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

2004－SLP－54（28）
2004／12／21

部分空間法による感情音声合成

森　真也†　　森山　剛†　　小沢　慎治†

† 慶應義塾大学理工学部情報工学科
〒 223-8522 神奈川県横浜市港北区日吉 3-14-1
E-mail: †{morishin,moriyama,ozawa}@ozawa.ics.keio.ac.jp

あらまし　音声の合成は，近年のハードウェアの進歩によって，録音編集方式が一般的となってきており，音素環境，感情，話者性といった音声における多様性を実現できる技術が望まれている．従来提案されていた方法では，その多様性の数だけ音声波形を用意し，それらを結合することによって合成していたが，考慮する要因が増えると，それに伴ってより多くの音声波形を必要とし，また，蓄積された波形以外の合成音を得ることはできなかった．本研究では，主成分分析により音声の多様性を平均の音声からの分散で表現する統計モデルを構築し，含まれる感情を連続的に変化させた音声を文節単位で合成し，それを結合することで，自由なテキストから合成音声を得る手法を提案する．学習サンプルの主成分分析結果及び統計モデルによる音声合成実験の結果より，本手法の有効性を確認したので報告する．

キーワード　音声合成，感情，主成分分析，PSOLA

A Synthesis of Emotional Speech Using *Eigenspeech*

Shinya MORI†, Tsuyoshi MORIYAMA†, and Shinji OZAWA†

† Keio University
Hiyoshi 3-14-1, Kouhoku-ku, Yokohama-shi, Kanagawa, 223-8522 Japan
E-mail: †{morishin,moriyama,ozawa}@ozawa.ics.keio.ac.jp

Abstract　Recent progress in hardware made tex-to-speech system general in waveform splicing. Technology to implement variations in speech, such as phonemic environment, emotion and speaker, is required. The main difficulty in the existing technology is that they require the same number of waveforms as the number of varieties considered. Thus, as the number of factors increases, the more waveforms the system requires. Furthermore, waveforms not stored in the database cannot be generated in principle. We propose a method to synthesize emotional speech from arbitrary text using principal component analysis(PCA). Developing the statistical model from the varience of the speech parameters by PCA, prosodic parameters of the target speech can be generated efficiently. Synthetic speech of a sentense is generated by concatenating the subphrases synthesized separately where emotional information has been generaged from the model proposed. We confirmed the effectiveness of our method by our examining the result of principal component analysis and the quality evaluation of synthesized speech generaged by our statistical model to report.

Key words　speech synthesis, emotion, principal component analysis, PSOLA

1. はじめに

近年，音声処理技術の進歩とともに，音声合成の品質は向上し，様々な音声合成システムが実用化されてきた．介護ロボット，音声案内といったインタフェースの実現のためには，より人間味のある豊かな表現，すなわち話者性，感情を考慮に入れた発話様式を実現する必要がある．

豊かな感情表現を含む自然な音声合成を実現するためにパラメータ合成による方法が検討されたが，近年では，ハードウェアの進歩により，録音編集方式が一般的である．これは，実際の音声波形をそのまま用いることが本質だが，切り貼りをうまくするために様々な種類の音声，すなわち，音素環境，感情，話者性といった要因の組み合わせに対し，膨大な音声波形を用意する必要があった [1].

しかし，ある単語レベルの長さの音声が様々な様式で話される状況を考えると，音声は一つの平均的な抑揚パターンが少しずつ揺らいでそれらの要因を伝達していると考えることができる．すなわち，中央極限定理に従って音声が変動することを仮定し，様々な様式で話された音声群をそれらの平均とそこからの分散によって表現し，構築された統計モデルを用いて合成音を得る方法が有効であると考えられる．

そこで本研究では，様々な感情を込めて発話した音声それぞれについて，韻律パラメータの時系列パターンを算出し，その集合から主成分分析を用いて構築される統計モデルを用いることによって，与えたテキストに所望の感情を付与した音声を合成することのできる手法を提案する．

2. 音声パラメータの統計モデル

様々な様式で話された音声の平均とそこからの多様性が中心極限定理に従うと仮定し，相関行列に基づく主成分分析を用いた統計モデルを提案する．

2.1 部分空間の構築

学習サンプルを分析して得られる韻律パラメータの時系列パターン（以下，韻律パターン）の集合から，次の式 (1) の関係を満たす部分空間を構築することができる．

$$\mathbf{C} = \mathbf{L} \cdot \mathbf{P} \tag{1}$$

ここで，$\mathbf{C}(m \times n)$ は主成分得点行列，$\mathbf{L}(m \times m)$ は主成分負荷量行列，$\mathbf{P}(m \times n)$ は韻律パラメータ行列，m は韻律パラメータ数，n はサンプル音声の数である．

2.2 韻律パターンの合成

音声を合成するための韻律パターンは，次の式 (2) を用いることによって，与えられた主成分空間上の一点から逆変換して求めることができる．同様の手法が顔画像解析等でよく用いられている [2]．

$$\mathbf{p} = \bar{\mathbf{p}} + \sum \mathbf{c} \cdot \mathbf{v} \tag{2}$$

$\mathbf{p}$ は得られる韻律パターン，$\bar{\mathbf{p}}$は学習サンプルの平均韻律パターン，$\mathbf{c}$ は係数（主成分得点），$\mathbf{v}$ は固有ベクトルである．ここで，固有ベクトル $\mathbf{v}$ は次式 (3) で求められる．

$$\mathbf{L} = \sqrt{\mathbf{e}} \cdot \mathbf{V} \tag{3}$$

ここで $\mathbf{e}(m)$ は m 番目の主成分の固有値，$\mathbf{V}$ は $\mathbf{e}(m)$ に対応する固有ベクトルを列ベクトルに持つ固有ベクトル行列である．

2.3 主成分数の決定

次の (4) 式で与えられる主成分の累積寄与率 r は，現象の分散のうち，第 k 主成分までで占める割合を表しているが，r に閾値を設定することで，分散を十分に反映する最適な k が求められる．

$$r = \sum_{k=1}^{m} \frac{e_k}{m} \tag{4}$$

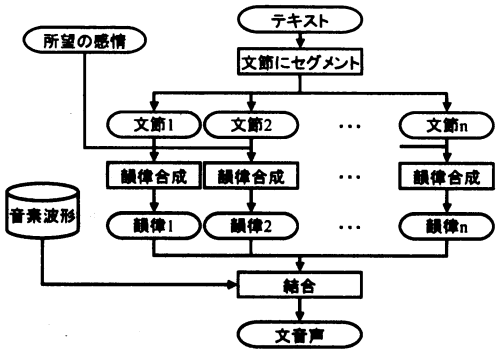


図 1 提案する統計モデルを用いた音声合成方式

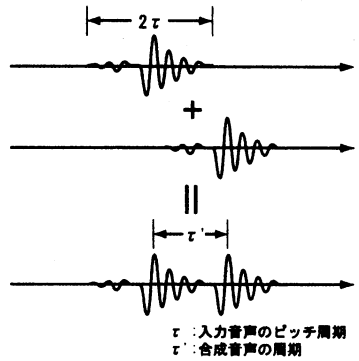


図 2 波形重畳の概念図

3. 統計モデルを用いた音声合成方式

表 1 に，3〜5 モーラについて可能な全てのアクセント型を示す．本研究では，モーラ数とアクセント型の組み合わせごとに韻律の変動が異なると考え，表 1 の 12 通りそれぞれについて統計モデルを構築する．

図 1 に提案する統計モデルを用いた音声合成方式を示す．入力として文テキストと含ませたい感情を与えると，システムはまずテキストを文節単位に分割する．文節テキストそれぞれについて，言葉に固有のモーラ数及びアクセント型に合致したモデルを用いて，所望の感情を含んだ文節音声を合成する．そして，PSOLA(Pitch Synchronous OverLap Add) 法 [3] を用いて全文節音声を結合して文音声とする．

本研究で用いた音声合成手法である PSOLA は，録音編集方式の一つで，図 2 のようにピッチ波形を合成すべき周期 τ' だけずらして重畳することで，周期 τ の音声を合成することができる．また，パワー，ピッチなどといった韻律を自在に操作できるため，比較的少量の音声波形をデータベースに蓄積するだけで任意のテキストについて音声合成ができる．

表 1　実験に用いた音声の発話内容

	モーラ数	アクセント型		発話内容
3-1	3 モーラ	HLL		まえは
3-2	3 モーラ	LHL		いやな
3-3	3 モーラ	LHH		おまえ
4-1	4 モーラ	HLLL		みどりが
4-2	4 モーラ	LHLL		あなたが
4-3	4 モーラ	LHHL		あおぞら
4-4	4 モーラ	LHHH		おまえは
5-1	5 モーラ	HLLLL		どうしたら
5-2	5 モーラ	LHLLL		よまわりは
5-3	5 モーラ	LHHLL		やまざくら
5-4	5 モーラ	LHHHL		いもうとは
5-5	5 モーラ	LHHHH		となりむら

表 2　47 の感情語

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

4. 実　　験

本手法で提案した統計モデルが，音声の自然性を損なわずに元の音声を復元できるかを検証し，さらに構築したモデルで所望の感情を合成する実験を行った．

4.1 学習サンプル

男性話者 1 名により，表 1 の 12 種類の一文節の言葉それぞれに対して，表 2 の 47 種類 [4] の感情をこめて発話された全 564 音声である．音声は全てサンプリング周波数 16kHz，16bit 線形量子化した．

4.2 韻律パラメータ算出

韻律パラメータは，短時間平均パワー軌跡，ピッチ軌跡，全発話長とし，これらで構成される 1 次元ベクトル（韻律パターン）を全ての学習サンプルで求めた．分析条件はフレームの切り出しにハミング窓（窓長 32ms）を用い，フレームシフト幅は 2ms とした．

4.3 主成分分析結果

まず，4.2 で得られた韻律パターンを，全ての学習サンプルでフレーム数が等しくなるように短時間平均パワーとピッチの軌跡を時間方向に伸縮して 250 フレームに標準化した．これに対して単位の影響がなくなるようにデータの標準化を行った後にそれぞれの言葉別々に，パラメータ数 501（パワー 250，ピッチ 250，発話長 1），サンプル数 47（47 感情）のデータの主成分分析を行った．

表 3　HLLL 型における主成分の累積寄与率

	寄与率	累積寄与率
第 1 主成分	59.9%	59.9%
第 2 主成分	10.7%	70.6%
第 3 主成分	7.4%	78.0%
第 4 主成分	4.8%	82.8%
第 5 主成分	3.6%	86.4%
第 6 主成分	3.0%	89.4%
第 7 主成分	2.7%	92.1%
第 8 主成分	1.8%	93.9%
第 9 主成分	1.5%	95.5%
第 10 主成分	0.9%	96.4%
⋮	⋮	⋮

表 3 に HLLL 型のアクセント型の「みどりが」という音声について 4 に従って寄与率・累積寄与率を示した．第 9 主成分までで 95%の累積寄与率を示しており，501 次元のデータが 9 次元までほとんど表現できていることが示された．

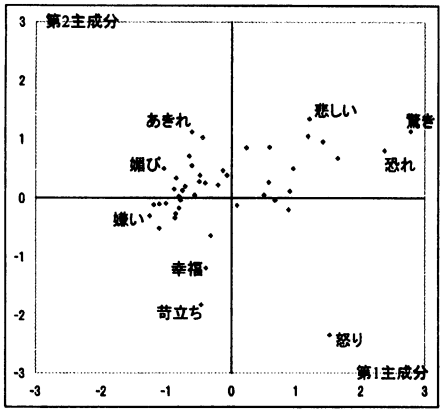
各感情音声の，第 1 主成分得点，第 2 主成分得点の値を同一のアクセント型の言葉について，モーラ数を変えて比較した．その結果を図 3，4，5 に示す．横軸を第 1 主成分，縦軸を第 2 主成分にとり，それぞれ固有値 e(1) 及び e(2) の平方根で与えられるそれぞれの標準偏差の-3 倍から 3 倍の範囲を示している．「怒り」「憤慨」「苛立ち」「叱責」に注目すると，これらは空間中で近距離に位置していることがわかる．

4.4 本モデルによる合成音声の自然性確認実験

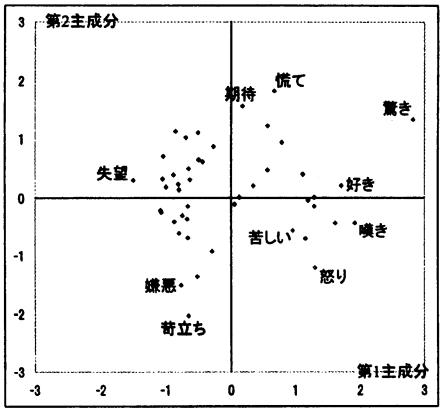
3 章で述べたように，式 (2) で韻律パターンに変換し，音素データベースの音素片を PSOLA 法で結合して文節音声を復元した．この際，第 9 主成分までで累積寄与率が 95%を超え，韻律パターン 501 次元を 9 次元で復元できることが示された．(閾値を 80%とすると 4 次元)

4 モーラのアクセント型 HLLL 型の「みどりが」について，式 (2) に従って主成分得点を主成分空間上で変化させた結果を図 6 に示す．(a) は平均音声，(c) は主成分空間上で「怒り」に位置する音声，(b) は主成分空間上で (a) と (c) のちょうど中間点に位置する音声である．自然な音声波形，パワー，ピッチが得られていることが確認された．

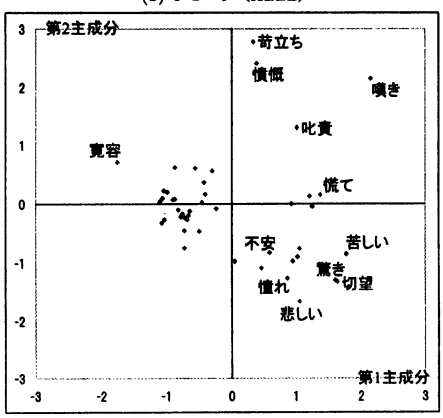
また，5 モーラのアクセント型 LHHHL 型の「いもうとは」について第 1 主成分のみを変化させた結果が図 7 である．(a) は平均音声，(b) は第 1 主成分のみを標準偏差の 1 倍の位置に変化させたもので，(c) は第 1 主成分のみを標準偏差の 2 倍の位置に変化させたものである．第 1 主成分を増加させると，パワーとピッチも同時に増加しているので，これらの第 1 主成分に対する寄与が大きいことがわかる．



(a) 3 モーラ（HLL）

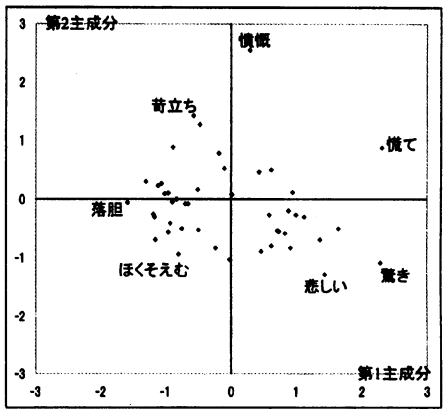


(b) 4 モーラ（HLLL）

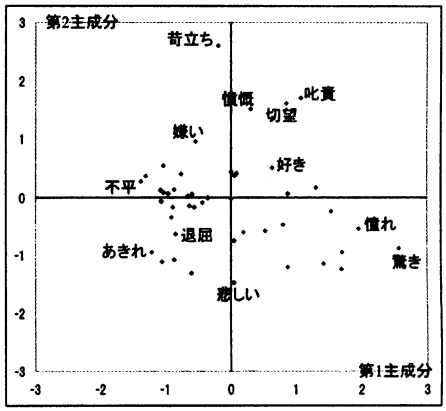


(c) 5 モーラ（HLLLL）

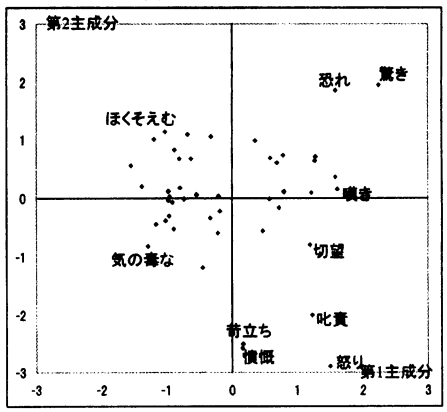
図 3　頭高型アクセント



(a) 3 モーラ（LHL）

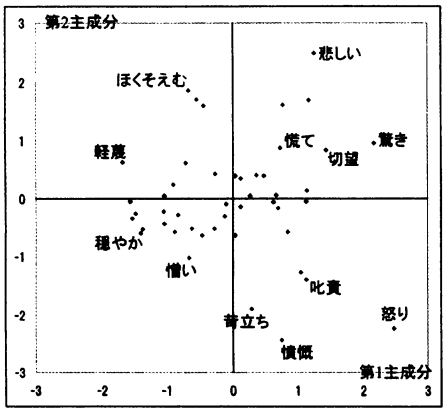


(b) 4 モーラ（LHLL）

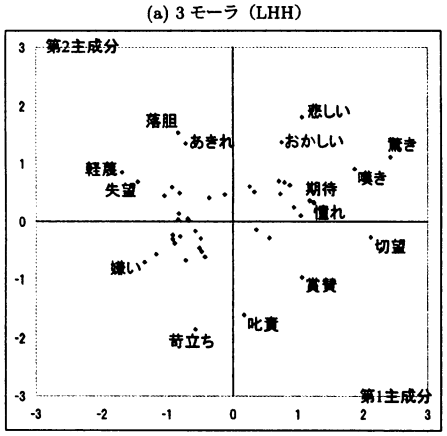


(c) 5 モーラ（LHLLL）

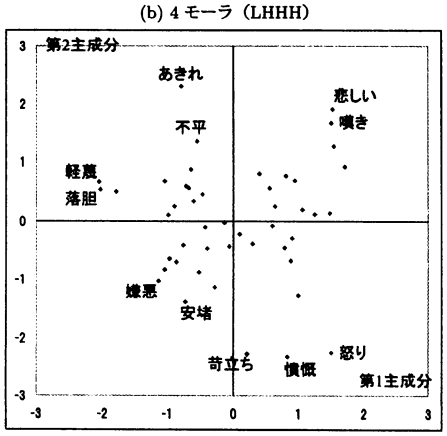
図 4　中高型アクセント



(a) 3 モーラ（LHH）

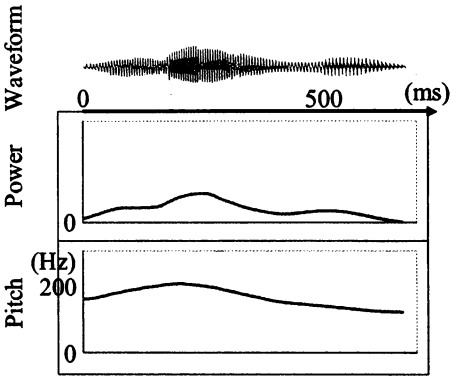


(b) 4 モーラ（LHHH）

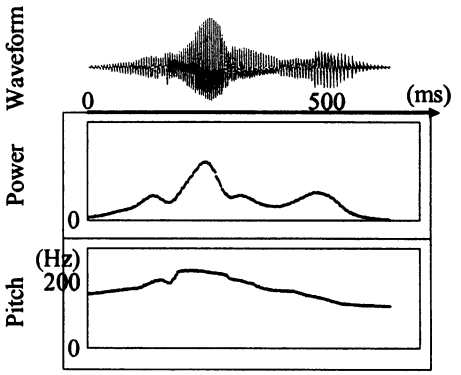


(c) 5 モーラ（LHHHH）

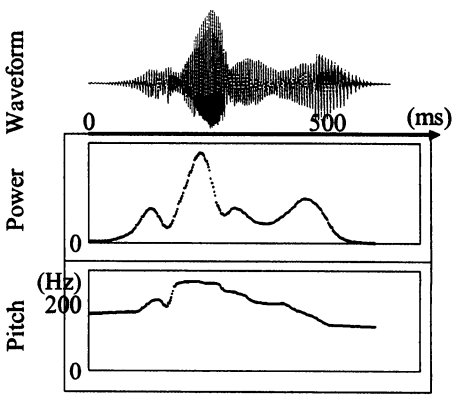
図 5　平板型アクセント



(a) 平均音声

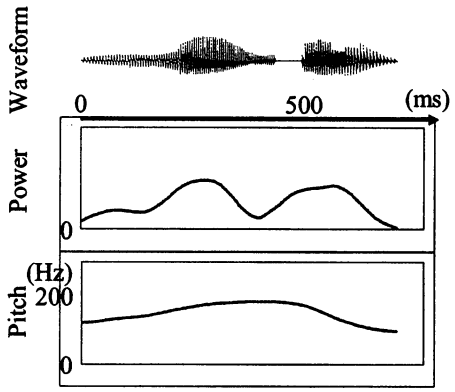


(b) 主成分空間上で平均音声と「怒り」の中間点に位置する音声

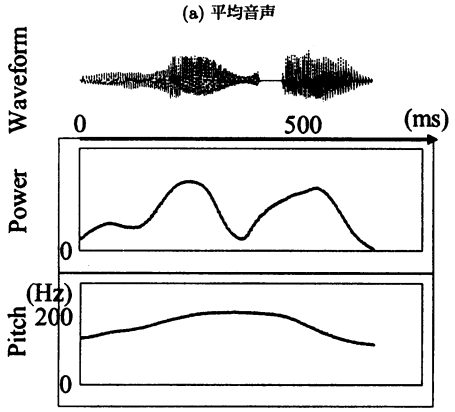


(c) 主成分空間上で「怒り」の位置にある音声

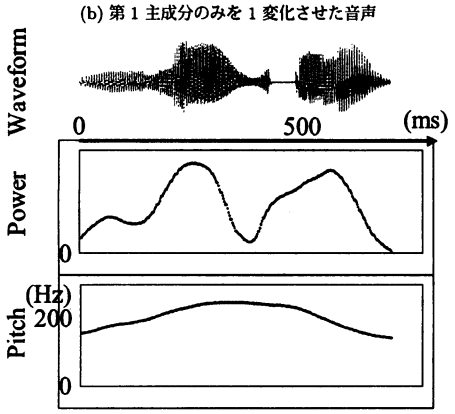
図 6 「みどりが」の波形・パワー・ピッチ



(a) 平均音声



(b) 第 1 主成分のみを 1 変化させた音声



(c) 第 1 主成分のみを 2 変化させた音声

図 7 「いもうとは」の波形・パワー・ピッチ

このように，音声波形及び韻律パターンを復元することができ，予備的に行った簡単な視聴実験においても，自然な感情音声を本統計モデルで復元できることを確認した．

5. ま と め

様々な感情を込めて話された音声の韻律パターンを統計モデルで表現し，与えられたテキストから所望の感情を含んだ音声を合成する手法を提案した．韻律パラメータ間の相関を明らかにし，互いに独立な主成分を求めた結果，95%の累積寄与率までで 501 次元空間を 9 次元の部分空間に圧縮できることがわかった．構築した統計モデルで音声合成実験を行った結果，本モデルで自然な感情音声を合成することができ，本手法の有効性が示された．主成分分析の結果，モーラ数及びアクセント型に関わらず，第一主成分（寄与率約 60%）にピッチとパワーの全体的な上昇/下降が主に寄与することがわかり，主成分と音声生成の解剖学的知見 [5] との関連性が示唆された．今後，提案した統計モデルによって，音声パラメータすなわち音声生成機構の感情表出のメカニズムを明らかにしたい．また，音色を考慮するために，韻律以外の音声パラメータを導入する．学習サンプルとして複数話者のものを用いて話者性を考慮し，合成音の主観評価を行うために聴取実験を行う．

文　献

[1] ニック・キャンベル，"音声合成 CHATR のしくみ" 信学技報，SP98-84，pp47–54，Nov. 1998.

[2] M. A. Turk and T. Veter, "A morphable model for the synthesis of 3D faces," in SIGGRAPH '99, pp.187–194, 1999.

[3] 阪本正治，斉藤隆，鈴木和洋，橋本泰秀，小林メイ，"波形重畳法を用いた日本語テキスト音声合成システムについて" 信学技報，SP95-6，1995.

[4] 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，" 信学技報，SP95–67，pp.9–16，Oct. 1995.

[5] I.R.Murray, J.L.Arnott, "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, Feb. 1993.