情報処理学会論文誌 Vol. 50 No. 3 1181–1191 (Mar. 2009)

韻律の部分空間を用いた感情音声合成

森山 剛[†1,*1] 森 真也[†1,*2] 小沢 慎治[†1,*3]

様々な感情を含んだ音声の韻律パターンを主成分分析して得られる部分空間を用いて，所望の感情を含んだ任意単語音声を合成する手法を提案する．本手法は，感情による韻律パターンの変動を部分空間に集約することによって，学習音声に十分近い新しい韻律パターンも合成することができる．また，従来独立に制御していたピッチ軌跡，パワー軌跡，モーラ長といった韻律成分を，それらの相関関係を保持しながらすべてを同時に制御することができる．部分空間は，アクセント型とモーラ数の組ごとに算出し，学習音声に対する感情の主観評価値ベクトルと重回帰式で対応付ける．音声を合成する際には，所望の感情から部分空間を経由して韻律パターンを生成し，TD-PSOLA 法によって音声波形を合成する．実験により，4 次元程度の部分空間で韻律パターンの変動の約 90%を表現でき，さらに感情の主観評価値と精度良く対応付けられることを確認した．学習内および学習外のアクセント句を用い，感情を段階的に与えた合成音声に対して聴取実験を行った結果，「悲しい」「退屈」「怒り」「驚き」「落胆」「嫌悪」について，与えた感情が聴き手に知覚された．

A Synthesis Method of Emotional Speech Using Subspace Constraints in Prosody

Tsuyoshi Moriyama,[†1,*1] Shinya Mori[†1,*2] and Shinji Ozawa[†1,*3]

An efficient method of speech synthesis that uses subspace constraint between prosody components is proposed. Conventional methods that utilize unit selection concatenate speech segments, which require enormous number of waveforms stored in database for synthesizing arbitrary texts with various emotional expressions. The proposed method employs principal component analysis to reduce the dimensionality of prosodic components, which also allows us to generate new prosody that is similar to training samples. Experimental results demonstrated that only about 4 dimensions were sufficient for representing the prosodic changes due to emotion at over 90% of the total variance. Synthesized emotions in steps were successfully recognized by the listeners for “sorrow”, “boredom”, “anger”, “surprise”, “depression” and “disgust”.

1. まえがき

近年の情報メディアの普及にともない，人工的に音声を合成する技術は，音声案内や読み上げを行わせる従来の用途から，娯楽，通信，教育，福祉と広範な分野にわたって，人間のように豊かな感情表現を行わせる用途に推移してきた．それにともなって，感情を伝える音声の合成に関する研究がさかんに行われている[1)–3)]．

音声による感情の伝達には，音声の韻律成分が重要な役割を担うといわれている[4),5)]．コーパスを用いて韻律制御を行う方法[6),7)] では，実際の音声を録音した合成単位を切り貼りするため，自然性の高い合成音声が期待される．その反面，感情のように種類と強度が多様な文脈的要因については，組合せの数に応じて，膨大な量の合成単位が必要となる．これを回避するために，統計モデルによる韻律の変形が試みられている[8)] が，モデルが複雑化するという問題がある．

また，ピッチや発話長といった韻律の各成分について，従来はこれらを独立に制御していた[6),9)]．高い自然性を実現するためには，すべての韻律成分を互いの相関を考慮して合成する必要がある．

音声の韻律パターンが感情を含むことによって変形する現象は，変形可能なパターンが，ある要因によって変形する事象の 1 つと考えられる．このような事象の例として，画像中の顔パターンが表情によって変形する事象があるが，主成分分析で求める統計モデル（部分空間）による表現法が有効であると報告されている[10)]．音声現象については，従来，韻律成分の次元圧縮[11)] や話者分散の表現[12)] に用いられてきた．感情という要因による韻律パターンの変形に対して部分空間による表現を用いることができれば，高次の韻律パターンを，その成分間の強い相関を利用して低次元で表現でき，また部分空間上で，学習した韻律パターンに十分近い新しい韻律パターンも合成できる．すなわち限られた学習音声から，コーパス

†1 慶應義塾大学理工学部情報工学科
Department of Information and Computer Science, Keio University
*1 現在，東京工芸大学工学部メディア画像学科
Presently with Tokyo Polytechnic University
*2 現在，株式会社リコー
Presently with Ricoh Company, Ltd.
*3 現在，愛知工科大学工学部情報メディア学科
Presently with Aichi University of Technology

に基づく方法以上の合成自由度を実現し，かつ，すべての韻律成分の時間パターンを，感情伝達における自然な相関関係を保持するように合成することができると考えられる．

本論文では，音声の韻律成分が平均のパターンから少しずつ変形して感情を伝達している点に着目し，様々な感情を含んだ音声のデータベースから抽出される韻律パターンの集合を，主成分分析を用いて低次元の部分空間で表現する．得られた部分空間と感情を対応付けることにより，所望の感情から音声の韻律パターンを合成する手法を提案する．アクセント句の種類ごとに対応付けを行うことにより，任意のアクセント句で感情を伝える音声を合成することができる．

2. 音声による感情の伝達

2.1 音声によって伝えられる感情

音声によって伝えられる感情には，話し手自身が表出したと自覚する「話し手の感情」と，音声の聴き手が音声から受け取る「聴き手の感情」の 2 つが考えられる．話し手の感情は，必ずしも「聴き手の感情」と一致するとは限らず，また，音声から観測することはできない．これに対し，「聴き手の感情」は，「話し手の感情」とは無関係に，音声のみから決定することができる．したがって，合成音声に含ませた感情が，聴き手に伝わるような音声合成規則を求めるためには，「聴き手の感情」と音声の物理的特徴の関係を定式化すべきと考える．以下，「聴き手の感情」を伝達する音声を感情音声と呼ぶ．

2.2 感情音声の物理的特徴

音声による感情の伝達には，従来，声の高さ，大きさ，発話長といった韻律成分が主に寄与していると報告されている[13)–15)]．

2.2.1 感情を含むことによる韻律の揺らぎ

図 1 は，アクセント句/arayuru/を，平静を含む様々な感情を含んで発話した音声からピッチ軌跡をそれぞれ抽出し，同じフレーム数に正規化した後，それらの平均軌跡とともに重畳表示したものである．図から，感情を含むことによる音声の韻律成分の変動が，ある平均からの揺らぎとしてとらえられる様子が分かる．重永[16)] も感情は心理的な状態の平静からの “ずれ” であり，同時に物理的特徴量の平静からの “ずれ” が，感情の判別に有効であると指摘している．

提案する手法では，韻律が正規分布に従って変動することを仮定し，様々な感情を含んで話された学習音声の韻律パターンを，それらの平均と平均からの変動分との和のモデルで表現する．正規分布に従うパターンの多様性を，主成分分析で抽出される互いに無相関な直交

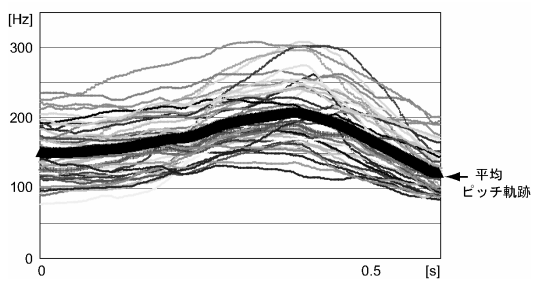


図 1　感情を含むことによるピッチ軌跡の揺らぎの例
Fig. 1　An example set of pitch contours that vary over various emotions conveyed.

基底（部分空間）で表現する手法は，顔画像認識[17)] や音声認識における話者適応[18)]，音声合成における話者性の表現[19)] で有効性が確認されている．韻律パターンを部分空間で表現することの利点には，学習音声の韻律パターンに含まれる冗長な成分を低次元な主成分に集約できる点，部分空間内で韻律成分間の相関関係が保持されることから，学習音声に十分近い新しい韻律パターンも合成できる点，すべての韻律成分の時間パターンを同時に合成できる点があげられる．

2.2.2 韻律の揺らぎのアクセント句への依存性

日本語の 1 文を構成する単位であるアクセント句は，それぞれアクセント型と，日本語に特徴的なモーラ（拍）数を有する．アクセント型は，全体的な抑揚のパターンを決定し，モーラ数は，単語を時間的に分割し，アクセント型で特定されたアクセント核が音声のどの時間範囲に存在するかを特定する．

アクセント型およびモーラ数は，音声が感情を含むときに生ずる韻律の変動の仕方を決定付ける要因と考えられる．たとえば，/naname/というアクセント句は，中高型のアクセント型，すなわち第 2 モーラ（2 番目の/na/）にアクセント核を持ち，モーラ数は 3 である．/naname/が「怒り」を含むと，そのピッチは語頭では低く，アクセント核に向かって高くなるのに対し，異なるアクセント型およびモーラ数を有する/naniyorimo/（第 1 モーラ/na/にアクセント核（頭高型），モーラ数は 5）では，語頭が高くなり，緩やかに低くなるというように，異なるピッチ軌跡の変動を示す．また，/naname/と同じアクセント型およびモーラ数を有する/amado/が「怒り」を含むと，/naname/とまったく同様の変動を示す．

そこで提案する手法では，韻律の変動規則がアクセント型とモーラ数の組合せで決まると仮定し，日本語に頻出する組合せ[20),21)] すべてについて，それぞれ独立に韻律の変動規則を

学習する．

3. 韻律の部分空間を用いた感情音声合成

アクセント型とモーラ数の組ごとに，未知の多様な感情を含む音声を収録する（学習音声とする）．学習音声に対して主観評価実験を行い，聴き手の感情を抽出する．聴き手の感情を互いに無相関な量に変換した感情パラメータを説明変数，学習音声の韻律パターンから抽出する部分空間に投影した韻律パラメータを目的変数として，双方を重回帰式によって対応付ける．**図 2** に本手法の概要を示す．合成するアクセント句のアクセント型とモーラ数から，用いる固有ベクトルと重回帰式を決定し，与えた K 個の感情それぞれの強度を要素に持つベクトル $\mathbf{e}$ から韻律パターン $\mathbf{p}$ を求め，ピッチ同期波形重畳合成（TD-PSOLA）法[22] を用いて音声波形を合成する．

3.1 韻律の部分空間の算出

各学習音声からピッチ軌跡，パワー軌跡，モーラ長を含む韻律パターン $\mathbf{p}_i = [f_{i1}, f_{i2}, \cdots, f_{iL},\ a_{i1}, a_{i2}, \cdots, a_{iL},\ l_{i1}, l_{i2}, \cdots, l_{in}]^T$（$i = 1, \ldots, N$，$N$：学習音声の総数，$L$：フレーム数，$n$：モーラ数）を抽出する．

主成分分析において韻律パターンの次元数を揃える必要があるため，フレーム数 L を L_0（$\propto n$）に正規化する．この際，ピッチおよびパワーの軌跡について，すべてのモーラの長さを均等化し，その際の伸縮比を保存する．また，ピッチ軌跡，パワー軌跡，モーラ長の単位を揃えるために，それぞれ平均 0，分散 1 を持つように正規化する．元の平均と分散の値は保存し，韻律を合成する際に用いる．

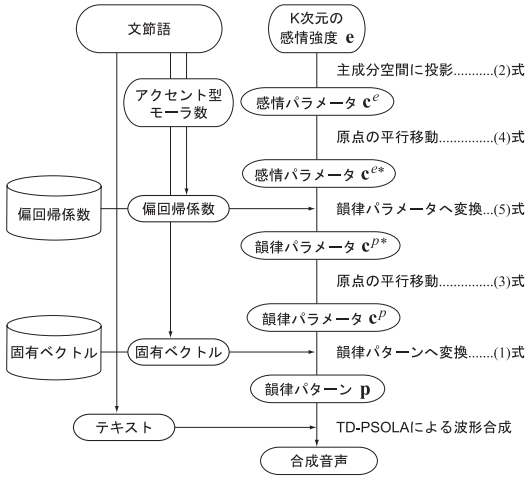


図 2　韻律の部分空間を用いた感情音声合成
Fig. 2　Emotional speech synthesis using subspace constraints in speech prosody.

韻律パターンの相関行列を主成分分析することにより，部分空間を構成する固有値 λ_j^p（$\lambda_j^p \geq \lambda_{j+1}^p$）および対応する固有ベクトル $\mathbf{v}_j^p$（$2L_0 + n$ 次元列ベクトル）を求める．$\bar{\mathbf{p}}$ を韻律の平均パターンとすると，韻律パターン $\mathbf{p}$ は次式で表すことができる．

$$\mathbf{p} = \bar{\mathbf{p}} + \sum_{j=1}^{m} c_j^p \cdot \mathbf{v}_j^p \tag{1}$$

c_j^p は j 番目の主成分の主成分得点（韻律パラメータ）である．m（$m \leq 2L_0 + n$）は用いる主成分の数である．

3.2 韻律の部分空間と感情の対応付け

多重共線性による対応付け精度の劣化を回避するために，主観評価実験によって求められる感情 $\mathbf{e} = [e_1, e_2, \cdots, e_K]^T$（$K$：合成する感情の数）についても主成分分析を行い，互いに無相関な K 次元の主成分得点ベクトル（感情パラメータ）$\mathbf{c}^e$ に変換した後に重回帰分析を行う．

平均感情ベクトルを $\bar{\mathbf{e}}$，主成分分析によって得られる固有ベクトルを $\mathbf{v}_k^e$ とすると，感情ベクトル $\mathbf{e}$ と感情パラメータ $\mathbf{c}^e$ の関係は次式のようになる．

$$\mathbf{e} = \bar{\mathbf{e}} + \sum_{k=1}^{K} c_k^e \cdot \mathbf{v}_k^e \tag{2}$$

式 (1) および式 (2) において，韻律パターンと感情の双方ともに，平均パターンからの変動分をモデル化しているが，主観評価値が平静音声を基準として得られるものであることを考えると，各々の部分空間内で，原点を平均から平静に平行移動した後に対応付けを行う必要があると考えられる．平行移動前の韻律，感情パラメータベクトルをそれぞれ $\mathbf{c}^p$，$\mathbf{c}^e$，平静音声の韻律，感情パラメータベクトルを $\mathbf{c}^{p_0}$，$\mathbf{c}^{e_0}$ とすると，平行移動後の韻律，感情パラメータベクトル $\mathbf{c}^{p*}$，$\mathbf{c}^{e*}$ は次式で求められる．

$$\mathbf{c}^{p*} = \mathbf{c}^p - \mathbf{c}^{p_0} \tag{3}$$

$$\mathbf{c}^{e*} = \mathbf{c}^e - \mathbf{c}^{e_0} \tag{4}$$

平行移動後の韻律，感情パラメータベクトルを重回帰分析によって対応付け，次式を得る．

$$\mathbf{c}^{p*} = \mathbf{R}\,\mathbf{c}^{e*} \tag{5}$$

ここで $\mathbf{R}$ は偏回帰係数行列である．

3.3　与えた感情からの音声合成

与えられた K 次元の感情強度 $\mathbf{e}$ から，式 (2) および式 (4) から感情パラメータ $\mathbf{c}^{e*}$ を求め，式 (5) によって韻律パラメータ $\mathbf{c}^{p*}$ に変換する．式 (3) により部分空間の原点を平行移動した後，式 (1) により韻律パターン $\mathbf{p}$ を求め，3.1 節で保存した平均と分散を用いて単位をピッチ，パワー，モーラ長に合わせることで，合成韻律パターンを得る．求めたパワーおよびピッチの軌跡は，L_0 フレームに正規化されているため，それぞれのモーラを l_{i1}，$\cdots$，l_{in} の長さになるように伸縮し，モーラ長 l_i1，$\cdots$，l_{in} は，先行子音と後続母音の組合せごとに，先行研究で提案されている平均比率[23] に従って，一律に子音部と母音部に分解する．子音部分におけるパワー軌跡の補正は，破裂音および摩擦音，有声子音について，それぞれ経験的に減少比率を定め，最大減少部を頂辺とする台形にパワーを減少させる．

式 (1) および式 (5) における固有ベクトル $\mathbf{v}_j^p$ と偏回帰係数 $\mathbf{R}$ は，合成するアクセント句のアクセント型とモーラ数の組合せに該当するものを用いる．

最後に，得られた韻律パターンを用いて，TD-PSOLA 法で平静音声の波形素片を接続して音声波形を合成する．

4.　実　　験

自ら収録した感情音声データベースから韻律パターンの部分空間を算出し，データベース中の音声に対する感情の主観評価値と対応付けを行った．対応付け精度の評価とともに，学習音声から抽出した韻律パターンと，学習音声の主観評価値から推定した韻律パターンとの誤差の評価を行った．アクセント型とモーラ数の組合せごとに，学習内のアクセント句に加えて学習外のアクセント句を用意し，段階的な感情強度を与えて合成した音声に対して主観評価実験を行った．

4.1　学習音声の収録

韻律パターンの部分空間を算出するために，含まれる感情のみが様々に異なる音声の集合を，アクセント型とモーラ数の組合せごとに用意する必要がある．男性話者 1 名に，**表 2** に示す 47 種類の感情語[24] を提示し，**表 1** に示すアクセント句 20 語を用いて発話したもの，計 940 音声を学習音声として収録した．

話者には，感情表現に熟達した舞台経験のある者を選び，感情を含むことによる韻律の十分な変動を学習するために，表 2 に示す感情語を感情表現の手がかりとして与えた．本手法では，主観評価実験で抽出する聴き手の印象と韻律との対応情報を用いるので，話者が各感情を必ずしも表現できていなくてもよい．

収録した音声は，16 kHz 標本化，16 bit 線形量子化で離散化した．ピッチ軌跡はケプストラム分析によって抽出し，無声もしくは無音部分は，前後の母音部に滑らかにつながるように手動で補正した．モーラ長は，波形の目視によって求めた．パワー軌跡は，短時間平均パワーの軌跡を求め，子音部で減少している箇所については，ピッチ軌跡と同様に，前後の

表 1　アクセント型とモーラ数の組合せごとに選んだ学習用アクセント句
Table 1　Words used for training prosody constraints, that are selected for each combination of accent position and the number of morae.

モーラ数	アクセント型	アクセント句
2	HL	/nama/
	LH	/nami/
3	HLL	/midori/
	LHL	/naname/
	LHH	/nagame/
4	HLLL	/arawani/
	LHLL	/amamizu/
	LHHL	/arayuru/
	LHHH	/omonaga/
5	HLLLL	/naniyorimo/
	LHLLL	/amamizuwa/
	LHHLL	/amanogawa/
	LHHHL	/yawarageru/
	LHHHH	/amarimono/
6	HLLLLL	/emoiwarenu/
	LHLLLL	/omoumamani/
	LHHLLL	/amagaeruwa/
	LHHHLL	/iwazumogana/
	LHHHHL	/oborozukiyo/
	LHHHHH	/warawaremono/

表 2　学習音声収録時に話者が表現しようと試みた感情
Table 2　Emotions that speaker tried to express in recording training samples of speech.

1.	怒り	17.	寛容	33.	満足
2.	喜び	18.	ほくそえむ	34.	退屈
3.	嫌悪	19.	失望	35.	苦しい
4.	侮り	20.	叱責	36.	期待
5.	おかしい	21.	悲しい	37.	幸福
6.	心配	22.	恐れ	38.	好き
7.	優しい	23.	憎い	39.	嫌い
8.	安堵	24.	軽蔑	40.	いや
9.	憤慨	25.	嬉しい	41.	落胆
10.	羞恥	26.	皮肉	42.	非難
11.	穏やか	27.	無関心	43.	不安
12.	憧れ	28.	賞賛	44.	驚き
13.	苛立ち	29.	誇り	45.	慌て
14.	不平	30.	愛	46.	あきれ
15.	切望	31.	嘆き	47.	平静
16.	気の毒な	32.	媚び		

表 3　/naniyorimo/（5 モーラ，HLLLL 型）に関する，韻律パラメータの累積寄与率と，感情ベクトルとの重回帰分析における決定係数（自由度修正済み）
Table 3　Cumulative proportions of the total variance and coefficients of determination for /naniyorimo/ (5 morae and HLLLL-accent).

韻律パラメータ（主成分）	累積寄与率 [%]	決定係数（自由度修正済み）
λ_1^p	63.5	0.784
λ_2^p	79.4	0.749
λ_3^p	88.8	0.410
λ_4^p	93.1	0.305

母音部に滑らかにつながるように手動で補正した．また，正規化フレーム数 L_0 は実験的に $100\,n$（n：モーラ数）とした．

4.2　韻律の部分空間の算出結果

部分空間の算出結果の例として，頭高型のアクセント型を持ち，5 モーラの/naniyorimo/に対し，表 2 を手がかりに発話した 47 音声から韻律パターンを抽出し，主成分分析によって求めた主成分を**表 3** に示す．この例では，韻律パターンはすべて $L_0 = 100 \times 5$（モーラ）$= 500$ フレームに正規化され，ピッチ軌跡（500 次元），パワー軌跡（500 次元），各モーラ長（5 次元）の計 1,005 次元が，累積寄与率が 90%を超える主成分までを考慮すると，4 次元で表現できるという結果となった．アクセント型とモーラ数のすべての組合せについて，第 4 主成分までの累積寄与率を**表 4** に示す．

表 4　韻律パラメータの累積寄与率 [%]（4 次まで）
Table 4　Cumulative proportions of the total variance.

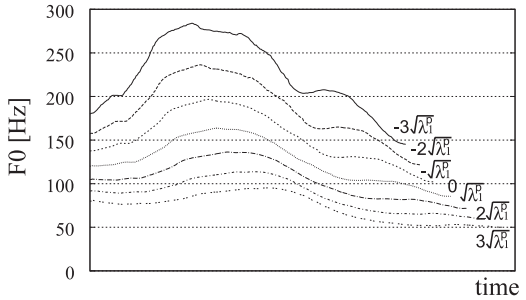
モーラ数	アクセント型	第 1	第 2	第 3	第 4
2	HL	59.3	81.3	89.5	93.3
	LH	56.6	82.3	88.6	92.1
3	HLL	68.1	79.8	88.0	92.2
	LHL	55.8	81.6	92.4	94.8
	LHH	57.1	82.7	90.8	94.0
4	HLLL	56.1	76.8	85.8	92.2
	LHLL	58.7	83.3	88.7	92.3
	LHHL	51.3	76.2	85.8	91.2
	LHHH	49.4	75.9	88.4	91.8
5	HLLLL	63.5	79.4	88.8	93.1
	LHLLL	61.4	77.0	87.7	91.9
	LHHLL	56.3	79.9	87.6	92.6
	LHHHL	64.0	79.0	89.1	92.4
	LHHHH	56.7	79.0	89.5	92.7
6	HLLLLL	72.5	85.4	91.0	94.0
	LHLLLL	63.8	82.6	88.6	93.2
	LHHLLL	60.1	81.5	89.4	93.6
	LHHHLL	47.0	72.3	84.4	89.7
	LHHHHL	61.1	77.6	86.7	90.5
	LHHHHH	54.9	78.5	90.5	94.3

また，/naniyorimo/の例において，式 (1) における韻律パラメータ c_1^p，c_2^p をそれぞれ $-3\sqrt{\lambda_1^p}$ から $3\sqrt{\lambda_1^p}$，$-3\sqrt{\lambda_2^p}$ から $3\sqrt{\lambda_2^p}$（λ_1^p, λ_2^p：第 1，2 主成分の固有値）と増減させたときに得られるピッチ軌跡を**図 3** (a)，(b) に示す．c_1^p が増加するに従ってピッチ軌跡は全体的に減少し，モーラ長が長くなることが分かる．一方，c_2^p が増加するとピッチ軌跡の語頭部分が上昇し，モーラ長は長くなることが分かる．

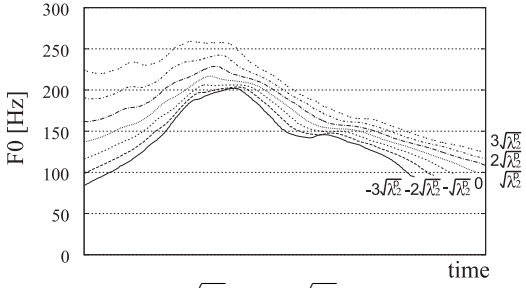
後の処理では，第 4 主成分までを用いることとした．

4.3　学習音声の主観評価実験

表 2 の平静を除く 46 感情語のすべてを用いることは，被験者の負担が大きく，被験者の集中力の低下にともなって評価に悪影響を与える可能性がある．そこでまず音声評価に慣れている被験者 3 名に，1 つのアクセント句（ここでは/arayuru/）の 46 種類の音声について，46 の感情語すべてをそのまま評価項目に用いて予備評価実験を行った．その結果を因



(a) c_1^p を $-3\sqrt{\lambda_1^p}$ から $3\sqrt{\lambda_1^p}$ まで変えた場合



(b) c_2^p を $-3\sqrt{\lambda_2^p}$ から $3\sqrt{\lambda_2^p}$ まで変えた場合

図 3　韻律パラメータを変えることで生成されるピッチ軌跡の例
Fig. 3　Example contours of F0 that are generated by varying prosody parameter.

子分析し，46 因子に関する因子負荷量を要素とする 46 次元ベクトルを，各感情語について求め，ベクトル間のユークリッド距離による最短距離法で樹形図を生成した．そして，樹形図上で隣り合う感情語の一方を削除しながら，一定の距離未満で隣り合う感情語がなくなるまで，樹形図を更新した．その結果，46 の感情語の中から**表 5** に示す 12 の感情語が選択された（最右列にそれぞれに統合された感情語を示す）．被験者は，評価項目の感情語それぞれについて，その感情が音声にどの程度が含まれるかを 7 段階評定尺度（最左：「含まれていない」，最右：「非常に含まれている」と教示し，アンケート用紙上はラジオボタンのみ 7 つ直線に配置）で評価した．また，感情音声のみを繰り返し聴かせると，慣れの効果が生じる恐れがある．そこで平静音声，感情音声の順に聴かせることとした．

次に予備評価実験の 3 名を含む被験者 20 名で，選択された 12 の感情語を評価項目として，表 1 のアクセント句ごとに 46 個の感情音声を評価する実験を行った．

表 5　本論文で合成対象とする感情
Table 5　Emotions synthesized in the current implementation.

選択された感情	統合された感情
e_1　怒り	憤慨，苛立ち，叱責
e_2　喜び	優しい，安堵，穏やか，憧れ，嬉しい，愛，満足，幸福，好き
e_3　嫌悪	不平，憎い，軽蔑，皮肉，嫌い，いや，非難
e_4　驚き	慌て
e_5　侮り	
e_6　誇り	寛容，ほくそえむ，賞賛，期待
e_7　落胆	失望
e_8　おかしい	
e_9　悲しい	心配，気の毒な，恐れ，不安
e_{10} 退屈	無関心
e_{11} 苦しい	切望，嘆き
e_{12} 羞恥	媚び

各学習音声について，20 名による主観評価値 [0 6] から反復切断法[25] により外れ値（平均値から標準偏差の 3 倍以上離れたもの）を除去した後，被験者間の平均を求めて感情ベクトル $\mathbf{e}$ とし，さらに式 (2) および式 (4) から感情パラメータ $\mathbf{c}^{e*}$ を求めた．また，序数尺度である主観評価値を，本実験ではそのまま間隔尺度と見なして用いた．

4.4　韻律の部分空間と感情の対応付け

4.3 節で得た感情パラメータ $\mathbf{c}^{e*}$ を説明変数，4.2 節で得た韻律パラメータ c_j^{p*} を目的変数として，アクセント型とモーラ数の組合せごとに 46 サンプルを用いた重回帰分析を行い，韻律の部分空間と感情の対応付けを行った．

表 3 に重回帰分析の精度を表す決定係数（自由度修正済み）を示す．韻律の主成分は，累積寄与率が約 80%を占める第 1，第 2 主成分に対して，それぞれ決定係数が 0.7 を超える値になっていることから，感情ベクトルによって韻律パターンを精度良く推定できているといえる．また F 検定の結果，第 1 主成分，第 2 主成分ともに，危険率 8%で韻律パラメータ c_j^{p*} を予測できていることを確認した．

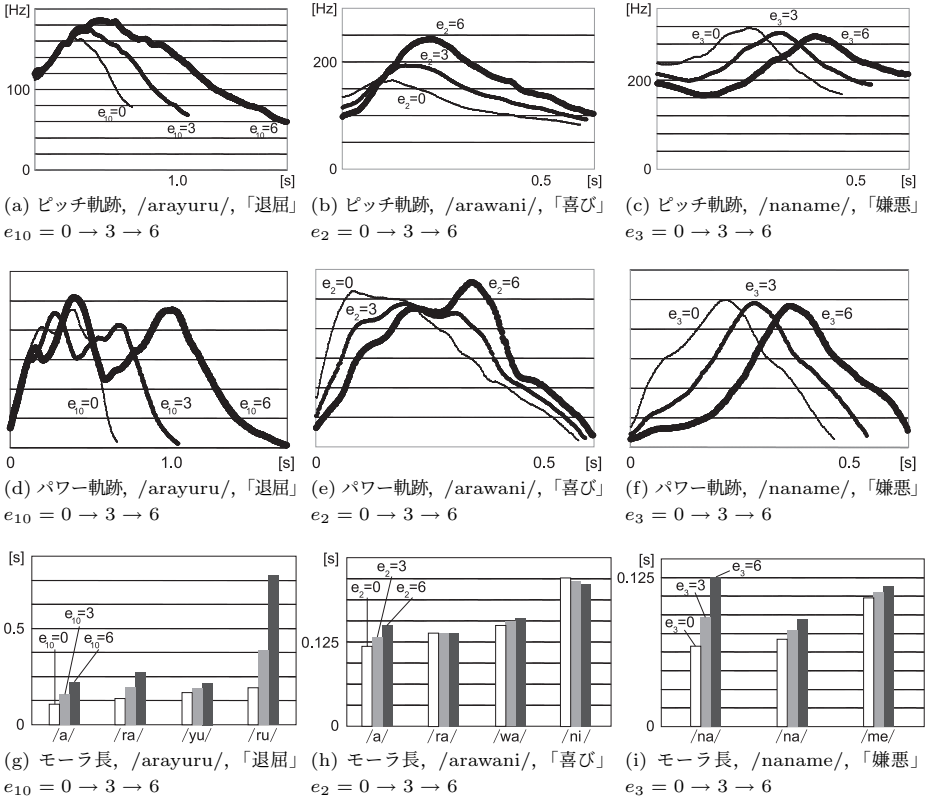
対応情報を用いた韻律パターンの推定精度を定量評価するために，学習音声それぞれから抽出した韻律パターンと，学習音声それぞれに対して得られている主観評価値の感情ベクトルから，対応情報を用いて推定した韻律パターンとで，ピッチ，パワーそしてモーラ長それぞれについて誤差（平均および標準偏差）を算出した結果を**表 6** に示す．ピッチ軌跡については自乗誤差 [oct.$\times 10^{-2}$]，パワー軌跡は式 (6) による原パワー軌跡対誤差比 ϵ_a [dB]，モーラ長は差の絶対値 [ms] を算出した．これらの誤差は，わずかに知覚される程度であった．

表 6　部分空間から推定した韻律パターンの誤差（平均（標準偏差））
Table 6　Reconstruction error of the prosody patterns.

モーラ数	アクセント型	パワー誤差 [dB]	ピッチ誤差 [oct.$\times 10^{-2}$]	モーラ長誤差 [ms]
2	HL	9.14 (5.18)	9.84 (7.88)	37.3 (32.7)
	LH	8.21 (5.06)	11.2 (9.35)	30.3 (36.7)
3	HLL	7.48 (5.52)	12.9 (9.57)	33.6 (39.8)
	LHL	7.65 (5.13)	10.9 (8.72)	26.1 (26.4)
	LHH	8.48 (5.54)	10.0 (8.78)	29.7 (26.5)
4	HLLL	8.50 (5.47)	9.64 (8.12)	19.5 (27.0)
	LHLL	7.72 (5.87)	10.4 (8.58)	21.8 (21.1)
	LHHL	9.35 (5.52)	8.61 (7.09)	19.4 (25.9)
	LHHH	9.01 (5.38)	9.37 (7.36)	20.2 (28.4)
5	HLLLL	8.34 (4.94)	9.56 (8.23)	18.7 (24.0)
	LHLLL	9.87 (5.63)	8.73 (7.22)	15.6 (19.9)
	LHHLL	9.11 (5.34)	8.30 (6.53)	14.7 (15.5)
	LHHHL	9.36 (5.43)	9.72 (7.56)	19.3 (17.3)
	LHHHH	9.14 (5.33)	9.76 (8.34)	16.9 (29.3)
6	HLLLLL	7.93 (5.43)	10.3 (8.80)	18.2 (23.0)
	LHLLLL	7.33 (5.08)	12.2 (10.3)	16.5 (19.1)
	LHHLLL	8.29 (5.43)	10.8 (9.14)	20.1 (27.1)
	LHHHLL	9.29 (5.33)	9.27 (7.18)	14.2 (15.6)
	LHHHHL	7.75 (5.39)	10.5 (8.66)	17.9 (20.9)
	LHHHHH	9.80 (5.42)	10.4 (7.97)	15.3 (18.0)

$$\epsilon_a = \sum_i^{L_0} \left(10 \times \log \frac{a_i}{|a_i - \tilde{a}_i|} \right) \tag{6}$$

与えた感情ベクトルから合成された韻律パターンの例を**図 4** に示す．図 4 (a)，(d)，(g) は，アクセント句/arayuru/に対し，感情ベクトルの「退屈」成分 e_{10} のみを 0→3→6（他の成分は 0 のまま）と増加させることで生成された韻律パターンである．「退屈」成分の増加にともない，ピッチ軌跡は，全体的にわずかに高くなり，モーラ長は，特に最後のモーラで長くなる傾向があった．図 4 (b)，(e)，(h) は，アクセント句/arawani/に対し，「喜び」成分 e_2 のみを 0→3→6 と増加させることで生成された韻律パターンである．「喜び」成分の増大にともない，ピッチ軌跡は，語頭でより低く語尾に向かってより高くなるように変動した．パワー軌跡は，アクセント核に存在したピークが弱まり，後半がより大きくなった．第 3 モーラ/wa/近傍でパワーの著しい増大が見られるのは，母音の音色がより明るく



(a) ピッチ軌跡，/arayuru/，「退屈」$e_{10} = 0 \to 3 \to 6$
(b) ピッチ軌跡，/arawani/，「喜び」$e_2 = 0 \to 3 \to 6$
(c) ピッチ軌跡，/naname/，「嫌悪」$e_3 = 0 \to 3 \to 6$
(d) パワー軌跡，/arayuru/，「退屈」$e_{10} = 0 \to 3 \to 6$
(e) パワー軌跡，/arawani/，「喜び」$e_2 = 0 \to 3 \to 6$
(f) パワー軌跡，/naname/，「嫌悪」$e_3 = 0 \to 3 \to 6$
(g) モーラ長，/arayuru/，「退屈」$e_{10} = 0 \to 3 \to 6$
(h) モーラ長，/arawani/，「喜び」$e_2 = 0 \to 3 \to 6$
(i) モーラ長，/naname/，「嫌悪」$e_3 = 0 \to 3 \to 6$

図 4　合成する感情を段階的に変えた時に合成される韻律パターンの例
Fig. 4　An example set of prosody patterns synthesized by varying emotions to synthesize.

変動することにより，スペクトルの高域が増大したためと考えられる．モーラ長は，最初のモーラのみわずかに長くなる傾向があった．図 4 (c)，(f)，(i) は，アクセント句/naname/に対し，「嫌悪」成分 e_3 のみを 0→3→6 と増加させることで生成された韻律パターンである．「嫌悪」成分の増大にともない，ピッチ軌跡は，語頭がより低くなり，モーラ長は，第 1 モーラが著しく長くなる傾向があった．

4.5　合成音声の主観評価実験

4.4 節で得られた対応情報を用いて，与えた感情から韻律パターンを合成し，TD-PSOLA

表 7　合成音声の主観評価実験で用いた学習外のアクセント句
Table 7　Words used in the subjective listening test, that were not included in the training speech.

モーラ数	アクセント型	学習外のアクセント句
2	HL	/mie/
	LH	/mizu/
3	HLL	/kagawa/
	LHL	/erabu/
	LHH	/omae/
4	HLLL	/borudoo/
	LHLL	/aomori/
	LHHL	/imamade/
	LHHH	/kumamoto/
5	HLLLL	/baieruN/
	LHLLL	/karugarii/
	LHHLL	/yamazakura/
	LHHHL	/hokakebune/
	LHHHH	/arakajime/
6	HLLLLL	/iNguraNdo/
	LHLLLL	/burukkuriN/
	LHHLLL	/hokkaidoo/
	LHHHLL	/airuraNdo/
	LHHHHL	/iiarawasu/
	LHHHHH	/aikawarazu/

法で音声を生成したものに対して主観評価実験を行った．

感情の強度を変えて合成できることを確認するために，表 5 の 12 感情それぞれを，7 段階 [0 6] の 3 および 5 の計 2 段階（他の感情は 0 に据え置き）を与えて音声を合成した．また，アクセント型とモーラ数の組合せに対して 1 通りのアクセント句で学習した結果が，他のアクセント句でも有効であることを確認するために，学習に用いた（学習内）アクセント句とは別に，名詞および副詞，動詞からなる学習外のアクセント句を用いた（**表 7**）．

本来であれば，12 感情各 2 段階それぞれに，学習内 20 語と学習外 20 語の計 40 語を用いて，計 960 個の合成音声に対して実験を行うべきであるが，ここでは被験者の負担を考慮し，12 感情（各 2 段階）ごとに，それぞれ学習内 5 語と学習外 5 語をランダムに選択し，計 240 個の合成音声を用意した．また，アクセント句ごとに平静音声も合成した．

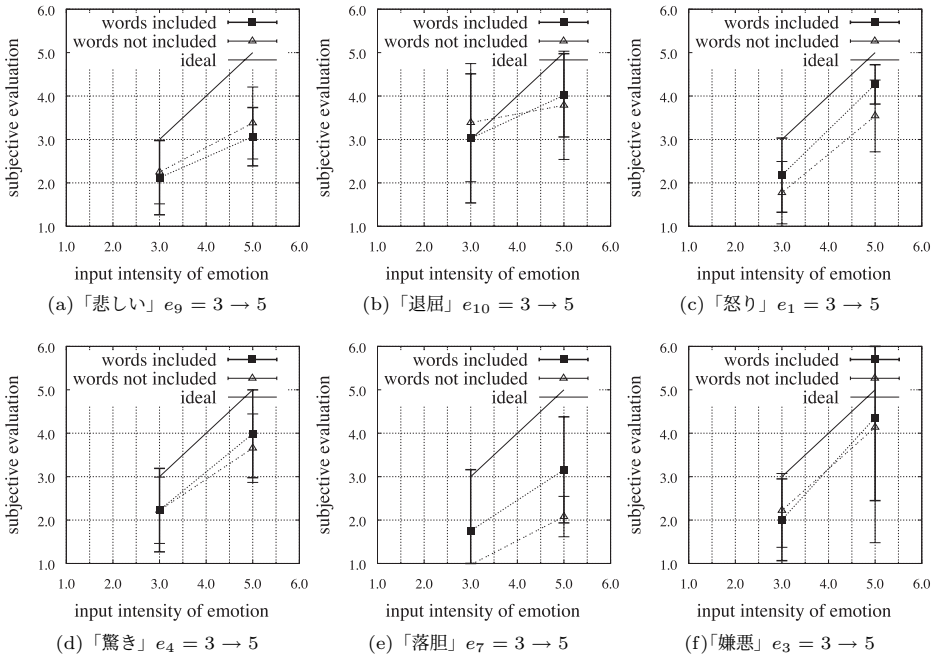
4.3 節と同じ被験者 20 名に，合成した平静音声と感情音声の組を聴かせ，4.3 節と同じ 7 段階評定尺度を用いて評価させた．聴かせる音声はランダムな順番で並べ，被験者は評価が

表 8　学習内アクセント句を用いて合成した感情（行）に対して，各感情（列）が被験者に最も強く感じられた音声数
Table 8　The number of speech that were perceived to contain the synthesized emotion (with words used for training).

	悲し	退屈	怒り	驚き	落胆	嫌悪	喜び	誇り	侮り	苦し	羞恥	おか
悲しい	**10**	0	0	0	0	0	0	0	0	0	0	0
退屈	0	**10**	0	0	0	0	0	0	0	0	0	0
怒り	0	0	**10**	0	0	0	0	0	0	0	0	0
驚き	0	0	0	**10**	0	0	0	0	0	0	0	0
落胆	0	2	0	0	**8**	0	0	0	0	0	0	0
嫌悪	0	0	1	0	2	**7**	0	0	0	0	0	0
喜び	0	0	2	1	0	0	**7**	0	0	0	0	0
誇り	2	1	0	2	0	0	1	**2**	1	0	0	0
侮り	0	0	3	6	0	0	0	1	**0**	0	0	0
苦しい	4	1	0	0	0	0	0	2	1	**1**	1	0
羞恥	0	3	0	0	3	0	2	1	0	0	**0**	1
おかしい	0	0	0	4	0	0	1	5	0	0	0	**0**

表 9　学習外アクセント句を用いて合成した感情（行）に対して，各感情（列）が被験者に最も強く感じられた音声数
Table 9　The number of speech that were perceived to contain the synthesized emotion (with words not used for training).

	悲し	退屈	怒り	驚き	落胆	嫌悪	喜び	誇り	侮り	苦し	羞恥	おか
悲しい	**10**	0	0	0	0	0	0	0	0	0	0	0
退屈	0	**10**	0	0	0	0	0	0	0	0	0	0
怒り	0	0	**9**	0	0	0	0	1	0	0	0	0
驚き	0	0	2	**8**	0	0	0	0	0	0	0	0
落胆	0	3	0	0	**6**	0	0	0	1	0	0	0
嫌悪	0	0	1	0	3	**6**	0	0	0	0	0	0
喜び	0	0	2	3	0	2	**4**	0	0	0	0	0
誇り	0	0	4	1	0	1	0	**4**	0	0	0	0
侮り	0	0	1	4	0	0	0	1	**4**	0	0	0
苦しい	1	1	4	0	2	0	1	1	2	**1**	0	0
羞恥	2	2	0	0	2	1	2	0	0	0	**1**	0
おかしい	1	1	2	3	0	0	1	0	2	0	0	**0**



(a)「悲しい」$e_9 = 3 \rightarrow 5$　(b)「退屈」$e_{10} = 3 \rightarrow 5$　(c)「怒り」$e_1 = 3 \rightarrow 5$

(d)「驚き」$e_4 = 3 \rightarrow 5$　(e)「落胆」$e_7 = 3 \rightarrow 5$　(f)「嫌悪」$e_3 = 3 \rightarrow 5$

図 5　合成する感情を段階的に変えた音声に対する主観評価
Fig. 5　Subjective evaluation for speech synthesized by two levels of emotion intensity.

終わるまで，平静と感情を含ませた合成音声の組を何度聴いてもよいこととした．

主観評価値は，反復切断法による外れ値処理を行った後，被験者間平均値を算出した．各合成音声について，最大の主観評価平均値を得た感情を，その音声から聴き手が知覚した感情とした．2 段階ともに，12 感情語のうちいずれかの感情についてのみ，大きな主観評価値が推定された．

学習内アクセント句，学習外アクセント句に対する結果を，それぞれ**表 8**，**表 9** に示す．各行の合成した感情に対し，各列が被験者が最も強く感じた感情であり（行方向の総個数が 10），対角要素が，感情音声の合成に成功した音声の数といえる．

表 8 を見ると，「悲しい」「退屈」「怒り」「驚き」「落胆」「嫌悪」および「喜び」の 7 感情で，合成した感情が，ほぼ合成したとおりに聴き手に伝わる結果となった．このことから，これらの感情については，音声の韻律成分が，各感情の伝達に有効な物理量であると考えられ，本手法で抽出した部分空間によって，与えた感情から韻律パターンを生成することができたと考えられる．また，表 9 に示すように，上位 6 感情について，学習外アクセント句に対する結果が，学習内アクセント句とほぼ同様の結果となったことから，これらの感情については，アクセント型とモーラ数の組合せで韻律の変動規則が決定付けられる，という本研究の仮説が裏付けられたと考えられる．

また，学習内アクセント句，学習外アクセント句の双方で合成に成功した 6 感情について，主観評価値の分布を**図 5** に示す．プロットは被験者間平均値を，誤差範囲（標準偏差）とともに示す．$(3,3)$ と $(5,5)$ を結ぶ実線（理論値）と比較すると，学習内外双方のアクセント句において，与えた感情の強さの度合いに応じて，段階的に強い感情が知覚されているのが分かる．感情の合成に失敗した残りの 6 感情については，段階的に強度を変えても知覚される強度に変化がなかった．

5. 考　　察

4.3 節で選択した表 5 の 12 感情の間に，弱いながらも相関が認められた．しかし，合成音声に対する主観評価実験では，合成しようとする感情のみに最大の度合いを与え，他を 0 としたため，不自然な音声になった可能性があった．今後，自然な感情を与えるために，感情間の相関を除去した主成分空間 $\mathbf{c}^e$ を直接用いる等の方策を検討する．

「喜び」「誇り」「侮り」「苦しい」「羞恥」「おかしい」については，良好な結果は得られなかった．その原因としては，これらの感情を伝達するのに重要な物理量が，韻律成分ではないことが考えられる[2),26)]．また，これらの感情については，音韻の違いに起因する韻律成分の分散の影響が無視できなかった可能性がある．今後，本論文で用いた韻律成分に加えて，音韻の影響の少なく，かつ声質を表す高域のスペクトル等を部分空間の算出に組み入れることを検討する[5),27)]．さらに特殊拍を含めた音韻の種類による影響について，検討を行う必要がある．

本実験では，男性話者 1 名の音声について，韻律成分の部分空間を生成し，同一人物の波形素片を用いて音声を合成した．これは，感情を含むことによって生ずる韻律成分の変動が，話者ごとに固有で，一貫したものであると仮定したためである．本手法を複数話者へ拡張するためには，話者ごとの学習音声を用いて，話者ごとに部分空間を生成する必要がある．今後，話者ごとの部分空間の間に関係を見い出すことができれば，限られた学習音声から，複数話者へ拡張できると考えられる．

6. ま と め

本論文では，韻律の部分空間を用いて感情音声を合成する手法を提案した．

感情音声の韻律パラメータに対し主成分分析を行うことで韻律の部分空間を算出し，それと主観評価実験から抽出した感情を対応付けることで，感情から韻律を合成する手法を提案した．

男性話者 1 名の単語音声に関して，韻律の部分空間を算出した結果，感情を含むことによる韻律の変動は，その 90%程度を第 4 主成分までで表現でき，累積寄与率が 80%程度になる第 1 主成分，第 2 主成分を，重回帰式によって感情ベクトルから推定できることを示した．実験により，獲得した対応情報を基に，与えた感情ベクトルから韻律パターンを合成できることを示した．合成された感情音声に対して行った主観評価実験の結果，「怒り」「嫌悪」「驚き」「落胆」「悲しい」「退屈」で，学習外のアクセント句でも合成したとおりに知覚されること，段階的な感情の強度が知覚されることを確認した．なお，本手法で合成した音声をウェブサイト[28)]に掲載した．

今後は，自然性の評価実験を行うとともに，声質に関する特徴量を加えて部分空間を算出する．また，文音声において，感情と同様に種類と強度が多様な他の文脈的要因に対しても本手法を用いることで，感情を含む文音声の合成を行う．

謝辞　本研究の一部は，文部科学省科学技術振興調整費「環境情報獲得のための高信頼性ソフトウェアに関する研究」の支援による．

参 考 文 献

1) Murray, I.R. and Arnott, J.L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustical Society of America*, Vol.93, No.2, pp.1097–1108 (1993).
2) Schröder, M.: Emotional Speech Synthesis – A Review, *Proc. 7th European Conference on Speech Communication and Technology* (*EUROSPEECH'01*), Dalsgaard, P., Lindberg, B. and Benner, H. (Eds.), Vol.1, pp.561–564, Aalborg, Kommunik Grafiske Losninger A/S (2001).
3) Bailly, G., Campbell, N. and Mobius, B.: ISCA Special Session: Hot Topics in Speech Synthesis, *Eurospeech 2003* (2003).
4) Donna, E.: Expressive speech: Production, perception and application to speech synthesis, *Acoustical Science and Technology*, Vol.26, No.4, pp.317–325 (20050700).
5) 石井カルロス寿憲，石黒　浩，萩田紀博：韻律および声質を表現した音響特徴と対話音声におけるパラ言語情報の知覚との関連（音声言語，＜特集＞情報処理技術のフロンティア），情報処理学会論文誌，Vol.47, No.6, pp.1782–1792 (20060615).
6) 桂　聡哉，広瀬啓吉，峯松信明：感情音声合成のための生成過程モデルに基づくコーパスベース韻律生成とその評価，電子情報通信学会技術研究報告，SP2002-184, pp.31–36 (2003).
7) 飯田朱美，ニック・キャンベル，安村通晃：感情表現が可能な合成音声の作成と評価，情報処学会論文誌，Vol.40, No.2, pp.479–486 (1999).
8) 能勢　隆，山岸順一，小林隆夫：重回帰 HSMM を用いた合成音声のスタイル制御，電子情報通信学会技術研究報告，Vol.105, No.572, pp.61–66 (2006).
9) Murray, I.R., Edgington, M.D., Campion, D. and Lynn, J.: Rule-based emotion synthesis using concatenated speech, *Speech and Emotion, ISCA Tutorial and Research Workshop* (*ITRW*), Newcastle, Northern Ireland, UK, pp.173–177 (2000).
10) Blanz, V. and Vetter, T.: A Morphable Model For The Synthesis of 3D Faces, *SIGGRAPH '99*, pp.187–194 (1999).
11) Lee, C., Narayanan, S. and Pieraccini, R.: Recognition of negative emotions from the speech signal, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding ASRU2001*, Trento, Italy, pp.240–243 (2002).
12) Kuhn, R., Junqua, J., Nguyen, P. and Niedzielski, N.: Rapid Speaker Adaptation in Eigenvoice Space, *IEEE Trans. Speech and Audio Processing*, Vol.8, pp.695–707 (2000).
13) Murray, I. and Arnott, J.: Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion, *J. Acoust. Soc. Am.*, Vol.93, No.2, pp.1097–1108 (1993).
14) 桂　聡哉，広瀬啓吉，峯松信明：感情音声合成のための生成過程モデルに基づくコーパスベース韻律生成とその評価，電子情報通信学会技術研究報告，SP2002-184, pp.31–36 (2003).
15) Iida, A., Campbell, N., Iga, S., Higuchi, F. and Yasumura, M.: A Speech Synthesis System for Assisting Communication, *Proc. ISCA Workshop on Speech and Emotion*, pp.167–172 (2000).
16) 重永　実：感情の判別分析からみた感情音声の特性，電子情報通信学会論文誌，Vol.J83-A, No.6, pp.726–735 (2000).
17) Kirby, M. and Sirovich, L.: Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.12, No.1, pp.103–108 (1990).
18) Kuhn, R., Junqua, J.-C., Nguyen, P. and Niedzielski, N.: Rapid Speaker Adaptation in Eigenvoice Space, *IEEE Trans. Speech and Audio Processing*, Vol.8, No.6, pp.695–707 (2000).
19) 小山晃俊，徳田恵一，小林隆夫，北村　正：固有声（eigenvoice）に基づいた音声合

成，日本音響学会講演論文集，Vol.1, pp.219–220 (1999).
20) NHK 放送文化研究所：NHK 日本語発音アクセント辞典，日本放送出版協会 (1998).
21) 近藤公久，天野成昭：日本語の語彙特性，三省堂 (1999).
22) Moulines, E. and Charpentier, F.: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones, *Speech Commun.*, Vol.9, No.5-6, pp.453–467 (1990).
23) 板橋秀一：音声工学，chapter 6, pp.156–159, 森北出版 (2005).
24) 森山　剛，斎藤英雄，小沢慎治：音声における感情表現語と感情表現パラメータの対応付け，電子情報通信学会論文誌，Vol.J82-DII, No.4, pp.703–711 (1999).
25) Hoffmann, R.: *New Clinical Laboratory Stanardization Methods*, Exposition Press (1974).
26) 岩見洋平，戸田智基，川波弘道，猿渡　洋，鹿野清宏：GMM に基づく声質変換を用いた感情音声合成，電子情報通信学会技術研究報告，SP2002-171, pp.11–16 (2003).
27) Mokhtari, P., Pfitzinger, H.R. and Ishi, C.T.: Principal components of glottal waveforms: towards parameterisation and manipulation of laryngeal voice quality, *Proc. ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis* (*VOQUAL'03*), Geneva, Switzerland, pp.133–138 (2003).
28) http://www.mega.t-kougei.ac.jp/contents-design/moriyama/projects/emotion-synthesis.htm

(平成 20 年 6 月 9 日受付)
(平成 20 年 12 月 5 日採録)

森山　剛（正会員）

1994 年慶應義塾大学理工学部電気工学科卒業．1999 年同大学大学院博士課程修了．1999 年東京大学生産技術研究所にて日本学術振興会特別研究員 PD．2001〜2004 年米カーネギーメロン大学ロボティクス研究所ポスドク．2004〜2007 年慶應義塾大学助手を経て，現在東京工芸大学工学部メディア画像学科助教．音声の感情情報処理，顔表情の画像解析，パターン認識の研究に従事．1998 年電子情報通信学会学術奨励賞受賞．また，テノール歌手として音楽活動にも従事．1990〜1993 年慶應義塾ワグネル・ソサィエティー男声合唱団，2001〜2003 年 Pittsburgh Camerata，Pittsburgh Fox Chapel Episcopal．博士（工学）．電子情報通信学会，日本音響学会，日本バーチャルリアリティ学会，IEEE 各会員．

森　真也

2004 年慶應義塾大学理工学部情報工学科卒業．2006 年同大学大学院修士課程修了．現在，株式会社リコーに勤務．音声情報処理，感情情報処理に関する研究に従事．

小沢　慎治（正会員）

1943 年生まれ．1967 年慶應義塾大学工学部電気工学科卒業．1974 年同大学大学院博士課程修了（工学博士）．1970 年同電気工学科助手，同教授，同理工学部情報工学科教授を経て，現在，愛知工科大学工学部情報メディア学科教授．画像と音声のディジタル情報処理に従事．道路画像の解析，スポーツ映像の解析に興味を持っている．計測自動制御学会理事，電気学会 ITS 技術委員会委員長．平成 18 年映像情報メディア学会会長等を歴任，IEEE，電気学会，電子情報通信学会，画像電子学会等各会員．