会話音声の韻律生成に向けた
F_0モデル推定、制御特性分析および時間聴覚特性測定

早稲田大学大学院　国際情報通信研究科

Global Information and Telecommunication Institute、 Waseda University

匂坂　芳典

Yoshinori Sagisaka

＜研究協力者＞

ATR 人間情報科学研究所

ATR Human Information Science Research Laboratories

早稲田大学大学院　国際情報通信研究科

Global Information and Telecommunication Institute Waseda University

加藤　宏明 Hiroaki Kato

武藤　牧子 Makiko Muto

小川　博正 Hiromasa Ogawa

山下　琢美 Takumi Yamashita

In this paper, we describe studies on (1) automatic extraction of F_0 control parameters, (2) analyses of F_0 control characteristics of conversational speech and (3) measurements of temporal perceptual characteristics. For (1), we improved a precision of the parameter-estimation of an F_0 model using statistical properties of contents. For (2), from the F_0 observation of short utterances consisting of an adverb and an adjective and their perceptual naturalness scoring test, we showed a consistency of F_0 generation and perception and the possibility of corpus-based F_0 control for conversational speech synthesis. For (3), we found the necessity of a context-dependent error measure for temporal control in speech synthesis from perceptual naturalness measurements using temporally distorted speech. These analyses from multiple viewpoints contribute to the understanding and the modeling of prosody generation, and perception mechanism for the conversational speech prosody control.

Key Words: speech synthesis, corpus-base, prosody control, conversational speech, model for F_0 generation

1. 研究の目的

自然なマンマシンインタフェースを実現する上で、会話音声の韻律の制御は不可欠である．しかしながら、制御特性についてはこれまで十分に調べられてこなかった．制御特性の定量的な分析を図るためには、制御モデル化と共に生成面、知覚面からの知識の拡充が急務である．我々は、これら多面的な理解を進めるため、①制御モデルによる自動分析を進めるためのF_0制御パラメータの自動抽出、②会話音声の発話語彙情報を用いたF_0制御可能性に関する生成・知覚両面からの分析、③時間制御の自然性に関する聴知覚特性の測定を進め、会話音声制御機構の解明をねらった．以下、これら3項目についての検討結果を述べる．

2. F_0制御パラメータの自動抽出

基本周波数制御(F_0)の本質を理解し、音声合成に用いる高性能なF_0制御規則を作成するには、生成モデルに基づくF_0制御パラメータの分析が有効と考えられる．生成モデルに基づいたF_0制御パラメータの自動抽出に関する研究は多くなされている．これらの研究では、F_0パターンのみからF_0制御パラメータを求めている場合が多い．発話内容がわからないまま抽出を行うことは、経験を積んだ専門家でも困難である．本研究では、入力された発話の情報を可能な限り利用し、F_0制御パラメータの抽出を行うことにした．

2.1 F_0制御パラメータとその自動抽出

F_0制御モデルとしては、藤崎らが生成モデルとして提案するものを用いる[1]．このモデルでは、F_0の概形をフレーズ成分とアクセント成分に分解する．これらの成分は、インパルス状のフレーズ指令とステップ状のアクセント指令に対する臨界制動2次線形系で近似される制御機構の出力として得られる．従ってF_0パターンは式(1)で表される：

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (1)$$

ただしここで、F_{min}は話者に依存するF_0パターンの基底周波数であり$G_p(t)$はフレーズ成分、$G_a(t)$アクセント成分であり、以下の式(2)、(3)で表される：

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ 0, & t \leq 0 \end{cases} \quad (2)$$

$$G_a(t) = \begin{cases} \min[1-(1+\beta t)\exp(-\beta t), \gamma], & t \geq 0 \\ 0, & t \leq 0 \end{cases} \quad (3)$$

ここでα、βはそれぞれの制御機構の固有角周波数でありそれぞれ3.0[rad/s]、20.0[rad/s]に固定でき、γはパラメータの上限値で同様に0.9に固定する．さらにA_{pi}、T_{0i}、A_{aj}、T_{1j}、T_{2j}はそれぞれi番目のフレーズ指令の大きさと生起位置、j番目のアクセント指令の振幅と立上り位置、立下り位置である．発話音声からのこれらのパラメータの抽出は、F_0パターンの効率的な表現において重要である．しかしながら観測されたF_0パターンからモデルのパラメータを抽出することはいわゆる逆問題であって、解析的に解くことは出来ないために、モデルのパラメータの初期値を出発点としたAnalysis-by-Synthesis法による逐次近似処理法を用いて抽出する手法が提案され[1]、多くの自動抽出に関する研究が始められている．これまでの研究では、F_0パターンのみからパラメータを抽出しようとする研究が多い．発話内容がわからないまま抽出を行うことは、経験を積んだ専門家でも難しいことから、発話内容から得られる補助的な情報は有用であることが推測できる．本研究では発話内容に関する情報を可能な限り用い、発話内容から直接得られる情報だけでなくデータベースにある他のサンプルから得られるF_0制御パラメータの統計的特性も用いることを考えた．

2.2 発話情報を用いた自動抽出

2.2.1 アクセント句情報を用いた抽出

日本語発話のF_0パターンの概形は、その発話を構成するアクセント句数、アクセント句の長さ、アクセント句のアクセント型から、おおよそ決定できることが知られており、F_0制御パラメータの抽出の際に、これらの情報が全くない場合に比べ、より容易に行うことが出来る．同時に、発話情報として抽出に用いる情報は、音声合成および音声認識の技術を用いることにより求めることが期待できるものに限定した．句境界として考えられる候補は、日本語テキスト・音声合成システムに搭載されている句境界検出モジュールを用いて、限定することが考えられる．また、構成されるアクセント句の情報も同様に得ることができる．文を入力とした自動音素アライナーの出力とF_0パターンからのおおよその境界検出から、フレーズ指令の位置を決めることができる．

2.2.2 抽出手順

フレーズ指令、アクセント指令の位置や大きさの探索を行なう際、これらの初期値が必要である．これらの初期値の予測には線形回帰モデルを用いた．これらの予測モデルは、手動で分析されたF_0制御パラメータとそれに対応する句構成情報を持つ学習データから予め学習する．線形回帰モデルでは、フレーズ指令の生起位置、大きさ、およびアクセント指令の立上り位置、立下り位置、振幅は、構成されるアクセント句の長さ（モーラ数）とそのアクセント型をカテゴリとする線形和により予測される．説明変数がカテゴリ変数であるため、モデルの作成には数量化Ⅰ類[2]を用いた．

自動抽出は以下のように行なわれる．まず、テストセットにおける抽出対象の文が構成する全てのアクセント句のモーラ数とアクセント型を入力値として、回帰モデルからF_0制御パラメータの初期値を予測する．次に、実測F_0パターンとF_0制御パラメータにより生成されるF_0パターンとの誤差が減少しなくなるまで、最適パラメータ値を探索する．探索方法は、藤崎、広瀬[1]の提案方法と同様にした．探索では、フレーズ指令の個数を I、アクセント指令の個数を J とすると、(2I+3J)次元の探索空間を山登り法により探索し、二乗誤差を得る．ここで、アクセント指令の位置は、隣接する指令の位置と重ならないことを前提とする．探索幅は、位置に関しては0.01s、大きさと振幅については0.01とした．

2.3 パラメータ自動抽出実験

2.3.1 抽出実験

最初の実験として、フレーズ成分が1つあるいは2つ、アクセント成分が2つあるいは3つ含まれる短文を対象とした．音声データは、ATR日本語合成用音声データベース[3]に含まれる、2つのアクセント句を含む800文と、3つのアクセント句を含む100文で構成される．アクセント句が2つと3つの場合において回帰モデルを学習する際に、700文および80文を学習データとしてそれぞれ用いた．残りのアクセント句2つを含む100文と、アクセント句3つを含む20文は、テストデータとして用いた．次に一般的なより長い文への抽出方法の適用を考えるために、1つのフレーズ成分、複数個のアクセント成分からなる発話区間を複数個含む一般的な文を対象とした自動抽出実験を行なった．この実験には、発話区間が合計で528含まれる150文を用いた．これらの文は、ATR日本語音声データベース [4]より抜粋した．回帰モデルの学習データとして130文、残りの20文をテストデータとして用いた．全ての学習データおよびテストデータに対して、F_0制御パラメータの手動抽出を行なった．回帰モデルはF_0制御パラメータの初期値を予測する目的で、自動抽出を行なう前に学習データを用いて学習される．初期値の予測が合理的に行なわれていることは、実測のF_0パターンと生成されたF_0パターンを比較することで確認できた．初期値を予測する際の発話情報の必要性を確認するために、初期値の予測は行なわずにパラメータ探索を行なう実験を行なった．初期値の予測を行なわずに探索を行なう際、初期値として学習データのF_0制御パラメータの平均値を用いた．

2.3.2　抽出結果

ここでは、アクセント指令の立上り位置、立下り位置が、手動抽出結果と比較して0.5モーラの範囲内である場合、正解とした．表1に、抽出実験結果を示す．

表1：自動抽出精度

テストセット	発話情報有り	発話情報無し
アクセント句2つ	91.5%	77.0%
アクセント句3つ	91.7%	80.0%
一般的な文	85.9%	72.5%

表1から、発話情報を用いた自動抽出精度は、発話情報を用いない場合に比べて高いことが分かる．しかしながら、一般的な文に対する抽出精度は短文に対する結果ほど高くない．この問題を改善し、抽出精度を向上させるためには、隣接する句の関係を表す情報をさらに与える必要があると考えられる．

2.4　まとめ

本節では、生成モデルに基づくF_0制御パラメータの発話情報を用いた自動抽出方法を提案した．抽出実験結果から、発話情報を最適パラメータ探索の初期値を予測する過程で用いた場合、抽出精度が向上することが確認できた．また、短文と一般的でより長い文に含まれる句単位での抽出精度の違いから、隣接する句の相互関係を示す情報が必要であることも確かめられた．F_0制御パラメータの完全な自動抽出のためには、本稿で提案した手法だけでなく、一般的でより長い文を句単位に自動分割する手法や、フレーズ指令の生起位置推定方法などと併せて改良を進めていく必要がある．

3．発話語彙情報に基づく対話音声韻律制御

テキストからの音声合成では、構文構造とアクセント型等による基本周波数制御が広く行われている．しかし、対話音声にこれらを用いた場合にはその自然性が大きな問題となることが知られている．対話音声の韻律制御に関しては、変化を生ずる原因の特定は無論、変化事象についての科学的分析さえ殆どなされていないのが現状である．我々は、対話に現れる語彙情報から得られる特徴に基づいた制御モデル構築の手始めとして、語彙自体がもつ韻律的な有標性が理解しやすい、程度副詞を用いた対話音声のF_0について、生成、及び知覚の観点から分析を行った．そして、これらから得られた知見に基づき、語彙情報による生成モデルを用いたF_0制御を試みた．

3．1　程度副詞を用いた発話文

日常の生活によく見られるような程度副詞に、形容詞を後続させた 2 文節文を用いて、対話調の音声を収録し、F_0の分析を行った．4 モーラ０型の程度副詞(非常に/相当/割合/そこそこ/普通に/あんまり(~ない))に、形容詞を後続させ 2 文節文を得た．形容詞には、それぞれ、ポジティブ(きれい/うまい/かわいい/優しい/おもしろい)、ネガティブ(汚い/まずい/ぶさいく/厳しい/つまらない)のイメージを伴う 5 対(計 10)の形容詞を集めた．

3.2 程度副詞を用いた対話音声の収録と分析

3.2.1　音声収録

より自然な対話音声を収録するため、被験者には、日常的な質問への返答として自然に発話するようお願いした．たとえば、「非常にうまい」という 2 文節発話文の音声を収録するときには、「味はどう？」という質問に対する返答としてこの発話をする状況をよくイメージさせ、十分イメージできた上で発話させた．

合計で45種類の2文節文を、上記の方法で４人の被験者に発話させ対話調の音声を収録し、その後に、比較用の音声データとして、同様の2文節文を読み上げ調で発話させた音声も収録した．ま

た、実験に用いられた程度副詞の、語彙としての主観的な強勢の強さを調べるため、4人の被験者に、1から10までの数字を使い10段階で評定させた.

3.2.1 母音中心F_0平均による分析の結果と考察

読み上げ調および対話調の音声における、副詞節の箇所のF_0を抽出した. 読み上げ調の音声におけるポジティブイメージを伴う形容詞を後続させた時の副詞節の箇所のF_0平均は、全体的にほぼ変わらない結果となった. つまり読み上げ調の音声においては、語彙の違いによらず同様のF_0制御傾向が示された. 一方、対話調の音声において、ポジティブ・ネガティブのイメージを伴う、それぞれの形容詞を後続させた場合、程度副詞の強勢の強さとF_0平均値との間に強い相関がみられ、後続語のイメージの違いにより、相関の符号が逆になる事が示された. ポジティブイメージを後続させた場合、程度副詞の強勢が強くなるに従い、F_0平均値が上昇しているのに比べ、ネガティブイメージの形容詞を後続させた場合は、F_0平均値は、下降していた.

3.2.3 F_0制御パラメータによる分析の結果と考察

読み上げ調および対話調の音声について、藤崎らの提案するF_0制御パラメータ[5]を抽出した. パラメータの抽出を行う際、程度の違いに伴い大きく変化するのはアクセント指令の大きさAaのみでフレーズの大きさにはほとんど変化が見られなかったため、フレーズの大きさApを固定してパラメータ抽出を行った. その結果、近似による誤差は、Apを固定せずにパラメータ抽出を行った場合とほとんど変わらず、程度の違いによるF_0の変化はAaのみで制御できることが分かった.

ポジティブイメージを伴う形容詞を後続させた場合には、どの被験者のデータについても、程度副詞の強勢の強さが大きくなるにつれ Aa がほぼ一貫して高くなっており、ほぼ同様の制御傾向を示していた. ４人の被験者それぞれについて、主観的な強勢の強さを表す評定スコアと各副詞ごとの Aa の平均との相関を算出したところ、被験者平均0.86(最弱0.81、最強0.93)という非常に強い正の相関を示した.

ネガティブイメージを伴う形容詞を後続させた場合には、ポジティブイメージと伴う形容詞を後続させた場合とは明らかに違う制御傾向を示した. 4人の被験者それぞれについて、主観的な強勢の強さを表す評定スコアと各副詞ごとのAaの平均との相関を算出したところ、被験者Ａについては、正の相関(0.75)を示し、他の3人については、被験者平均 -0.90(最弱-0.76、最強-0.97)という非常に強い負の相関を示した. これらの結果より、程度副詞の強勢の強さおよび後続後の属性に基づいた、対話調の音声合成におけるF_0制御の可能性が確かめられた.

3.3 程度副詞のF_0にみられる対話音声の自然性知覚

3.3.1 自然性評価実験

音声の聴覚特性を調べる実験においては、調べたいパラメータを操作して分析再合成した音声を資料として用いる方法が考えられる. しかし、分析再合成された音声はどうしても不自然さが残り、それらを用いた場合には、我々が調べようとしている人間の発話音声に対する評価が正確に調べられない. そのため、本研究では人間の発話によりF_0の異なる音声を収録し音声資料とした.

60種の発話文をそれぞれ12種の異なるF_0で読み上げ、合計720種の音声サンプルを収録し音声刺激とした. これら12種の音声は、発話時の最大F_0(副詞節の2モーラ目のF_0)が、与えられた楽器音に等しくなるように、男性話者によって発話された. 楽器音は、G音(98.00Hz)から半音ずつF#音(185.00Hz)までの12音である.

標準の聴覚力を持った10人の日本人被験者に、12種それぞれの音声サンプルを、自然さの観点から１(不自然)から5(自然)までの5段階で評価させた. 聞かせる順番は、発話文ごとに12種それぞれの音声のファイル名をパソコンのモニタ上にF_0の高さ順に表示し、各音声を何度でも自由に聞いてそれぞれを比較し相対評価できるようにした.

3.3.2 結果と考察

各音声刺激に与えられた自然性評価スコアの被験者平均を算出した. また、被験者平均において、後続する形容詞にポジティブイメージの形容詞を後続させた場合、ネガティブイメージの形容詞を後続させた場合のそれぞれについて平均スコアを算出した.

結果、用いる副詞の違いおよび形容詞のイメージの違いにより、自然性の高く評価されるF_0が異なることが分かった. ポジティブなイメージを持つ形容詞を後続させた場合には、用いる副詞の程度が強いほどF_0の高い発話が評価されるという相関が見られ、マイナスのイメージを持つ形容詞を後続させた場合には、その相関が逆になっていることが分かった.

3.4 制御モデル

3.4.1 制御モデル

図1に、我々の提案する、語彙情報を用いたF_0制御のあらましを示す. これまでの分析で語彙自体が持つ強さの主観値とAaの値に高い相関があることが判っているため、単調増加な写像関数により語彙自体が持つ強さの主観値からAaを得る.

この値を用いて生成モデルを駆動することにより最終的なF_0を得る.

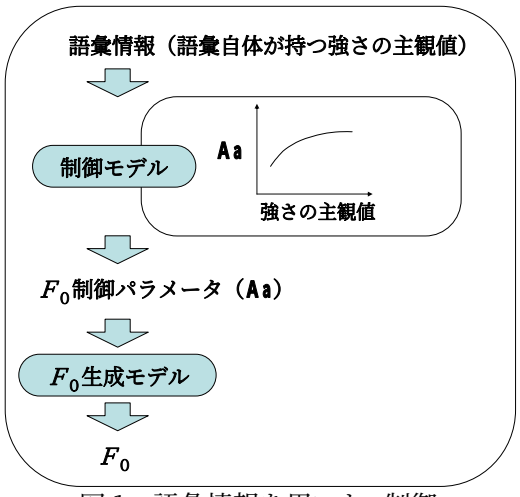

図 1：語彙情報を用いたF_0制御

3.4.2 F_0制御パラメータ予測実験

上述のF_0制御を実現するため、対話音声を用いた写像関数の学習を行った. 分析により得られた全120のデータを学習データとテストデータに分け、Aaの予測精度を調べた. また、制御の話者依存性に関しても検討を行った. 写像関数にはシグモイド関数を用い、全ての学習セットについて、予測値と実測値間の平均二乗誤差RMSEが最小となる基準により、関数のあてはめを行った.

3.4.3 オープンな語彙に対する予測性能の分析

実験は、サンプル数が比較的多いポジティブイメージを伴う形容詞を用いた場合についてジャックナイフテストを行った. すなわち、それぞれの程度副詞を用いた20ずつのデータをテストデータとして6分割し、各々その他5つの程度副詞(100データ)だけを用いた学習によるF_0制御パラメータの予測精度を調べた. 程度副詞の強勢の強さには、4人の被験者によるそれぞれの程度副詞に対する主観的な強勢の強さを表す評定スコアの平均値を用いた. Aaには、平均値と標準偏差について、予測する被験者のデータに合わせて正規化したものを用いた. 学習後の関数に、程度副詞の強勢の強さを代入して得られたAaの値を予測値とし、テストデータにおける実測値とのRMSEを計算した.

それぞれの程度副詞を用いたデータをテストデータとした場合について、(1)他の5つの程度副詞を用いたデータの学習による予測値と実測値とのRMSE(提案モデル値)、(2)程度副詞の主観的強度値の違いを考慮し自身のデータの副詞別の平均値を用いて制御した場合のRMSE(目標値)、(3)程度副詞の違いを考慮せずに自身の全データの平均値を用いて制御した場合のRMSE(従来の制御値)を比較した. その結果、(1)と(3)および(2)と(3)の差は、程度副詞の強勢の強さの違いを考慮することにより誤差を減少できる事が判ったが、(1)は、(2)による誤差の減少と比較し語彙平均70%の誤差の減少を行えておった. これは、Aaを予測する程度副詞についてのデータの学習を行わなくても、それ以外の程度副詞を用いたデータの学習により、学習を行った場合の70%の精度で誤差の減少を行えることを示す.

次に、学習データの少ない場合の予測精度を調べるため、程度副詞6つ中3つの副詞を用いたデータを学習データとした場合の、6つのそれぞれの程度副詞を用いたデータに対する予測精度を調べた. 強勢の強い方からの3つ、中程度の3つ、弱い方からの3つ、そして、強勢の最も強い「非常に」、中程度の「そこそこ」、最も弱い「あんまり」の3つ、以上4通りについて同様の実験を行った.

強勢の強い方3つの程度副詞を用いたデータの学習による予測では、強勢の弱い方の程度副詞を用いたデータに対する予測精度が低く、中程度および弱い方3つの学習による予測では、強勢の強い方に対する予測精度が低くなっていた. 一方、強中弱の3つ(非常に、そこそこ、あんまり)の学習による予測では、全体的に高い精度で予測が行われている. 語彙平均RMSEは0.658であり、全データの学習による予測を行った場合の0.655とほぼ同等の精度を示した. これから、強さの主観値のレンジをカバーできるAaの学習により、その他の強勢の強さを示す程度副詞に対しておおよその予測が行えることが分かった.

3.4.4 制御の話者依存性に関する分析

制御の話者依存性について調べるため、ポジティブイメージを伴う形容詞を用いた場合について、オープンな発話者に対するAaの予測精度を調べた. 4人の被験者によって得られた全120の音声データの中で、それぞれの被験者によって得られた30のデータをテストデータ、残り3人の被験者によって得られた90のデータを学習データとして、写像関数を学習しAaの予測を行った. Aaには、平均値と標準偏差についてテストデータに合わせて正規化したデータを用いた.

4人の被験者それぞれについて、(1)他の3人の学習による予測値と実測値とのRMSE(提案モデル値)、(2) 程度副詞の強勢の強さの違いを考慮し自身の副詞別のデータの平均値を用いて制御した場合のRMSE(目標値)、(3)程度副詞の違いを考慮せずに自身の全データの平均値を用いて制御した場合のRMSE(従来の制御値)を比較した.

その結果、(1)と(3)および(2)と(3)の差は、程度副詞の強勢の強さの違いを考慮することにより誤差を減少できることが示された. また、(1)と(2)との比較から被験者平均によっても92%誤

差減少が可能であることが判った．これらから、他被験者のデータによる学習によりオープンな被験者に対してほぼ予測できることが出来たと言える．発話者による Aa の大きさとばらつきの違いは平均値と標準偏差といった簡単な正規化で対処でき、程度副詞の強勢の強さと Aa との間に見られる制御傾向は発話者にほぼ独立であると思われる．

3.5 まとめ

語彙から得られる情報に基づいたF_0制御を目的とし、対話音声における程度副詞にみられるF_0変化について、生成、知覚、制御の3つの観点から分析を行った．まず、生成の観点から、ある発話コンテキストにおける、程度副詞に形容詞を後続させた2文節発話文において、その形容詞がポジティブもしくはネガティブなイメージを持つ場合に、程度副詞の持つ程度の強さとF_0との間に強い相関が見られ、その相関の符号は、形容詞の持つイメージの符号と一致するという結果を得た．知覚の観点からも、対話音声において語彙情報がF_0自然性に与える同様な影響を確認した．そして、これらの知見に基づき、程度副詞が持つ語彙としての主観的強さによるF_0制御を提案した．語彙および被験者についてオープンなデータに対して、アクセント句成分の大きさの予測実験を行い、この予測が効果的に行えることが判明した．これらの結果より、程度副詞の強勢の強さに基づいて、対話音声のためのF_0制御パラメータの制御が可能であることを確認し、語彙情報に基づくF_0制御の可能性を確かめた．

4. 文音声における音韻時間長伸縮に対する許容度

合成音声の音韻長制御の評価と自然性向上を目的として、音韻長伸縮に対する許容度の研究がされてきたが[6]-[8]、これまでは主として孤立発話の単語音声が対象とされてきた．この音韻長知覚の特性を音声合成の音韻長制御に応用することを考えると、文音声への適用が不可欠である．しかしながら、文における音韻長制御は様々な要因[9]の影響を受けるため、単語において見られた特性がそのまま当てはまらない可能性がある．そこで本研究ではまず単語より大きな単位の文節に注目して、伸縮位置の許容度への効果が単語と同様に見られるか調べた．次に音声合成において大きな課題である発話速度が、音韻長伸縮に対する許容度にどのように影響を与えるか調査した．さらにこれらの音韻長知覚特性が発話時の音韻長制御特性とどのような関係があるのか考察を行った．

4.1 音韻長伸縮の許容度に対する文節内位置の効果

4.1.1 刺激と実験手続き

「合成用日本語音声データベース」(ATR 自動翻訳電話研究所1988)から3モーラ1型の文節3種類を選択し、各文節に対して3文と文節単独発話したデータを使用した．高品質音声分析変換合成法 STRAIGHT[10]を用いて、対象文節の第1、2、3モーラの母音の準定常部分を一度に一箇所だけ -50 ms から+50 ms の範囲で伸縮した．

刺激は防音室内の被験者の両耳にヘッドフォンを通して提示し、その際に対象文節に下線を引いた刺激文をディスプレイに表示した．被験者は下線部の長さと速さにおける不自然さの程度を7段階(1～7)で評定 した．１人の被験者は1個の刺激に対して合計4回評定した．日本語を母語とし正常な聴力をもつ成人7名が実験に参加した．

4.1.2 結果

被験者１人刺激１個当たりの4回の評定値の平均をとり、さらに文ごとに異なる評定値のバイアスを除くため、各評定値から文ごとの最大評定値を差し引いた．以後は、このバイアスを除いた評定値を許容度とする．単語単独発話において許容度と音韻長伸縮が2次曲線で近似できることがわかっているので、今回も同様に2次曲線で近似した．図2に文節“ 春が ”の第1(initial)、2(medial)、3(final)モーラ母音における音韻長伸縮に対する許容度をプロットし、2次曲線で近似した例を示す．この2次曲線の2次係数の絶対値は音韻長伸縮に対する許容度の低下の度合いを示し、以後これを許容度低下指標と呼ぶ．許容度低下指標が大きいほど音韻長伸縮に対して許容度が低下しやすい．この許容度低下指標を刺激文ごとに調査した結果、位置による違いが見られた．つまり音韻長伸縮値が大きくなるにつれて、文節頭、文節中、文節末の順に許容度低下の度合い(許容度低下指標の絶対値)が小さくなっていることが分かった．今回用いたほぼ全刺激文・文節において同様の傾向が見られた．

4.2 音韻長伸縮の許容度に対する発話速度の効果

4.2.1 刺激と実験手続き

評価対象の文節は0、1型アクセントのものを5個ずつ、キャリア文に挿入したものと単独のものを用意した．これらの文はプロのナレータにより3段階の速度(速い、普通、遅い)で読み上げられた．最も遅い場合の評価対象の文節時間長は、 最も速い場合の約2倍となっている．高品質音声分析変換合成法 STRAIGHT[14]を用いて、 各母音部(全部で180個)を一度に一箇所、-50 ms から+50 ms の範囲で10 ms 伸縮した．

刺激の呈示と評定方法は4.1.1の実験と同様に行われた.被験者は日本語を母語とし正常な聴力をもつ成人で、 キャリア文有りの刺激に対する実験は14

名、 文節単独の刺激に対する実験には14名の内の7名が参加した. 1つの刺激に対して各被験者当たり4回の評定を求めた.

4.2.2 結果

許容度低下指標を4.1.2と同様に求めた．被験者毎に各文節の伸縮音韻における許容度低下指標を求め、それらを発話速度および各文節内位置ごとに全被験者にわたって平均した結果を図3に示す．発話速度が速い程許容度低下の度合が大きくなる傾向が見られた．また各発話速度で、文節頭、 文節中、 文節末の順に許容度低下の度合(許容度低下指標)が大きいことが観察された．キャリア文有無による系統だった差は見られなかった.

4.3 音韻長の制御特性と知覚特性

たとえば同じ音韻(たとえば“a”)であっても常に同じ時間長で発話されるわけではなく、一般に発話毎に長さは変動する．変動要因は当該音韻が置かれた位置や前後の音韻の種類など文脈に起因するものと発話制御の精度の甘さに起因するものとの2つに大まかに分けられる．ここでは後者を対象とするため、前者の文脈による変動を極力除く必要がある．そこで、数量化Ⅰ類を用いて音韻毎文脈毎に長さの期待値を求め、これと実際に発話された音韻長との差を制御精度に起因する変動の推定値とした．(実測の音韻長)-(数量化Ⅰ類により算出した推定音韻長)の標準偏差を制御精度とする．

4.3.1 資料と推定方法

文音声データベースとして「ATR 日本語音声データベース B」を用いた．このデータベースは10名の(男性6名、女性4名)アナウンサーにより1人あたり503文の音韻バランスされた文が読み上げられ、 音韻境界ラベルの他に形態素境界、品詞、文節境界、アクセント句境界、アクセント型の情報が付与されている．

分析対象を母音とし、数量化Ⅰ類を用いて期待値を求めた．文脈要因は先行研究[11]を参考にし、これに文節情報を加えた．文脈要因は当該音韻の種類、近傍の音韻(前、後、2つ前、2つ後)、促音の位置(前、後)、当該音韻の区分(単語、文節、アクセント句、呼気段落、文)内位置、当該音韻を含む区分のモーラ数、アクセント核の有無、品詞とした．先行研究では区分内位置は区分頭(第1モーラ)、区分中(1モーラと最終モーラを除く2モーラ目以降)、区分末(最終モーラ)というカテゴリー化を行なっていたが、今回はモーラでカウントした値をそのまま用いた．分析は話者毎に行なった．重相関係数に関して先行研究と比較し、同じまたはそれ以上の値が得られたことから、今回のモデルの精度は妥当であると判断した.

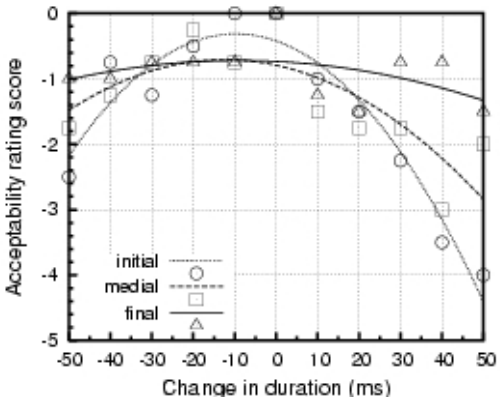

図2：音韻長伸縮に対する許容度低下.

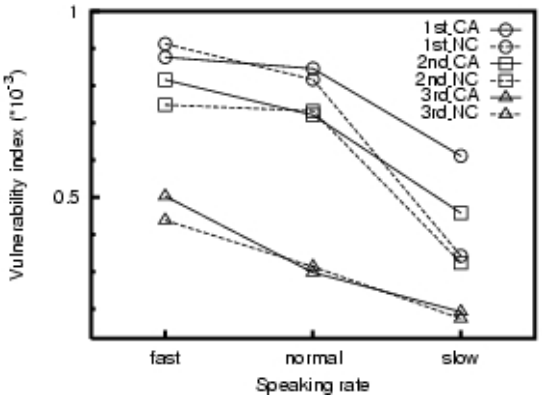

図3：許容度低下指標(Vulnerability index)の発話速度による違い.

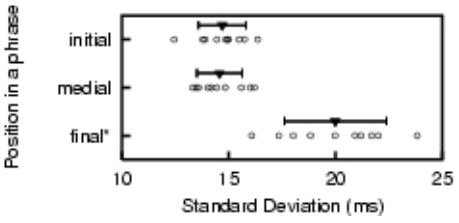

図4： 文節における全話者の音韻長制御精度の分布

4.3.2 音韻長制御精度推定

図4に文節における話者毎の制御精度の分布に示す．縦軸は各区分内の位置、 横軸は制御精度をあらわす標準偏差である．プロットは1話者あたりの変動の推定値の平均を示す．“initial”は区分内の1モーラ目、 “medial”は2モーラ目以降(1モーラと最終モーラを除く)、 “final”は最終モーラ目を示す．2モーラ目、 3モーラ目、 4モーラ目以降の間には有意差がなかったのでそれらを “medial”のカテゴリーにまとめた．図3をみると“initial”、“medial”の位置にくらべて“final”は全話者の標準偏差が大きくなっている．この

傾向は全区分について見られ、単語、 文節、 アクセント句、 呼気段落、 文のすべての区分において最終モーラの母音が他の位置の母音に比べて制御精度が悪いことがわかった．また、“ initial ”と“ medial ”の精度差は単語、 文節、 アクセント句ではほぼ同等で、 呼気段落では差が見られるものの有意差なし、 文では有意差ありという傾向が見られた．

文節内の音韻長伸縮に対する許容度の実験結果から、文節頭、文節中、文節末の順に許容度低下が大きいことがわかった．特に文節末の許容度低下度合いは他の位置に比べて顕著に低かった．このことから文節末においては、音韻長制御精度が悪いところでは音韻長伸縮に対する許容度低下が大きいという、 制御特性と知覚特性の対応が見られたと考えられる．

4.4 まとめ

合成音声の音韻長制御の評価と自然性向上を目的として、新たに文音声を対象として音韻長伸縮に対する許容度を調べた．その結果、文節頭、文節中、文節末の順に許容度低下が大きいことがわかり、この位置の効果は発話速度が変化した場合でも見られることがわかった．また、発話速度が速くなるにつれて、音韻長伸縮に対する許容度低下が大きくなることがわかった．さらに、音韻長伸縮に対する位置の効果が、発話時の音韻長制御特性と関連があることが示唆された．これらの知見は例えば、音韻長制御の際に重み付けをすることで合成音声の品質を向上させるなどの応用が期待できる．今後更なる時間長伸縮に対する知覚特性を調査することでこのような応用が可能になると考えられる．

5. 結論と今後の予定

以上述べたように、①F_0制御パラメータの自動抽出、②会話音声の発話語彙情報を用いたF_0制御特性のモデル化、③時間制御の自然性に関する聴知覚特性の測定を行った．①のF_0制御パラメータ自動抽出に対しては発話音声の内容を利用して推定精度を向上した．②の会話音声のF_0制御では副詞と形容詞からなる短発声を用い、使用される語彙の特性に基づく制御の存在を示し、聴覚的整合性、コーパスに基づく制御可能性を定量的に立証した．③時間制御の自然性については、合成音声の時間制御に対する自然性の測定を行い、聴知覚特性を反映したコンテキスト依存の制御誤差尺度の必要性を明らかにした．制御に関する研究成果は直接的に会話音声の韻律生成に使える．また、生成・知覚面からの知見を役立て、今後の大きな課題である、会話韻律により話者が取り交わしている言語コンテキスト外情報の本質的な理解、定量的制御モデル作成を進めたい．

参考文献

[1] H.Fujisaki, K.Hirose, “Analysis of voice fundamen- tal frequency contours for declarative sentences of Japanese,” J.Acoust.Soc. Japan(E), vol.5, No.4, pp. 233-242(1984).

[2] C. Hayashi, “On the Quantification of Qualitative Data from the Mathematico-Statistical Point of view,” Annals of the Institute of Statistical Mathematics, Vol. 2(1950).

[3] M. Miyatake, Y. Sagisaka, “Japanese speech database for Synthesis,” ATR Technical Report TR-I-0056(Nov. 1988).

[4] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara, “Speech Database,” ATR Technical Report TR-I-0166(Sep. 1990).

[5] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Japan (E), Vol.5, No.4, pp. 233-242(1984).

[6] H. Kato, M. Tsuzaki and Y. Sagisaka, “Acceptability for temporal modification of consecutive segments in isolated words,” J. Acoust. Soc. Am. 101, pp. 2311-232(1997).

[7] H. Kato, M. Tsuzaki and Y. Sagisaka, “Acceptability for temporal modification of single vowel segments in isolated words,” J. Acoust. Soc. Am. 104, pp. 540-549(1998).

[8] H. Kato, M. Tsuzaki and Y. Sagisaka, “Effects of phoneme class and duration on the acceptability of temporal modifications in speech,” J. Acoust. Soc. Am. 111, pp. 387-400(2002).

[9] 匂坂芳典, 東倉洋一, “規則による音声合成のための音韻時間長制御,” 電子通信学会論文誌, J67-A, pp. 629-636(1984).

[10] H. Kawahara, I. Masuda-Katsuse and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication 27, pp. 187-207(1999).

[11] 海木延佳, 匂坂芳典, “言語情報を利用した母音継続時間長の制御,” 電子通信学会論文誌, J75-A, pp. 467-473(1992).