研究では，平静の感情から他の感情の特徴量の距離を考慮する．そのため平静の感情の特徴量の平均により，話者ごとに正規化を行う．

2.3 Teager Energy Operator

感情音声認識では，従来から F0 やパワーといった特徴量が使用されてきた．しかし，これらは発話の長さに依存してしまう．短い発話（3-4 語程度）に限定した音声感情認識も行われているが，これは短い発話を想定している場合には有効であるが，発話の長さが不定長である場合には応用しにくい．

感情は外部からの負荷を受けてあらわれると考えられ，ストレスの一種と言える．直井ら [4] は，言語に依存しない特徴量として，周波数帯域に注目した．周波数帯域での特徴量は，ストレスを検出するのに有効である．TEO は非線形演算子でありピーク部分をより強調することができる．この強調部分の感情ごとの違いを音声感情認識に応用する．TEO は連続信号 $x(t)$ で表される．

$$\begin{aligned}\Psi_C &= (\frac{d}{dt}x(t))^2 - x(t)(\frac{d^2}{dt^2}x(t)) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t)\end{aligned} \tag{1}$$

これは離散信号 $x(n)$ では次のようになる．

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{2}$$

ここで，$\Psi[\cdot]$ が TEO であり，非線形である．この演算子によって得られる瞬時的エネルギーの変化がストレス検出及び感情発声認識に有効である．本研究では，TEO を応用した TEO-CB-AutoEnv[4] を音響特徴量として用いる．この特徴量は F0 推定の精度に影響を受けず，F0 に依存しない点で安定した特徴量である．図 1 に TEO-CB-AutoEnv の計算プロセスを示す．人間の聴覚システムは，可聴な周波数の範囲を多くの

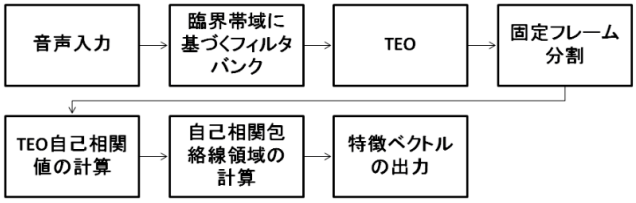

図 1. TEO-CB-AutoEnv 特徴量抽出手順

臨界帯域に分割するフィルタリング操作を行っていると考えられている．そこで TEO-CB-AutoEnv は，人間の臨界帯域に基づいたフィルタバンクを使用する [4]．

さらに，サブバンドごとに TEO を計算し固定フレームで分割を行う．フレームは，シフト幅がフレーム長の半分になるように分割を行う．文献 [4] では 19 フレームに音声データを分割していた．その後，各サブバンドで自己相関関数の計算を行う．もしフレーム内のピッチに変動がなければ，出力される TEO は一定であり，自己相関関数は時間領域で (0,1) から (N,0) への減衰直線となる．ここで N はフレーム長である．そのため，フレーム内での自己相関包絡線領域の面積は N/2 となる．しかしフレーム内にピッチの変動がある場合，正規化された自己相関包絡線は理想的な直線にならないため，自己相関包絡線領域の面積は N/2 より小さくなる．以上の計算を行うことで，各サブバンドの励起変動の具合を反映できる．このようにして，フレーム長 N で正規化された自己相関包絡線領域パラメータを，特徴量として扱う．結果として TEO-CB-AutoEnv では，固定分割フレーム数×バンド数の特徴ベクトルが得られる．

本研究では，固定フレーム数 18，サブバンド数 16 の 288 次元の特徴ベクトルを用いる．また本研究で用いるサンプルは不定長のため，固定フレーム分割時に感情分類に有効な特徴が得られないことも考えられる．そのため，フレーム分割を行わない特徴ベクトル 16 次元（サブバンドの個数）を追加し，本研究では TEO-CB-AutoEnv に関する，のべ 304 次元の特徴ベクトルを用いる．

2.4 感情発声の時間構造

従来研究では，発話を時間構造に分割するために，パワーや F0 の変化が用いられている．文献 [11] では，F0 やパワーの特徴量の概形が右上がり，右下がりの場合で時間構造を分割し，各区間で平均，レンジ，最大値などの特徴量を求めている．

また光吉 [6] は，パワー変化に基づく発話の時間構造モデルを提案した．このモデルは，発話を 3 つの区間に分割する．“Attack” と呼ばれる区間は，パワー領域において発話の開始からピークまで継続する区間である．“Keep” と呼ばれる区間は，パワーレベルが一定に保たれている区間である．最後に “Decay” と呼ばれる区間は，パワーレベルが下降し続けている区間である．このように，光吉は Attack-Keep-Decay を単位として音声を区分し，これを A-K-D 区間と呼んだ（図 3）．

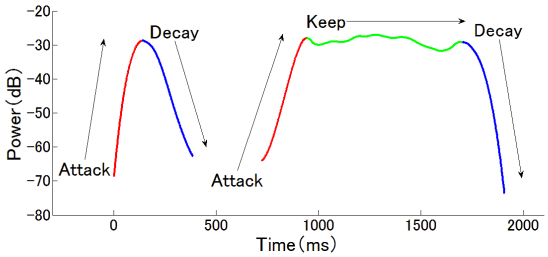

図 3. A-K-D 区間に構造分割された発話信号

感情発声では，特にパワーの立ち上がり（Attack）と立下り（Decay）に感情認識の要因があるとされている．Attack 区間では，“喜び” や “怒り” の感情発声の F0 の最大値や平均などの特徴量が，他の感情と比較して非常に大きい値を取る．Keep 区間では，“悲しみ” の感情発声において持続時間が短い傾向にあり，パワーとピッチがあまり変化しない．Decay 区間では，“悲しみ” の感情発声において，F0 やパワーの傾きが小さい傾向にある．このように A-K-D 区間に基づく構造モデルは，感情発声分類に有効であると考えられる．

文献 [6] では，特徴量として Attack と Decay 区間では傾き，最大値，区間の長さが用いられていた．また Keep 区間では，区間の長さ，パワーの平均，Δ パワーの平均と分散を用いており，これらの特徴量を用いて自然発話及び意図的発話の感情音声の分類が行われている．本研究では，A-K-D 区間の推定アルゴリズムを新たに提案する．なお A-K-D 区間は，パワー変化を基に推定するされるため，A-K-D の各区間中で推定する特徴量は，パワーに関連した特徴のみであり，F0 や TEO-CB-Auto-Env などの周波数領域に関する特徴は含まない．

2.4.1 従来手法による A-K-D 区間推定

Attack 開始点は，以下に示す式のパワー差分が閾値を超えた点とする．

$$\Delta p = p_n - p_{n-1} (n = 1, 2, 3, 4, ...) \tag{3}$$

開始点検出後，$\Delta p > 0$ が続いている間を Attack 区間とする．$\Delta p \leq 0$ の検出があった時，Attack 区間を終了する．Attack 区

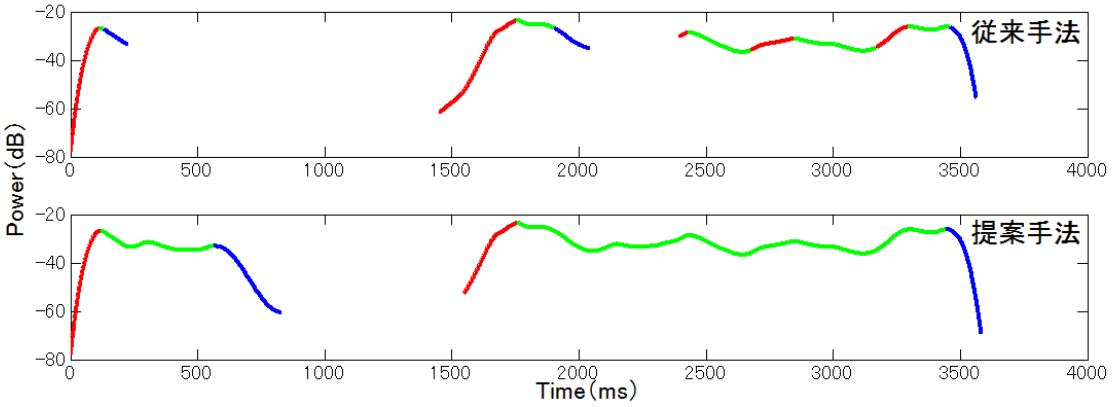

図 2. 従来アルゴリズムと提案アルゴリズムの A-K-D 区間推定結果 (赤が Attack, 緑が Keep，青が Decay)

間の終了点から Keep 区間が開始される．Keep 区間の連続条件は，フレーム間において Δp の絶対値が閾値を超えないこととする．なお文献 [6] では閾値についての詳細な記載がなかったため，本論において従来手法における A-K-D 区間推定の際に用いる閾値には，1 発話全体の Δ パワーの平均値を用いている．閾値を超えた場合に Keep 区間を終了し，Decay 区間が開始される．Decay 区間は $\Delta p \leq 0$ と定義される．$\Delta p \geq 0$ の検出があった場合，Decay 区間を終了し，一つの A-K-D 区間が終了し，次の A-K-D 区間推定が開始される．以上のアルゴリズムが，A-K-D 区間検出の従来手法である．

しかし従来手法では，パワー全体の流れが単一のピークを持つ場合（図 3 の左の概形），Decay として認識されるべき区間が Keep として誤認識されてしまう可能性がある．また，パワーの流れに単一のピークを持たない場合（図 3 の右の概形）では，Keep として認識されるべき区間が複雑な波形をしていることで，Attack 区間や Decay 区間として誤認識されてしまう可能性もある．本論では，これらの課題の改善を目指す．

2.4.2 提案手法による A-K-D 区間推定

A-K-D 区間検出の事前処理として，始めに有声区間推定を行う．有声区間推定には，STRAIGHT [12] の V/UV 判定を用いる．しかし STRAIGHT の有声区間推定のみでは，発話の立ち上がりと立下りが，非音声として誤認識する可能性がある．そのため，STRAIGHT で有声区間が検出された時，前後の 100msec 以内のフレームを 1 つの有声区間として扱う．なおここでは日本語の音節の長さを 100msec として考えている．求められた有声区間中の信号から，A-K-D 区間推定及び音響特徴量の算出を行う．有声区間推定後，有声区間中のパワー変化に基づき A-K-D 区間検出を行う．また次数 3，フレーム幅 201ms の最小二乗平均（Savizky-Golay）フィルタを用いてパワーの平滑化を行う．

有声区間の開始点を Attack の開始点とする．10ms 毎のパワーの傾きを後ろのフレームと比較することで Attack の終了点を求める．次式を満たす点を Attack の終了点とする．ここで R は 10ms 毎のパワーの回帰係数である．

$$|R^n| > 0, R^{n+1} < 0 \tag{4}$$

Attack の終了点を Keep の開始点とする．次にパワー変化の概形判定を行う．本論では，1 つの有声区間のパワー変化において，単一のピークを持ち Keep 区間が存在しないと場合と，単一のピークを持たず Keep 区間が存在する場合に分類できると仮定する．1 つの有声区間の Attack 開始点から Attack 終了点までのパワーの傾き（R_{Attack}）と，Keep 開始点から有声区間の終了地点までの傾き（R_{KD}）を比較することで，それぞれのケースに分類する．

$|R_{KD}| < R_{Attack}$ ならば，Keep 開始点からのパワー変化の傾きが大きいので図 2 の左のパワー変化の形ように単一のピークが存在すると判定する．つまり有声区間中に Keep 区間が存在しないと判定し，Attack 終了点が Decay 開始点となる．

$|R_{KD}| \geq R_{Attack}$ ならば，Keep 開始点からのパワー変化の傾きが小さいので図 2 の右のパワー変化の形のように単一のピークが存在せず，Keep 区間が存在すると判定する．このように判定された場合，式（5）の R_{peak} の条件を満たす点の中で，式（6）を満たす点 n を Keep 終了点とする．

$$R_{peak} = \left\{R^n > 0, R^{n+1} < 0, \left|R^{n+1}\right| > |R^n|\right\} \tag{5}$$

$$R^n \geq R_{peak} \tag{6}$$

そして Keep 終了点が Decay 開始点となり，有声区間の終了点を Decay 終了点とする．以上のアルゴリズムにより 1 つの A-K-D 区間を推定し，次の A-K-D 区間推定を行う．図 2 に，従来手法と提案手法による A-K-D 区間推定の例を示す．図 2 から，提案手法では 2.4.1 であげた課題の改善が達成できたと考える．これらの推定手法の評価は，本論の 4.3 章で行う．

2.4.3 F0 の変化を用いた時間構造分割

本研究では，F0 の変化を基にした時間構造からも特徴量を得る．F0 とパワーは同じ時間軸上でも異なる変化をみせるため，パワー変化を基に推定される A-K-D の各区間で，F0 に関する特徴を抽出するのは適切ではない．文献 [11] では，F0 軸上で F0 の上昇区間と下降区間の 2 つの時間構造に分割し特徴量を求めている．本研究では，上昇区間を Attack 区間，下降区間を Decay 区間と定義する．A-K-D 区間推定と同様に，STRAIGHT で求めた有声区間中の信号から，Attack 区間と Decay 区間を求める．

$$Rise - fall : F_{0(t_{k-1})} < F_{0(t_k)} \quad \& \quad F_{0(t_k)} > F_{0(t_{k+1})} \tag{7}$$

$$Fall - rise : F_{0(t_{k-1})} > F_{0(t_k)} \quad \& \quad F_{0(t_k)} < F_{0(t_{k+1})} \tag{8}$$

式 (7) (8) における t は時変数である．式 (7) は，F0 が Attack から Decay に変化する点を示しており，発話中の F0 の Decay 開始点及び Attack 終了点を示す．式（8）は，F0 が Decay から Attack に変化する点を示しており，発話中の F0 の Attack 開始点及び Decay 終了点を示す．式（7）で観測された点から式（8）で観測された点までを，Decay 区間，式（8）で観測された点から式（7）で観測された点までを，F0 軸における Attack

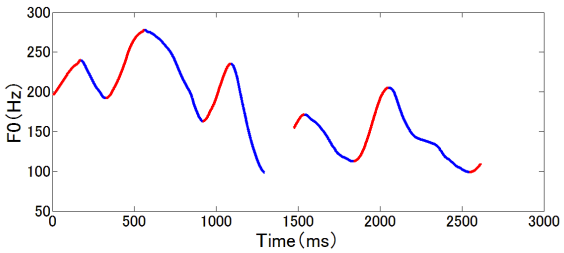

図 4. F0 軸における時間構造分割された発話信号（赤が Attack 区間，青が Decay 区間）

区間と定義する．このように単純ではあるが，以上のアルゴリズムにより F0 軸における Attack 区間及び Decay 区間を推定する．推定された Attack 区間及び Decay 区間を図 4 に示す．

3 多クラス SVM による分類器の構築

従来研究 [2][3][5][13] では, K-nearest neighbour method（K-NN），Probabilistic Neural Network（PNN），Support Vector Machine（SVM），Gaussian Mixture Model（GMM）などの分類器が用いられてきた．本論では従来研究 [2][7][14] などで，最も有効な認識結果が得られている SVM による認識を行う．SVM では，特徴次元が増えても分類器が有効に働くことが知られている．人間は様々な音響特徴から発話の感情を推定しており，次元数の増加に強い SVM は感情発声に適した分類器だと考える．

文献 [15] では，8 種類の感情ごとに認識に有効な特徴量を絞り込んでおり，各感情における有効特徴量が異なることが報告されている．そのため，本論では特徴選択 [16] を行い認識率の向上を図る．また文献 [17] では，独立特徴選択という多クラス SVM における特徴選択の手法が提案されている．本研究では，独立特徴選択を DAGSVM に適応させた分類器を提案する．なお本論では，LIBSVM[18] を用いて SVM の実装を行う．

3.1 SVM の多クラス化

二クラス識別器である SVM を多クラス問題に運用する方法としては，あるクラスと他のクラス間で識別を行う OneVsAll(一対多分類器)，すべてのクラス対で識別器を構成し，それらを組み合わせて識別を行う OneVsOne(一対一分類器) などがある [17].

OneVsAll は N クラスのデータを学習する場合，すべてのデータを用いた N 個の SVM を構築する必要があり，そのすべての結果を統合する．この方法はクラス数とサンプル数が多いときは，非常に計算量が多い．

一方 OneVsOne の場合は, N クラスのデータを学習する際には，学習に 2 クラス分のデータを用いた N(N-1)/2 個の SVM が必要となる．そのため，OneVsOne は構築しなければならない SVM の個数は OneVsAll よりも多くなるが，各 SVM を構築する際に用いるデータ量は OneVsAll よりも少ないため，OneVsAll よりも高速な学習が可能であることが知られている．それぞれの手法を比較すると，OneVsOne は各 SVM モジュールが二クラス問題を担当するため，構造がシンプルであり計算量の少なさや次元数を多く削除できる．

3.2 Max-Win アルゴリズム

多クラス SVM における OneVsOne アルゴリズムの代表的な統合手法として，Max-Win アルゴリズム [19] がある．Max-Win アルゴリズムは多数投票制を用いており，テストサンプル x に対する 2 クラス（l,m）の SVM に関する判別関数 $D_{lm}(x)$ が正であればクラス l に投票され，そうでない場合はクラス m に投票される．SVM の出力は各クラスの投票数によって決定され，最大票を得たクラスがテストパターンの出力クラスとなる．また，最大票を得たクラスが複数存在する場合，TMDF（Total magnitude of discriminant function）によって出力クラスが決定される．クラス l に対する TMDF の式を以下に示す.

$$TMDF_l = \sum_m |D_{lm}(x)| \tag{9}$$

ここで l は観測クラス，m は l に対する全てのクラスである．TMDF が最大値のクラスが出力される．

3.3 特徴選択 (Feature Subset Selction)

パターン認識で用いるデータベースは，一般的に事例とそれを表す特徴量で表され，高次元ベクトルで構成されることが多い．そのため特徴次元やクラスが多い場合は，計算コストが高くなりやすいという問題がある．さらに，識別に関してあまり意味を持たない不要な特徴により，最良の識別を行うことができないこともある．しかし，このような場合でも特徴選択を行うことで，識別精度を向上させることができる [9].

特徴選択は，与えられた特徴集合から識別器にとって最も効果のある特徴集合の候補を与える．元の D 個の特徴を持つ特徴集合 S から識別に有効な情報をもつ d 個の特徴から構成される部分集合 s を探す処理を行う．評価関数の出力値が高いほど，良い特徴集合である．

特徴選択には様々な方法があり，一方向の探索アルゴリズムなどは容易に用いられるが，可能な全ての特徴の組み合わせについて実行できるわけではないので，最良の特徴集合を求められるとは限らない．そこで Pubdil らにより SFFS(Sequential Floating Forward Search)[16] が提案された．SFFS は，Forward 型アルゴリズムと Backward 型アルゴリズムを組み合わせた Floating 型アルゴリズムである．SFFS は Forward 型を基本としているが，特徴を増やした後，それまでに選択された特徴集合の中から 1 つを削除し，評価関数が良くなった場合は削除し続け，評価関数が悪くなった場合は削除を中止し特徴量を増やす作業に戻る．SFFS は SFS や SBS などの一方向探索アルゴリズムと比較して，より高いパフォーマンスが得られるが，特徴数やクラスが大きいと計算量が膨大になる．本研究では，認識性能の高い特徴集合を求めるために SFFS アルゴリズムを用いる．

3.4 独立特徴選択（Split feature selection）

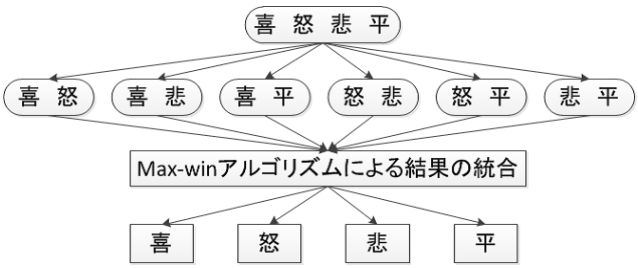

図 5. OneVsOne-SVM による感情クラス決定経路

パターン認識では，全体の入力空間で特徴選択を行い，その結果を識別機に入力する手法が広く用いられており，グローバル特徴選択（Global Feature Selection）と呼ばれている．それに対し文献 [17] では，多クラス SVM の OneVsOne のアルゴリズムを用いてクラス対ごとに識別器を構成し，各 SVM モジュールにおいて特徴選択を行う手法が提案されている．このような特徴選択手法を独立特徴選択（Split Feature Selection）と呼んでいる．

独立特徴選択は，各部分空間で最適化された特徴選択が行うため，識別時間を抑えつつ高精度な識別が可能である．OneVsOne-SVM の各 SVM モジュールで，SFFS アルゴリズムを用いて特徴選択を行い，結果を統合してクラスが決定される．OneVsOne-SVM のクラスの決定経路を図 5 に示す.

3.5 DAGSVM（Directed Acyclic Graph Support Vector Machines）

3.2 章の Max-Win アルゴリズムの弱点として識別時間の長さがあげられる．この問題を解決するために提案された手法が DDAG（Decision directed acyclic graph）アルゴリズムを用いた DAGSVM（Directed Acyclic Graph Support Vector Machines）[10] である．DAGSVM は高速な識別時間を実現しており，良好な結果を得ることができる．文献 [20] では，OneVsOne-SVM と DAGSVM を含めた 5 つの多クラス SVM の評価を行っており，3 クラス及び 4 クラス分類問題における OneVsOne-SVM と DAGSVM が最も性能が高かった．

DAGSVM は学習時に，OneVsOne-SVM を $M(M-1)/2$ 個構築する．また評価時には，根の OneVsOne-SVM は $M(M-1)/2$ 個の内部ノードと M 個の葉ノードをもつ．各内部ノードはペアワイズの SVM に対応している．テストサンプル x のクラスを決定するために，各ノードにおける SVM の判別関数をもとに木構造を進んでいく．到達した葉ノードのクラスがテストサンプルの出力に対応する．DAGSVM は M-1 回の比較を行うため，識別時間の観点から OneVsOne-SVM よりも効果的なアルゴリズムである．本研究における DAGSVM を用いた感情の決定経路を図 6 に示す．本研究では，DAGSVM に独立

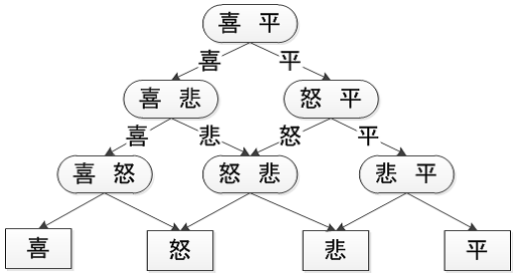

図 6. DAGSVM による感情クラス決定経路

特徴選択を適用した分類器を提案する．OneVsOne-SVM の場合，各 2 クラス分類器では特徴選択によって高い性能を得られるが，認識誤りを起こした SVM によって，各 2 クラス分類器の統合後に正しく認識できないことが考えられる．DAGSVM は，Max-Win アルゴリズムのように複数の SVM モジュールの識別関数に依存せず，単純な 2 クラス分類器の性能によって分類が求まる．そのため，独立特徴選択により DAGSVM の各 SVM モジュールの性能を高めた分類器は，高い性能が期待される．

4 実験

4.1 特徴選択による各 SVM モジュールの特徴推定

各感情クラス対及び 4 感情の全体空間に対する有効な特徴セットを求めるために，グローバル特徴選択と独立特徴選択を行った．特徴選択のために，2.2 章で提示した 378 次元の音響特徴量を用いた．A-K-D 区間から得られる特徴量については，ここでは提案手法で推定された区間における特徴量を用いた．

4.1.1 実験結果

4 感情の全体空間，喜びと怒り，喜びと悲しみ，喜びと平静，怒りと悲しみ，怒りと平静，悲しみと平静の計 7 つの SVM モジュール対して特徴選択を行った．4 感情の全体空間の識別誤り確率は 23.5% であり，A-K-D 区間特徴が 6 次元，TEO 特徴が 17 次元，パワーのレンジ，Δ パワーの最小値，1 次元目の MFCC の計 26 次元の特徴セットが得られた．喜びと怒りの識別誤り確率は 17.2% であり，A-K-D 区間特徴が 6 次元，TEO 特徴が 4 次元，パワーの標準偏差，Δ パワーのメディアン，ΔF0 のメディアンの計 13 次元の特徴セットが得られた．喜びと悲しみの識別誤り確率は 8.3% であり，Attack 区間特徴が 2 次元，TEO 特徴が 4 次元の計 6 次元の特徴セットが得られた．喜びと平静の識別誤り確率は 7.4% であり，A-K-D 区間特徴が 4 次元，TEO 特徴が 5 次元，1 次元目の MFCC，F0 のメディアン，Δ パワーの最小値の計 12 次元の特徴セットが得られた．怒りと悲しみの識別誤り確率は 2.9% であり，ΔF0 のメディアン，TEO 特徴が 29 次元の計 30 次元の特徴セットが得られた．怒りと平静の識別誤り確率は 7.8% であり，A-K-D 区間特徴が 6 次元，パワーの標準偏差，レンジ，メディアン，1 次元目の MFCC，Δ パワーの最小値，TEO 特徴が 13 次元の計 24 次元の特徴セットが得られた．悲しみと平静の識別誤り確率は 10.3% であり，ΔF0 のメディアン，TEO 特徴が 9 次元の計 10 次元の特徴セットが得られた．

4.1.2 考察

怒りと悲しみ，悲しみと平静，喜びと悲しみの SVM は，F0 に関する特徴と TEO に関する特徴セットが選択されていた．怒りと悲しみの SVM は，識別誤り確率が最も低いクラス対であり，特徴セットの大半が TEO 特徴であった．そのため TEO 特徴が怒りと悲しみの分類に有効であると考えられる．悲しみと平静のペアについても TEO 特徴は非常に重要な特徴量といえる．また発話の立ち上がりの F0 が，喜びと悲しみの分類に効果的であると考えられる．これら 3 つのクラス対では特徴セットでは TEO 特徴が多く選択されていた．そのため，悲しみと他の感情を分類する際には，TEO 特徴が有効だといえる．各サブバンドの励起変動に悲しみ特有の特徴がある可能性が考えられる．

怒りと平静，喜びと平静のクラス対は，パワーと TEO 特徴で感情を分類している傾向がある．そのため発話者は，平静と怒り，喜びを区別するために，発声の強さを意識的に大きくしており，また瞬間的なエネルギーの変化（励起変動）により各感情の発声を応変していると考えられる．

喜びと怒りの SVM は，他の感情と比較して特徴セットに傾向がみられなかった．また識別誤り確率も最も高い値であり，効果的な特徴が元の特徴集合に含まれていなかった可能性が考えられる．また怒りは喜びと比較して F0 やパワーが強いといわれているが [21]，本研究のサンプルにおいて主観で聴取したが，大きな差はなかった．

4.2 特徴選択後の感情発声分類

選択された特徴セットを用いて認識実験を行う．OneVsOne-SVM と DAGSVM について，グローバル特徴選択（GFS）及び独立特徴選択（SFS）を行ったモデルと, 特徴選択を行わなかった（None）のモデルの認識率を算出した．なお leave-one-out 交差検定を用いて性能を算出している．

4.2.1 実験結果

表 1. 特徴選択後の特徴セットを用いた各モデルの認識結果

分類器 ＼ 特徴選択手法	None	GFS	SFS
OneVsOne-SVM	44.9%	76.2%	70.6%
DAGSVM	44.9%	76.2%	**80.1**%

表 2. 特徴選択を行ったモデルの各感情の認識結果

分類器 ＼ 感情	喜び	怒り	悲しみ	平静
GFS-SVM	69.6%	71.6%	76.5%	87.3%
SFS-OneVsOne	71.6%	62.7%	68.6%	79.4%
SFS-DAG	70.6%	78.4%	81.4%	90.2%

表 3. 独立特徴選択を用いた DAGSVM の認識結果

出力 ＼ 入力	喜び	怒り	悲しみ	平静
喜び	**70.6**%	12.8%	5.88%	0.98%
怒り	15.7%	**78.4**%	2.94%	1.96 %
悲しみ	5.88%	2.94%	**81.4**%	6.86 %
平静	7.84%	5.88%	9.80%	**90.2**%

表 1 に各モデルの認識結果を示す．SFS を用いた DAGSVM が最も性能が高かった．GFS を行った結果，特徴選択を行わないモデルよりも 69.7% 性能が向上した．また DAGSVM は，特徴選択の有無にかかわらず，OneVsOne-SVM と同程度かそれ以上の認識率であった．表 2 に特徴選択を行ったモデルの各感情の認識率を示す．なお GFS を行った OneVsOne と DAGSVM は同じ認識結果のため，表 2 では 2 つのモデルを GFS-SVM と表記している．表 3 に最も性能が高かった提案分類器の認識結果を示す．平静が最も高い認識率であり，喜びが最も低かった．また，喜びと怒りが互いに誤認識しやすかった．

4.2.2 考察

提案分類器の性能が最も高かった．従来研究 [17] で提案された SFS を行った OneVsOne-SVM と比較して 13.5% 性能が向上し，GFS を行ったモデルと比較して 5.1% 性能が向上した．特徴選択を行った場合，DAGSVM は OneVsOne-SVM と同等かそれ以上の性能であった（表 1）．

SFS を行った場合，OneVsOne-SVM では，ある SVM モジュールで正しく認識されても，他の SVM モジュールでの識別誤りや識別関数の出力によって統合後の識別誤りを起こす可能性があり，SFS の直接的な恩恵を受けにくい．一方，DAGSVM は 2 クラス分類問題に帰結し，各分類器の性能に認識率が左右される．そのためクラス対の SVM の性能が向上すれば，全体の性能も同時に向上するモデルであるといえる．以上の理由から，提案分類器が最も高い性能であったと考えられる．また全ての感情において高い性能を得られており（表 2），感情発声分類に有効なモデルといえる．

OneVsOne において，GFS を行ったモデルと比較して SFS を行ったモデルは性能が 7.35% 低かった．OneVsOne-SVM

における SFS では，統合後の識別誤りが起こる可能性があることが原因として考えられる．また OneVsOne-SVM における GFS では，全体の特徴集合で特徴選択を行っており，あるSVM モジュールで識別誤りを起こしても，他の SVM モジュールとの識別関数の統合によって正しいクラスが選ばれる特徴集合が得られるため性能が高かったと考えられる．

表 3 から，喜びと怒りは互いに誤認識されやすかった．これは最も誤り確率が高いクラス対である喜びと怒りの SVM の認識誤りが多いためだと考えられる．今後は全体のモデルの性能向上のために喜びと怒りの SVM の性能向上が必要である．喜びと怒りは言語情報で分類できる可能性がある．そのため，今後は言語情報を付加したモデルにより，高精度な分類ができると考える．

4.3 A-K-D 区間推定の性能評価

従来手法と提案手法で推定した A-K-D 区間の特徴量を用いた分類器における認識率の比較で性能の評価を行った．本研究で用いた A-K-D 区間特徴は 24 次元である．分類器は 4.2 章で最も高い性能を得られた提案分類器を用いた．

表 4. A-K-D 区間特徴を用いた認識結果

	喜び	怒り	悲しみ	平静	全体
従来手法	19.6%	52.9%	52.0%	55.9%	45.1%
提案手法	**35.3%**	**72.6%**	**61.8%**	**63.7%**	**58.3%**

提案手法により全体の性能は 29.3% 向上した．また全ての感情においても性能が高かった（表 4）．これは，提案手法が時間構造を正確に分割できており，各区間で特徴量を効果的に抽出できているためだと考えられる．そのため提案した A-K-D 区間推定は，従来手法よりも感情分類に有効な時間構造分割を可能にしている．これは発話による Keep 区間の有無の推定が有効であることを示している．

しかし，喜びの認識率は提案手法においても 35.3% と低い認識率であった．主観評価において，喜びと比較して他の感情の時間構造分割の精度が極端に増減することがなかったため，時間構造分割が喜びの分類に有効でない，もしくは時間構造中の音響特徴量（F0，パワーなど）が有効でない可能性がある．

また，本研究では，A-K-D 区間を求める際に有声区間を検出しているため，誤推定が存在すると時間構造分割に影響を及ぼす．より精度の高い時間構造分割を行うために，ロバストな有声区間推定を行う必要がある．

5 あとがき

本研究では，声優の感情発声分類を行った．喜び，怒り，悲しみ，平静の 4 感情について，378 次元の音響特徴量を用い，感情発声分類に有効といわれているサポートベクターマシンによる分類を行った．また本論では，従来研究で感情発声の分類に効果的な時間構造モデルである A-K-D 区間推定の手法と，独立特徴選択を適用した DAGSVM を提案した．

提案手法による時間構造分割特徴量は，従来手法と比較して全体の性能が 29.3% 向上した．全ての感情においても提案手法による時間構造分割特徴量の方が性能が高かった．そのため，提案手法は精度の高い時間構造分割が可能であり，効果的に時間構造特徴量を推定できると考えられる．

また本研究で提案した独立特徴選択を適用した DAGSVM は，最も性能が高く 80.1% の認識率が得られた．従来手法である独立特徴選択を適用した OneVsOne-SVM と比較して性能が 13.5% 向上し，グローバル特徴選択を行ったモデルと比較して性能が 5.1% 向上した．DAGSVM は全体の認識率が各分類器の性能に依存すると考えられる．そのため提案分類器は，各クラス対の性能が向上すれば全体の性能も同時に向上するモデルであるといえる．以上の理由から，提案分類器が最も高い性能であったと考えられる．しかし本研究で用いたコーパスのみでは，分類器の評価が十分ではないと考えられる．コーパスによって認識性能が異なる可能性もあり，今後は複数の感情発声コーパスを用いて評価する必要がある．

本研究では，文献 [6] のような意図的発話における 4 感情分類の認識率（60%）よりも，高い性能を示すことができた．これは多くの音響特徴量及び提案分類器による効果が大きい．また多くの従来研究では，サンプル収集の際に一定の言語列をもとに発話が収集されるが，本研究ではアニメから感情発声を抽出しており，サンプルは不定長であり言語列も一定ではない．そのため，提案モデルは実環境での自然な会話で用いられる感情発話においても，高い性能を持つモデルといえる．

喜びと怒りは互いに誤認識が多く，それらの感情を分類するために必要な特徴量が元の特徴集合に含まれていない可能性がある．今後は喜びと怒りを分類できる特徴量について考慮する必要がある．また，言語情報でこれらの感情を分類できる可能性があり，今後は音響特徴量に言語情報を付加したモデルにより認識率の向上を目指す．

参考文献

[1] Shunji Mitsuoka et al.”Emotion Recognition”, IEEJ,125,3,pp641-644,2005.

[2] B.Schuller et al.”Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture”, ICASSP ’04, vol.1, pp.I-577-80, 2004.

[3] W.J.Yoon et al.”A Study of Emotion Recognition and Its Applications”, Modeling Decisions for Artificial Intelligence, vol.4617/2007, pp.455-462, 2007.

[4] 直井 克也他,”Teager Energy Operator を使用した音声感情認識”, IEICE technical report. Speech 105(572), pp.1-6, 2006.

[5] D.Ververidis,”Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm”, ICME, pp.1500-1503, 2005.

[6] S.Mitsuyoshi et al.”NON-VERBAL VOIVE EMOTION ANALYSIS SYSTEM”, IJICIC, vol2, pp.819-830, 2006.

[7] N.Amir,”Characterizing emotion in the soundtrack of an animated film: Credible or incredible?”, Affective Computing and Intelligent Interaction, vol.4738/2007, pp.148-158, 2007.

[8] B.Bird (Director),”The Incredibles [motion picture]”, United States: Walt Disney Pictures, 2004.

[9] I.Guyon,”Gene Selection for Cancer Classification using Support Vector Machines”, Machine Learning, vol.46, pp.389-422, 2002.

[10] John C. Platt et al.”Large Margin DAGs for Multiclass Classification”, MIT Press, pp.547-553, 2000.

[11] C.F.Huang et al.”A three-layered model for expressive speech perception”, Speech Communication, vol50(10), pp.810-828, 2008.

[12] K.Hideki et al.”FIXED POINT ANALYSIS OF FREQUENCY TO INSTANTANEOUS FREQUENCY MAPPING FOR ACCURATE ESTMATION OF F0 AND PERIODICITY ”, Proc.EUROSPEECH’99, vol.6, pp.2781-2784, 1999.

[13] W.Ser et al.”A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition ”, ICPR, pp.1-4, 2008.

[14] Olusola Olumide Aina et al.”Extracting Emotion from Speech: Towards Emotional Speech-Driven Facail Animations”,Smart Graphics,Volume 2733/2003,pp.65-80,2003.

[15] 有本泰子他,”感情音声のコーパス構築と音響的特徴の分析 : MMORPG における音声チャットを利用した対話中に表れた感情の識別”,IPSJ SIG Notes 2008(12), pp.133-138, 2008.

[16] P.Pudil et al.”Floating search methods in feature selection”, Pattern Recognition Letters, vol.15, pp.1119-1125, 1994.

[17] 胡 欣他,”多クラスサポートベクターマシーンにおける各 SVM モジュールの独立特徴選択”, 電子情報通信学会技術研究報告. NC, ニューロコンピューティング 105(457),pp.31-36,2005.

[18] C.W. Hsu et al.”A Practical Guide to Support Vector Classification ”, 2010.

[19] J.H.Friedman,”Another approach to polychotomous classification”, Technical report, Stanford, Department Statistics, 1996.

[20] Naotoshi Seo,”A Comparison of Multi-class Support Vector Machine Methods for Face Recognition”, Department of Electrical and Computer Engineering, Maryland, 2007.

[21] Dimitrios Ververidis et al.”Emotional speech recognition: Resources, features, methods, and applications,” Speech Communication,vol.48,issue 9,pp.1162-1181,2006.