韻律の多様性

－多様性の要因・音高知覚・普遍韻律・発話意図の推定・極端な韻律－

Variability of Prosody

- Factors of Variability, Pitch Perception, Universal Prosody, Intension Extraction, Extreme Prosody -

同志社大学　工学部
Faculty of Engineering, Doshisha University

柳田益造
Masuzo YANAGIDA

＜研究協力者＞

情報通信研究機構
けいはんな先端研究センター
Keihanna Advanced Research Center,
National Institute for Communication Technology
白土　保
Tamotsu SHIRADO

同志社大学大学院　工学研究科,
現：龍谷大学　理工学部
School of Engineering, Doshisha University,
currently with Fac. of Eng. Sci., Ryukoku Univ.
三浦雅展
Masanobu MIURA

大阪電気通信大学　第２工学部
Faculty of Engineering II,
Electro-Communication University of Osaka
光本浩士
Hiroshi MITSUMOTO

同志社大学大学院　工学研究科，現：ATR
School of Engineering, Doshisha University,
currently with ATR
安田圭志
Keiji YASUDA

Prosody has been treated as an important theme in speech synthesis as it has to be concretely specified anyway at synthesis. In speech recognition, however, prosody has not been used so much because prosody cannot be the key for resolving uncertainties in speech recognition as it shows a wide variability in speakers' individuality, locality, physical and mental conditions and so forth. Recently, aiming at naturalness of synthetic speech and for detecting speakers' intension or emotion, prosody gathers attention in many aspects in speech information processing. Necessity of grasping realities of prosody arises for improving the quality of synthtic speech and performance of speech recognition, then the present research group has been assigned to investigate variabilities of prosody. This paper summarizes achivements of four-year research works of the authors group concerning variability of prosody, investigating into classification of factors that affect prosody, distinction between language-depedent prosody and language-independent prosody, extraordinary prosody, indeces for representing easiness of pitch perception, application of prosody to irony/praising discrimination with performance comparison between acoustic parameters and prosodic parameters, recognition of stuttering speech, and residual prosody in singing.

Key words: prosody, variability, emotion, intension, individuality, dialect, pitch perception, ironies, stuttering, singing

1　研究の目的

韻律に関する研究は音声合成の分野では古くから重要なテーマとして扱われてきた．それは，音声合成に際しては，何らかの韻律制御をせざるを得ないからで，さしあたり使う必要のない音声認識では，韻律はむしろ積極的には使わないような風潮があったが，最近は合成音声にも自然性が求められ，音声認識では発話者の意図や感情の推定が望まれるようになって，人間の発話における韻律の様態を把握し

ておく必要が生じた．当多様性班はそのような状況下で，韻律の多様性を種々の面から調査・研究することを目的として設定され，本研究の目的はその中で，韻律の多様性を調べ，例えば認識に使う場合の利用のための定量的な表現方法を開発することであった．

この研究目的に関して，本研究では**自然な韻律の採集**とその中の代表的なあるいは特異な音声データの分析結果を示し，また**韻律の利用**としては，皮肉発声と賞賛発声の判別を韻律分析によって行い，その際に音響パラメータと韻律パラメータを用いて動作評価を行い，認識のための**記述能力の評価**を行う．

ここでは，このほかに，韻律知覚の基本的な要件である音高知覚の精度に関する検討，音高知覚のしやすさを表す指標の検討，音声認識での発話スタイルによる音響モデルの選択，音楽関係では，指定された韻律での発声としての歌唱音声における残された自由度，なども検討項目に含めた．

2　韻律の多様性とその要因・現象

2.1　韻律に関与する要因

韻律に関与すると考えられる要因を表 1 に示す．表 1 のように，音声には発話者の年令層や性別，生育環境，生育地域，家庭環境，両親の使用方言，話者の体調や心理状態，発話意図，作為性の有無，聞き手との社会的関係，聞き手の人数，聞き手からの反応の有無，発話の制約として，読み上げか自発発話か，発話環境の音響条件などが言語的・心理的・音響的に複雑に関与し，それは当然ながら韻律にも表れる．ただし，どの要因がどのような影響をどの程度及ぼすのかの定量的な検討は，これまでほとんど行われていない．以下，これまでほとんど取り上げられることのなかった感情の可制御性，作為性，音響フィードバックについて触れておく．

表 1: 韻律に関与すると考えられる要因

事項	分類
話者	
話者の属性	性別, 年齢, 生育環境, 生育地域, 両親の方言
話者の条件	健常／吃音・嗄声など
話者の状態	体調，心理・精神状態，感情
感情の種類	喜び／怒り／悲しみ／驚き／中立／その他
感情の程度	激しい／中程度／穏やか／なし
感情の可制御性	制御不能＝感情音声／制御可＝感情抑制音声
発話意図	陳述／質問／命令／賞賛／非難／叱責／他
作為性	なし＝通常発話，あり＝演技・意識的発話
意識	録音されている／いない，聞き手の聞き方
発話言語	母語／非母語，母方言／非母方言
社会的関係	話者＞聞き手，話者＝聞き手，話者＜聞き手
聞き手	
人数	0=独白,1=対話, 少数=座談, 多数=講演放送
聞き手の応答	可能＝対話・談話，不可能＝講演・放送
制約	
発話の状況	読み上げ／自発的発話，通常／非常事態
明示的な韻律規定	なし＝通常音声／あり＝歌唱音声
適用される韻律則	言語固有の韻律則／言語に依存しない一般則
音響条件	
騒音条件	騒音の大きさ，騒音の種類・特性
音場条件	残響の長さ，主な反射音の遅延時間
幾何学的条件	聞き手への距離，部屋の広さ
音響フィードバック	程度（音量，遅延時間）

2.1.1　感情の可制御性と作為性

感情については，無意識的・不可避的に感情が韻律に出てしまう場合と，意識的にあるいはもっと極端な場合として作為的に韻律に反映させる場合がある [1]．無意識的な場合，その感情が発話に表れても構わないと思う場合，および発話への表出を排除あるいは抑制しなければいけないと思っても出てしまう場合がある．また，作為的な感情には，舞台演技のような場合と現実の場面での意識的演技があり得る．前者の場合は，実際の感情で発話するのではないが，通常は表出が過剰になりやすい．これに対して後者では，特に日本では，発話への表出を控えめにすることが多いように思われる．

2.1.2　音響条件

騒音下での発話には Lombard 効果が現れる．当然韻律も通常のものとは異なったものになる．また，大きなホールなどで，長い時間遅れ（100〜200ms）で自分の発話が返ってくる場合は，韻律はおろか，発話の流暢性にも影響が現れる [2]．

2.1.3　試験的収録についての条件の絞り込み

韻律に対する表 1 の要因の影響をすべて調査することは実験規模と音声データ収集の実現可能性を考えるとほとんど不可能であり，本研究ではそのごく一部について調査・研究を進めたに過ぎない．韻律の多様性に関する調査・研究を進めるための試験的なデータの収集として，韻律を抽出することができるということを唯一の条件にして，自然発話音声の収録可能性を試みた．

2.2　韻律データの収録と分析

韻律データ収録の難しさと解決手段

韻律の実体を調査するには種々の条件での自然発話データを集める必要がある．ところが，同じ文章をいろいろな人がいろいろな条件下で自然に発声す

ることはほとんど期待できないし，基本周波数を抽出できさえすればいいという収録条件でも，その声以外の音が混入しない状態で収録するということは不可能に近い．

そこで，小型MD（Sony MZ-R909）を常時携帯して録音して貰い，回収後，使える場所を探した．現在までに韻律抽出ができたデータ数の一覧を表 2 に示す．このデータの韻律の一部を 2.4 で示す．

表 2: 韻律抽出可能なデータの詳細

データセットID	話者ID	人数	状況	収録場所	主な話題	韻律抽出可能な発話数
1	MS	3	雑談	防音室	単位	33
2	FL	2	対話	下宿	生活	67
3	FL	2	対話	下宿	高校	61
4	FL	2	対話	下宿	授業	36
5	FH	2	対話	自宅	学校	26

2.3 言語依存韻律と普遍韻律

2.3.1 言語依存韻律と普遍韻律

言語に依存した「個別言語の韻律」と言語に依存しない「普遍的な韻律」を別物として扱うべきであることを提唱し，普遍韻律として，発声器官の物理的制約による現象と最小努力の原則に基づく現象を示す．

2.3.2 言語に依存しない韻律則

発声機構を考慮すると，**藤崎モデル**のような形のモデルが導かれる．これに加えて，韻律に関しても**最小努力の法則**が働くであろうということが予測される．以下，中国語の２つの現象を述べる．

一発話内での韻律の下降傾向

中国語に関して，３文字から成る文字列に関するすべての声調の組み合わせ（4^3 種）のそれぞれ３つの３文字列について，北京語話者 (F) １名が発声した音声資料について，各声調の字の母音部分の始点での基本周波数 F_{bi} が，３音節の中の基本周波数の最大値 F_{0max} から見て何セント下がっているかを単語の中での各位置（i = 語頭，語中，語末）について調べた平均値を表 3 に示す [3]．どの声調についてもすべて語内での位置が下がるにしたがって始点の基本周波数が下がっていることが分かる．

韻律のなまけ

人間は何らかの行動する場合，同じ効果を生じることが期待できるならその中でなるべく労力が少なくて済む方略を採る．これは「最小努力の原則」と呼ばれる．これが韻律上にも表れる．これを「韻律のなまけ」と呼んでおく．中国語ではこれが本来の声調からの「なまけ」として表れる．

表 3: 中国語３音節語の各声調の文字の始点の基本周波数 F_{bi} の相対的高さの位置（i = 語頭，語中，語末）依存性（語中の基本周波数の最大値 F_{0max} から見たセント値）

声調	第１声	第２声	第３声	第４声
語頭	-58.8	-268.0	-345.2	-51.2
語中	-120.9	-382.5	-619.2	-140.9
語末	-285.2	-653.0	-1005.7	-166.8

表 3 と同じ３音節語の発声資料について，当該音節内での基本周波数の変化を語中の位置別にセント値で表 4 に示す．値は音節の始点の基本周波数に対する音節尾の基本周波数の比をセント値で表した（したがって第２声は正，第４声は負）ものの平均である．第３声は孤立発声では下降＋上昇であるが，連続発声では後半の上昇部分は観察されない．

表 4: 中国語３音節語における各声調の当該音節内での語中位置別の基本周波数の変化幅（セント）

声調	第１声	第２声	第３声	第４声
語頭	-7.7	141.3	-131.9	-182.2
語中	-7.5	131.1	-178.7	-206.6
語末	-17.1	116.3	-180.2	-269.7

2.4 個人内の F_0 変動と平均値の使い分け

2.4.1 女声の基本周波数の上限と下限

基本周波数の上下限は抽出の際に重要なので，把握しておく必要がある．表 2 の ID=3 の中に，通常考え得る値を越えた例（約 700Hz）があった．個人内での変動幅は 150~700 であった．

2.4.2 言語・状況による F_0 の平均値の使い分け

人は使用言語によって韻律を使い分けている可能性がある．聞き手の数とか発話状況によって，韻律が異なる可能性もある．ここでは極端に異なる状況での発話における F_0 の分布の違いを示す．図 1 は日産のカルロス・ゴーン氏が TV のインタヴューに答えた発話と社員に対して日本語で行った演説での

F_0 の比較である．平均値がかなり異なることが分かる [4].

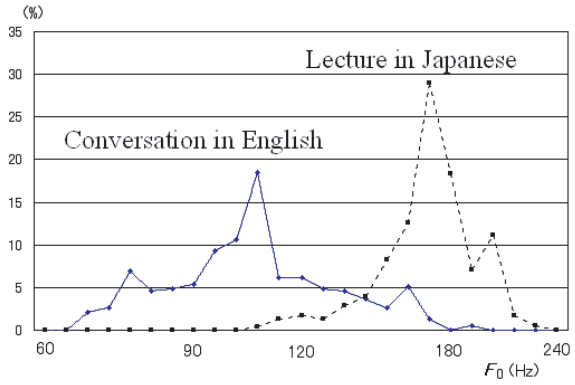

図 1: 母語に近い英語での対話と母語から遠い日本語での講演での F_0 の違い（日産 C. Ghosn 社長）

2.5 特殊状況での韻律

特殊状況での韻律として，幼児の叫び声と高校野球の選手宣誓の例を示す．図 2 は幼児の叫び声「キャー」の “a”の部分で，F_0 が約 1200Hz に達している．図 3 は 2003 年の高校野球の選手宣誓の冒頭部分の F_0 の時間変化を，図 4 は F_0 の頻度分布を示す．これは，声量一杯に発声するので基本周波数はその話者としての最高値にへばりついてしまうためと考えられる．[4]

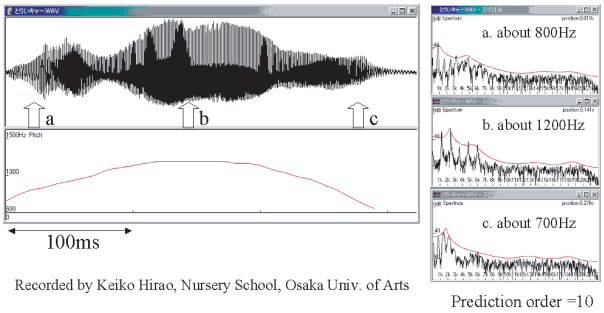

図 2: 幼児の叫び声．中央付近で F_0 が 1200Hz を越える．窓長=10ms, フレームシフト間隔=10ms

3 韻律の知覚

韻律の知覚では，音の高さ（ピッチ）の知覚が基本となる．しかも，音声では連続した音のピッチ知覚によって韻律の判定が行われる．高さは，成分エネルギーが集中している周波数領域の位置あるいは，

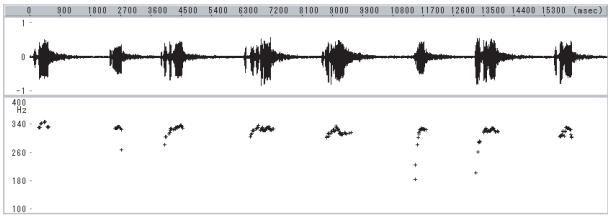

図 3: 2003 年度高校野球の選手宣誓の冒頭の F_0.

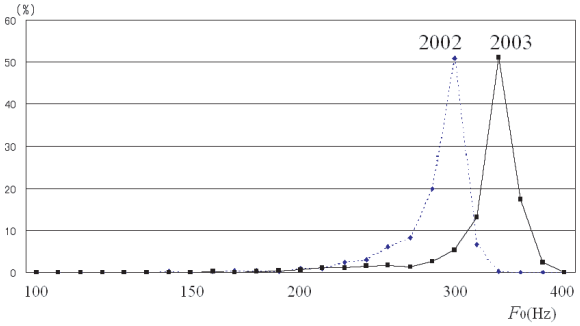

図 4: 高校野球における選手宣誓の F_0 の頻度分布.

周波数スペクトルが調波構造をとる場合にはその成分間の間隔が基本周波数に対応する明確な高さとして知覚される．ここでは．高さの知覚のしやすさを表す指標についての検討と，ごく近い時間間隔で聞こえてくる音の高さの弁別実験を紹介する．

3.1 ピッチの知覚しやすさを表す指標

ヴァーチャルピッチ（以下，VP と略）を含む「ピッチ知覚のしやすさ」を表す簡単な物理指標として，ケプストラル・プロミネンス・ファクター (CPF) を提案し，その妥当性を検討する．

3.1.1 ケプストラル・プロミネンス・ファクター

元の信号 $s(t)$ のサンプル値すべてにペデスタル $P = A \times p \geq 0$（A は信号の振幅を表す）を加えた信号 $s_p(t) = s(t) + P$ に対するケプストラムの，ある注目するケフレンシ τ_m でのケプストラム値 $C_p(\tau_m)$ と，その比較対象とするケフレンシ τ_c でのケプストラム値 $C_p(\tau_c)$ とのレベル差を周波数スペクトルのばらつき具合によって調整したものとして，**ケプストラル・プロミネンス・ファクター** (CPF) $R_p(\tau_m)$ を，次のように定義し [5]， $R_p(\tau_m)$ がピッチ知覚のしやすさを表すことを示す．

$$R_p(\tau_m) = \{20 \log_{10} | C_p(\tau_m)/C_p(\tau_c) |\}/(\bar{d} + 1) \quad (1)$$

3.1.2 CPF の妥当性

現時点での結論として，ペデスタルを調整する値 を 0.4〜0.8 程度としたときの CPF の値が約 0〜0.9dB 以上ならば，ピッチを知覚することができ，また，その大きさによってピッチ知覚のしやすさを表すことができるという結果が得られた．これによって，CPF がピッチ知覚のしやすさを表すことができる指標としての妥当性が検証できた．

3.2 ほぼ同時に聞こえる音の周波数弁別

音声あるいは音楽のように，短い時間間隔を置いて継時的に呈示される音に関する先行音との相対的な高さの弁別について検討した [6].

周波数弁別特性を表す指標

ふたつの純音をいろいろな周波数比で継時的に呈示する．先行音に対して後続音が高く聞こえたかどうかを，反応カテゴリ：“高い,” “同じ,” “低い，” を用いて回答するよう求めるものである．“高い” に関する反応率は累積正規分布関数として，更に反応率を Z-score に変換することにより，反応曲線は直線で表すことができる.

実験結果

立ち上がりの差 d が 30 msec 程度以内ならば，ほぼ同時的なふたつの純音間の立ち上がりのずれを被験者は知覚できないことと，立ち上がりの差 d=30 msec における周波数弁別力は d=0（同時）よりも良いことがわかった.

4 韻律の表現とその利用

韻律を表す従来の表現法としては，基本周波数，パワー，時間などを組み合わせて韻律を表すようにした「音響パラメータ」と，後述の藤崎モデルによるフレーズ指令・アクセント指令の大きさとタイミングという「韻律パラメータ」がある.

4.1 韻律の表現

4.1.1 音響パラメータによる表現

音響パラメータを以下 1. から 8. に示す [4]．発話データのモーラごとの区切り位置の決定は自動化が難しく，マニュアルで行うことが多い.

1. 最終モーラの基本周波数の平均 $\overline{F_{0H}}$ から，発話全体の基本周波数の平均 $\overline{F_0}$ を差し引いて，標準偏差 $\sigma(F_0)$ で正規化した値.

$$(\overline{F_{0H}} - \overline{F_0})/\sigma(F_0) \tag{2}$$

2. 最終モーラの基本周波数の標準偏差 $\sigma(F_{0H})$ を，発話全体の基本周波数の標準偏差 $\sigma(F_0)$ で正規化した値.

$$\sigma(F_{0H})/\sigma(F_0) \tag{3}$$

3. 最終モーラの持続時間長 D_H を，発話全体の 1 モーラあたりの平均持続時間 $\overline{D}$ で正規化した値.

$$D_H/\overline{D} \tag{4}$$

4. 最終モーラの始点における基本周波数 F_{0H_s} から，発話全体の基本周波数の平均 $\overline{F_0}$ を差し引いて，標準偏差 $\sigma(F_0)$ で正規化した値.

$$(F_{0H_s} - \overline{F_0})/\sigma(F_0) \tag{5}$$

5. 最終モーラの中点における基本周波数 F_{0H_c} から，発話全体の基本周波数の平均 $\overline{F_0}$ を差し引いて，標準偏差 $\sigma(F_0)$ で正規化した値.

$$(F_{0H_c} - \overline{F_0})/\sigma(F_0) \tag{6}$$

6. 最終モーラの終点における基本周波数 F_{0H_e} から，発話全体の基本周波数の平均 $\overline{F_0}$ を差し引いて，標準偏差 $\sigma(F_0)$ で正規化した値.

$$(F_{0H_e} - \overline{F_0})/\sigma(F_0) \tag{7}$$

7. 最終モーラのパワーの平均 $\overline{A_H}$ から，発話全体のパワーの平均 $\overline{A}$ を差し引いて，標準偏差 $\sigma(A)$ で正規化した値.

$$(\overline{A_H} - \overline{A})/\sigma(A) \tag{8}$$

8. 最終モーラのパワーの標準偏差 $\sigma(A_H)$ を，発話全体のパワーの標準偏差 $\sigma(A)$ で正規化した値.

$$\sigma(A_H)/\sigma(A) \tag{9}$$

4.1.2 藤崎モデルによる表現

藤崎は，いかなる言語にも適用可能な韻律生成モデルとして，基本周波数 F_0 の時間変化を近似するモデルを提案した．それは次の式で表される.

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{p_i} G_p(t - T_{oi}) + \sum_{j=1}^{J} A_{a_j}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \tag{10}$$

ただし

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \tag{11}$$

および

$$G_a(t) = \begin{cases} \min[1 - (1+\beta t)e^{-\beta t}, \gamma] & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \tag{12}$$

である.

$G_p(t)$ と $G_a(t)$ は，それぞれフレーズ指令とアクセント指令の大きさを表し，それらの中の定数は本論

文では α=1, β=10 and γ=0.9 としている．

式の中の各記号は以下の意味を表す．

F_b：アクセント成分がない場合の F_0 の漸近値

I：フレーズ指令の数

J：アクセント指令の数

A_{p_i}：第 i フレーズ指令の大きさ

A_{a_j}：第 j アクセント指令の大きさ

T_{0i}：第 i フレーズ指令のタイミング

T_{1j}：第 j アクセント指令のオンセット時刻

T_{2j}：第 j アクセント指令の終了時刻

4.1.3 韻律パラメータの推定

式 (10) は $3+2I+3J$ 個の推定すべきパラメータを持っている．しかも，I および J そのものも未知である．これらの値の中で，I, J および T_{**} は，F_0 の時間波形から視察によって決めることができる．また，ここでは文末の主辞の韻律だけが興味の対象であるから，F_b は無視できる．したがって，実際に推定しなければならないパラメータは A_{p_i} と，A_{a_j} だけということになる．

A_{p_i} と A_{a_j} は，次式で表される評価関数を最小化する値の組として，各フレーズおよび各アクセントについて推定することになる．

$$F(F_b, I, J, A_{p_i}, T_{0i}, A_{a_j}, T_{1j}, T_{2j})$$

$$= \int_{\text{voiced}} || \ln \hat{F}_0(t) - \ln F_0(t) ||^2 \, dt \qquad (13)$$

ただし，$\ln \hat{F}_0(t)$ は，$i = 1, 2, \cdots, I$ のフレーズ指令および $j = 1, 2, \cdots, J$ のアクセント指令について想定した値に基づいて生成した F_0 パターンの対数値である．

A_{p_i} と A_{a_j} は，実用的には表 5 に示すいくつかの離散的な値で代表させて，それらの中のどの値をとるかを決めれば十分であるという報告がある．したがって，それらの値の推定は，表 5 の中のどの値になるかという簡単な組み合わせ問題に縮退させることができる．

4.2 皮肉発話と賞賛発話の判別

ここでは，文末につく「ね」音に着目した判別実験について報告する．文末の「ね」は皮肉文の多くにみられる特徴であり，異なる皮肉文にも適応可能である．さらに，一般に，文末音の抽出は，言語処理を用いる必要がなく，工学的に実現しやすいという利点がある．

表 5: フレーズ／アクセント指令の大きさの代表値

指令		記号	大きさ
フレーズ指令		P_0	-0.50
		P_1	0.35
		P_2	0.25
		P_3	0.15
アクセント指令	平坦	F_{H}	0.50
		F_{M}	0.25
		F_{L}	0.10
	起伏	D_{H}	0.50
		D_{M}	0.25
		D_{L}	0.25

ここでは，韻律を定量的に表す方法を 2 種とりあげる．一つは従来の音響的特徴の組み合わせで，もう一つが藤崎モデルに基づく韻律パラメータである．本研究では，それらを用いた皮肉／賞賛の判別実験により，韻律の定量的な記述の効率を比較する．

4.2.1 音声資料

文字で表現した場合には同一の表記になるが，文字通りの意味のほかに，言外の意味を持たせることのできる 20 文について，賞賛と皮肉の 2 様に発声したものを音声資料とした．発声者は 10 人 (内女性 3 人) で，全員東京圏の生まれ，育ち，在住の舞台俳優である．予備的な聴取テストを行い，皮肉と賞賛とがわかりやすい 3 文を表 6 のように選んだ．表 6 には，聴取テストによって聞き手の印象で「友好（賞賛）型」と「反発誘発（皮肉）型」に明確に判断できるもののサンプル数も示した．

表 6: 識別実験で用いた皮肉文と各サンプルの数

ID	文	皮肉発声	賞賛発声
1	お早いお帰りですね.	25	7
2	君は整理が上手だね.	30	28
3	君は頼りになるね.	29	19

4.2.2 判別能力の比較

友好／反感誘発の判別を文末の「ね」の韻律によって判別を行う．文末の「ね」の韻律を用いる理由は，日本語では一般に文末が文の主辞で，そこの韻律は文の性格を決定する上で重要であると考えられるからである．

a. 音響的特徴パラメータを用いた判別

前節で説明した８個の音響パラメータを用い，友好／反感誘発の判別を Mahalanobis 距離によって試みた．判別をする発話データ以外を用いて識別空間を作り，これによって判別した．判別実験は，判別対象を順次取り替えてすべてのデータについて行った．表 7 に結果を示す [7].

表 7: Mahalanobis 距離による判別結果

	反感誘発	友好	合計
正認識数/文章数	70/84	38/57	108/141
正解率	83 %	67 %	77 %

b. 韻律パラメータを用いた判別

藤崎モデルを用いて求められた韻律メータの内，文末の「ね」に関するものをここでの判別のためのパラメータとし，１次元のスカラー量に基づいた皮肉／賞賛の判別を試みた．
終助詞「ね」に関する A_a の平均値は，皮肉発話と賞賛発話の間で，危険率 1%以下で有意差があることが分かった [8] ので，A_{a_j} の境界値を 0.14 として，皮肉／賞賛の判別を行った．結果を表 8 に示す．調査したデータ数が少ないが，正判別率は約 76%で，音響パラメータを用いた場合とほとんど同程度であった．

表 8: A_a による判別

	皮肉	賞賛	計
正判別／全数	17/21	15/21	32/42
正判別率	81 %	71 %	76 %

4.2.3 音響パラメータと韻律パラメータの比較

皮肉発話／賞賛発話の判別のための韻律の記述法として，音響的特徴に基礎をおいた方法と，韻律パラメータによる方法を比較した．音響パラメータは８種をセットで用いたが，韻律パラメータは文末のアクセント指令の大きさ１個だけで動作比較したにも拘わらず，判別率はほとんど同じであった．

4.3 発話スタイルの品詞依存性とその利用

発話スタイルは同一話者であってもそのときの気分とか発話内容によって変わる．しかし，ここではもっと細かく，単語レベルで発話スタイルが異なることを考慮した音響モデルの使い分けを考える [9].

品詞による発話スタイルの違い

名詞・接頭辞・連体詞などについては朗読発話の音響モデルを用いた方が，また感動詞・助動詞類・間投詞などは自然発話の音響モデルを用いた方認識率が上がることが分かった．この事実から，音声認識に際して，単語候補の品詞によって音響モデルを使い分けることによって，認識率を改善できる期待がもてる．

音響モデルの使い分け

自然発話音響モデル優位率の値によって音響モデルを使い分けるようにした．切り分けは，単語単位になるが，その想定した単語の品詞によって適切な方の音響モデルを用いることになる．選択には SVM を用いることが可能である．

認識率の向上

常に最適解を選ぶことができるなら単語誤り率を 13.6%に抑えることができる．これが認識率の上限である．逆に，常にまずい方を選択するなら単語誤り率は 21.7%になる．これが認識率の下限である．SVM を使った方法では，単語誤り率を 16%にすることができた．単一の音響モデルを用いた場合の単語誤り率はどちらの場合も 17.7%であるので，改善率は 1.7 ポイントであった．

4.4 吃音の認識

4.4.1 吃音

発話において，発声が詰まったり上ずったりすることを発話の「非流暢性」という．普通の人でもしばしば観察されるが，この生起頻度が高く，慢性的なものが「吃音」と呼ばれる言語障害の一つの形態である．発症して間もない頃は，音節や音素の繰り返しや音が出ないブロックとかブレーク，音の引き延ばしなどが見られる．これらは吃音の基本的な症状である．[2]

4.4.2 福祉の観点から

計算機の高速化，音声認識技術の発達などによって，近い将来日常生活においても音声認識による機械操作や各種の自動販売機が普及することが見込まれる．しかしながら，音声認識をする装置が吃音を

含む音声をうまく認識できないとなると，自分の声を機械が認識してくれるまで何度も繰り返し言い直さなければならなくなり，機械そのものが吃音者に精神的な苦痛を与えることになりかねない．

本稿では，吃音の機械認識を，単語発声の場合と連続音声の場合について，それぞれに対応した手法を試みた結果を述べる．

4.4.3 吃音の認識

吃音の認識について，単語認識には語頭音節を重複させた逆ＤＰによって，また連続音声には繰り返し部分を削除することによって，吃音にある程度対応できる方法を開発した．以下，概要を述べる．[10]

a. 単語認識のための語頭音節重畳逆行ＤＰ

「第１モーラ音の繰り返しが何回か起きる可能性が高い」ことを積極的に使い，さらに接頭語の問題もクリアーするために，標準パターンの第１モーラを数段重ねたものを改めて標準パターンとして使う．これを「語頭音節重畳逆行ＤＰ（Augmented Backward DP, AB-DP）」と呼んでいる．これによって語中のブレイクや特定音素の「引き伸ばし」にも対処できるようになる．

AB-DP の動作を，従来の端点フリー前向き DP と比較した．AB-DP では，第１モーラを重ねる回数は 6 回とし，入力音声方向の始端および終端は，認識に使う範囲のうちの最初 20%と最後 20%に制限し，DP 処理の終端は，第１モーラを重ねた部分の各先頭ならどこでも構わないとした．

対象語彙は近鉄の駅名 30 で，標準パターンは健常話者の発声 1 回である．サンプリングレートは 16ksamples/sec，距離尺度には 16 次の LPC ケプストラム距離を使い，テストデータは，吃音者１名による「繰り返し型の吃音になった音声」と「繰り返しにはならなかった発声」，標準パターンとは別の健常話者３人の「普通の音声」を各話者について 100 個ずつ用意し，各方法による認識率を調べた．結果を表 9 に示す．

表 9: 端点フリー DP と提案法の動作比較
語彙：駅名 ×30　認識率（%）

	端点フリー DP	提案法
吃音	11	53
吃音者の非吃音	50	72
健常音声	56	85

b. 連続音声認識のための重複部削除法

吃音の単語認識では実音声の終端をＤＰの始点としてＤＰを使えるが，連続音声ではこの手法は使えない．そこで．これに代わる方法を検討した．

ここで述べる吃音処理は，隣接する有音区間について LPC ケプストラム距離を用いて局所的なＤＰマッチングを行い，距離が一定値より近ければ，先行部分を吃音による繰り返しとみなして削除するものである．

実際の削除方法は，有音区間を抽出しておいて，DP を用いてその中で同じようなパターンが隣接して続いている部分を吃音部分として検出することになる．この方法を**「重複部削除法」**と呼ぶ．この方法により，吃音の約 90 %を占める繰り返し型吃音の中の約 60 %を削除することができる．

5　韻律指定発声としての歌唱

歌唱は韻律が離散的に指定された発声であると見なすことができる．ただ，いくら細かく韻律を指定・規定しても，なにがしかの自由度が残る．ここではその一例としてビブラートを取り上げる．また，与えられた歌詞に旋律を付けて和声付けするという過程では，旋律以降は自動処理が可能である．しかし，歌詞からそれに相応しい旋律を，しかもそれを歌える旋律を作り出す方法は確立されていない．我々は，その手法を開発したので，その機構の概略を述べる．

5.1　邦楽と洋楽の比較

5.1.1　音韻明瞭性と響きのバランス

洋楽歌唱においては，そこで使われる声の高さ長さと強さは離散的に指定されている．しかし，細かく指定されていない部分がある．その１例がビブラートである．歌唱は洋楽と邦楽でその目指すところが違う．このことから，邦楽と洋楽ではビブラートの様態がかなり異なる．

5.1.2　ビブラートの比較

ここでは韻律に関わる邦楽と洋楽歌唱の比較として，ビブラートを取り上げる．邦楽（能と狂言）と洋楽の専門家（能８人（内３人は人間国宝），狂言６人（内１人は人間国宝），テノール４人，バリトン５人）に５母音をそのジャンルの唱法で発声してもらった音声 [11] についての，ビブラートの深さと速度および定常に達するまでの立ち上がり時間の比較を表 10

に示す．もう少し詳しく見てみる必要があるが，洋楽のビブラートは半音 (平均律で 100cents) あるいは全音 (200cents) で揺れているように思われるのに対して，能では半音・全音の他に長３度 (400cents，純正律で周波数比が 5/4) あるいは完全４度 (500cents，周波数比は 4/3) の巾で揺れるビブラートがある．

表 10: ５母音のジャンル別唱法によるビブラート

	能	狂言	洋楽
深さ (cents)	329(148)	0()	170(61)
速度 (Hz)	4.7(0.8)	*()	5.2(0.5)
立上がり (sec)	0.28(0.20)	0.19(0.10)	0.14(0.12)

( ) 内は標準偏差

この科研を契機として，大阪芸術大の中山を中心にした邦楽歌唱の研究グループと連携し，当班の力丸と協同で，歌唱時の調音器官のダイナミックスに関する MRI を用いた調査研究を立ち上げた [12].

5.2 歌詞に基づいた旋律の生成

和声法を計算機にインプリメントし，音楽大学における和声課題（バス課題とソプラノ課題）に対処できるシステム BDS [13] および SDS [14] を構築し，BDS については回答の評価システムも開発した [15]．SDS は与えられた旋律に対して和声付けをするシステムで，これをポップス系の旋律にも対応できるようにしたものが AMOR [16]，それを携帯電話に搭載したものが「着つく」[17, 18] である．「着つく」には２つの旋律入力方法がある．一つは携帯端末のボタンによって音符情報を入力する方法，もう一つが歌詞として入力された文章に基づいて，それに相応しい旋律を自動生成する方法である．以下，これらについて説明する．

旋法の決定

調性の枠内で旋律を生成するには，まず旋法（長調／短調）を決定する必要がある．ここでは，入力された文章から旋法を決定するのに，楽しそうな単語と悲しそうな単語の出現頻度によって決めている．

曲構造の決定

通常の場合は，２小節の動機を２つ繋いで小楽節，それを２つ繋いで大楽節とし，８小節の大楽節単位で構成しておくと，新鮮味はないが安定感が得られるので，特に指定がない場合は８小節単位での構成を標準としておく．最後は２重終止のカデンツ型をとるようにする．

和音進行の決定と旋律音の絞り込み

カデンツ型から各部分の和声機能が決められ，各機能について音度表記を選択していく．ここでは和音間遷移の確率を考慮して決定する．音度表記まで進むと，その和音区間で使用できる旋律音の階名が限定される．旋律はこの限定された音群の中から何らかの方法で選んで繋いでいくことになる．方法としては確率を用いるか，既存曲のデータベースを参考にするかになる．

モーラと音符との対応

入力された文章に基づいて，それを歌詞として歌える曲を作るには，音楽のフレーズ境界を歌詞の文節境界と一致させ，各モーラの音符への割り当てに際して単語アクセントを強拍に持ってくるというのが，日本歌曲のやり方であった．音楽が歌詞の制約を受けるのはこの部分である．

5.3 鼻歌の採譜

上のシステムでは，ユーザが自分で思いついた旋律を入力する場合，携帯端末からボタン入力しなければならなかったが，これは譜面に旋律を書くよりも遙かに困難である．この状況を改善する方法として，鼻歌あるいは歌唱のような方法によって旋律入力を行うことを考えた．しかし，絶対音感のない一般ユーザが歌うと，音高の絶対的な高さが一般には規定値（現在では A4=442Hz）になっていないし，声の基本周波数はフラつく．これに対処するための工夫が必要である．また，歌い方の粗雑さや duty 比によって本来の音価（音符の譜面上の長さ）からはほど遠い値になるので，音価の判定も極めて難しい．採譜は，話者（歌い手）が意図した離散化された韻律（音高と持続時間）の推定である．この問題に対して，テンプレートを用いる方法 [19] を開発したので，それを略述する．

音高の判定

調性的な旋律が入力されていると想定して，その基音を判定するため，基本周波数の生起頻度の分布を求め，それと基音周波数を可変とした全音音階テンプレートとの整合が全体としてとれるように基音の周波数を求める．それに基づいて，各階名音の周波数を求める．

音価の判定

音価についてもテンプレート整合法を用いる．すなわち，各音の立ち上がり間の時間長の生起頻度の分布を求め，それと，単位時間の２のベキ乗の間隔のテンプレートを作り，それらの間の整合度を見て実際の時間単位を推定し，それに基づいて音価を判定する．

動作評価

このシステムは声だけではなく楽器音に対しても動作可能なので，動作確認のためにMIDIの鍵盤楽器による演奏，アコースティック・ギターの演奏，および鼻歌の３種の音に対する動作を，音高の認識率と音価の認識率に分けて表11に示しておく．なお，ここでのギターは意図的にチューニングをA4=430Hzとしたものである．音価の認識率が鼻歌よりもギターで悪いのは，減衰による誤判定で，これを救済するために持続判定用の閾値を下げると雑音による誤動作が増える．持続判定をせずにIOI（オンセット間間隔）を音価とすると，休符の抽出が困難になる．

表 11: 自動採譜の動作

	MIDI Piano	ギター	鼻歌
音高の認識率 (%)	100	98	76
音価の認識率 (%)	100	89	93

6　むすび

韻律の多様性とその定量的表現についての４年間の研究をまとめた．

・多様性の調査として，韻律の多様性の要因分類と感情の可制御性など，真の韻律の採集の困難さとその対策，極端な韻律・特殊な韻律の具体例，

・韻律に関する考察として，言語依存の韻律と言語独立な韻律，ピッチの知覚しやすさの指標と隣接音間のピッチ知覚の精度，

・認識への応用として，韻律の定量的表現のための音響パラメータと韻律パラメータの比較，皮肉発声と賞賛発声の韻律に基づいた判別，音声認識における発話スタイルに関する調査に基づいた音響モデルの使い分け，吃音の認識，

・韻律が離散的に指定された発声としての歌唱での残された韻律的自由度であるビブラートに関する邦楽と洋楽の比較，歌詞の韻律に整合した旋律生成，鼻歌からの離散的韻律抽出としての自動採譜

についてまとめた．

参考文献

[1] 柳田益造，滝澤 修，“皮肉発話における話者の心理状態と韻律の関係，” 音講論，2-4-4，1995.

[2] 柳田益造, “音声言語発達の観点から見た吃音　－その原因と治療－,” 信学技報, SP2000-41 (2000-7).

[3] 張燕，柳田益造, “中国語連続音声における声調パターンの変形現象とその規則性,” 信学技報　SP2000-117 (2001.1).

[4] Takasumi Murakami and Masuzo Yanagida, “Variability of Prosody -Extraordinary Cases and Comparative Studies-,” 文科省科研費　特定領域 (2)「韻律」H15 年度成果報告書，pp.19-22 (2004-1).

[5] 熊野他：「ピッチの聞こえやすさを表すケプストラム・プロミネンス・ファクターの提案」，音講論，3-3-3，2001.10.

[6] T. Shirado and M. Yanagida, “Dependency of off-scale sensation on the complexity of melodies,” J. Acoust. Soc. Jpn (E) **17**, 37-39 (1996).

[7] 光本浩士，柳田益造，大多和寛，田村進一，“皮肉音声の音響的特徴に基づいた判別，” 信学技報，SP96-36,pp.17-24,1996.

[8] 辻晋佑，光本浩士，柳田益造, “発話の最終モーラの韻律による知覚的印象の推定,” 音講論　3-6-5 (2001-3).

[9] K. Yasuda, K. Aono, T. Takezawa, S. Yamamoto and M. Yanagida, “A Comparative Study on the Characteristics of Spoken Style Linguistic Information for Accurate Speech Recognition,” WESPAC-8, Melbourne (2003-4).

[10] 柳田益造, “吃音音声の認識,” 情処 第 65 回全国大会講論 (5), T6-3-04，pp.267-270 (2003-3).

[11] Kenji Kojima, Masuzo Yanagida and Ichiro Nakayama, “Variability of Vibrato - A Comparative Study between Japanese Traditional Singing and Western Bel Canto -,” Speech Prosody 2004, Nara, 1b-26, pp.151-154 (2004-3).

[12] 中山一郎，柳田益造，力丸裕, “歌唱時における MRI 画像,” AFIIS Sympo. 2003, pp,110-115 (2004-4).

[13] 三浦雅展，下石坂徹，斉木由美，柳田益造，“和声学におけるバス課題についての回答確認システムの構築とその評価,” 信学論D-　，Vol.J84-D-II, No.6, pp.936-945, (2001-11).

[14] 三浦雅展，柳田益造, “ソプラノ課題の全許容解列挙システムの構築,” 音学誌，第 60 巻，３号，pp.105-114 (2004-03).

[15] 三浦雅展，山田真司，柳田益造, “四声体和声の音楽美を評価するシステム MAESTRO,” 音学誌，第 59 巻，３号，pp.131-140 (2003-3).

[16] 三浦雅展, 黒川誠司, 青井昭博, 尾花　充, 柳田益造, “ポップス系の旋律に対する和声付与システムＡＭＯＲ,” 日本音響学会　MA 研資　MA2003-7 (2003-6).

[17] 柳田益造，三浦雅展，小西秀文，黒川誠司，青井昭博, “音楽データ生成システム、音楽データ生成装置、および音楽データ生成方法,” 特願 2003-097483（平成 15 年 3 月 31 日）.

[18] 読売新聞，“ＩＴ探検隊「メール内容に合わせ作曲」,” 2003.6.6　33 面.

[19] 清水　純，丸山剛志，三浦雅展，柳田益造，“ハミングからの階名と音価の推定,” FIT2004, G-016 (2004-9).