2P1-5in1

会話による情報伝達における発話系列の韻律分析

Prosodic analysis for utterance sequence in spoken dialogue based information

福岡維新 *1 Ishin FUKUOKA　高津弘明 *1 Hiroaki TAKATSU　藤江真也 *2 Shinya FUJIE　林良彦 *1 Yoshihiko HAYASHI　小林哲則 *1 Tetsunori KOBAYASHI

*1早稲田大学 Waseda University　*2千葉工業大学 Chiba Institute of Technology

Prosody control for speech synthesis used in speech based information delivery is discussed. Conventional speech synthesis systems generate natural and clear speech of written sentences. In conversation for information delivery, however, a sentence is divided into several information units and they are synthesized one by one. To synthesize speech that a listener can easily understand, the voices of all the units should be harmonized as a sentence. This study focuses on prosodic information of such synthesized speech. Subjective evaluation is conducted to compare synthesized speech with normal prosody and speech with modified prosody where some of the parameters, such as F0 average, speech rate, are adjusted to human voice. The results show that the speech rate should be controlled to make a speech more understandable.

1. はじめに

人同士の対話を参考に合成音声の韻律を変化させ，会話における合成音声に必要な韻律制御の検討を行った．

近年，Siri や Pepper といった音声対話システムの普及に伴い，システムの音声にも人の発話と同様に意図を伝える [1]，局所的な強調を行う [2] といった表現力の向上が求められている．HMM 音声合成方式によるシステム音声の生成は多様な韻律表現を可能とするが，それらは文に対する制御を基本としている．一方で，我々が日常で行うような対話では，一つの発話が一文に相当するほど長くなることは少なく，いくつかの発話が複数連なってまとまった内容を伝達していることが多い．

我々は，ニュース記事に代表されるまとまった量の情報をユーザに伝達する音声対話システム [3,4] を指向している．このシステムはニュースの内容をユーザとのインタラクションを交えながら伝えるため，記事の一文に相当する内容であっても，会話のテンポを崩さない適切な単位の発話に分割して伝える必要がある．この適切な単位とは，佐藤 [5] が定義した「話節」とほぼ同義のものである．佐藤は「話節」を，「それが意図的（心理的）にしろ無意図的（生理的）にしろ，発話の途中の音声の途切れによってできる最小の節である」としており，本稿でも同様の定義を以って以降話節という用語を用いる．また，「発話」についても話節と同様の単位を指し，一文に相当する複数の発話を「文発話」と定義する．

日常会話では，文をそのまま読み上げるように話すことはなく，図 1 のように複数の話節が連なって内容が伝えられる．これによって会話におけるテンポが良くなり，話者間のインタラクションが増え，より円滑に情報伝達が実現できると考えられる．このようなシステム音声を生成しようとすると，二通りの方法が考えられる．一つ目は，全ての発話を繋げて生成し，話節の単位で音声を区切るという方法である．二つ目はあらかじめ話節単位で発話を生成し，それらをポーズを挟みながら繋げる方法である．

しかし，このような文発話では，聞き手が聞きやすいように各話節内，話節間の韻律のバランスが取れていなければならない．そのため一度に発話したものを区切る，あるいは独立に発話を生成するといった方法は適切ではなく，システムの発話の生成には，全体のバランスを踏まえた韻律の制御が求められる．一方，現在の合成音声においてこのような発話系列での韻律の制御は想定されていない．

そこで，本論文では人同士の対話を収録し，そこに現れる韻律特徴を合成音声に揃えることで，どのような韻律情報が話の聞きやすさに影響を与えるかを調査する．収集した人同士の対話を参考に，HMM 音声合成によって生成された発話の幾つかの韻律情報を，文節などの小さい単位で変化させる．作成した合成音声について印象評価実験を行った．

野生のコアラなんだけどさあ、オーストラリアの南東部にある、ビクトリア州ってとこで、数を減らす目的で、保護担当者の人が注射を使って、安楽死させてたんだって

図 1: 文発話の例．読点の箇所が発話が区切れる単位となる

2. 関連研究

対話においてパラ言語・韻律情報が果たす役割の重要性は様々な研究によって述べられており [6–8]，特にナレーションのような読み上げ口調と，日常会話のような対話口調では，その韻律情報が大きく異なっているとされている．

対話システムにおける音声合成の実現を目的とした韻律の研究も盛んになされている．中島ら [9] は読み上げ口調と対話口調における基本周波数 (F_0) の構成成分を utterance 成分，phrase 成分，local 成分の三つに分け，分析を行っている．分析の結果， utterance 成分・phrase 成分では F_0 平均・範囲共に読み上げ口調と比べて差があることを確認し，対話における音声合成ではこれらの値を予測する機能が必要であると述べている．

また，Campbell [10] らは様々な状況での人同士の対話を収集し，公共の場と私的な場の対話では，正規化振幅係数や基本周波数の値に差が見られることを確認した．話者同士の関係性が発話の音響的特徴に影響を与えることから，システムがユー

連絡先: 早稲田大学基幹理工学研究科，東京都新宿区喜久井町 27，fukuoka@pcl.cs.waseda.ac.jp

ザと気軽な対話を行うためには，より細やかな韻律情報の制御が必要であると述べている．

伊藤ら [11] は HMM 音声合成の韻律制御の精度改善を目的とし，韻律推定に利用する制御要因の数と種類の違いが「話し言葉としての自然性」に与える影響について検討を行い，当該／後続アクセント句間のポーズ長や，アクセント句ごとの基本周波数 (F_0) の平均値のばらつきなどが「話し言葉らしさ」に大きな影響を与えることを示している．

このように対話音声合成を指向した韻律分析の研究は多く行われているが，いずれも発話系列といった観点からの問題は扱っていない．

3. 音声試料の作成

3.1 情報伝達対話の収集

手本となる文発話の音声を用意するにあたり，人同士の音声対話の収録を行った．対話はシステム役とユーザ役の二人からなる．タスクはシステム役による，ユーザ役へのニュース記事の内容の伝達とした．

通常，書き言葉と話し言葉には大きな乖離が存在する．図 2 のように，書き言葉における「宮間や大儀見らが先発した日本」といった連帯修飾の形式は,「日本は宮間や大儀見らが先発した」という平文に直されて発話される．このような口語化の技術 [12] は，対話による情報伝達を実現する上で非常に重要である．そのため対話収録においてもシステム役にはニュース記事を読み，内容を理解した上でどのように口語化すると自然に内容を伝えることができるか意識して対話に臨むように指示を出した．

宮間や大儀見らが先発した日本は開始直後に先制された →「日本は宮間や大儀見が先発したんだけど，開始直後に先制されたんだって」

図 2: ニュース記事の一文と口語化の例

3.2 手本となる発話の収録

本研究では，人発話内容のテキスト情報から合成音声を作成し，その韻律特徴を人の韻律特徴に合わせ込むことで音声試料を作成する．

3.1 の収録では，記事に書かれている文章にしばられることなく，日常会話のような対話が収集された．しかし，日常会話に近づけば近づくほど人の発話には言い淀みや言い間違いが生じてしまう．今回収録した対話でも同様の現象が見られたため，その発話内容を利用して再度合成音声の手本となる発話を収録した．**3.1** の対話の書き起こしからフィラーや言い間違いといった要素を除いた原稿を作成し，その原稿に基づいて再度対話口調での発話データを収録した．このようにしてニュースの伝達を行う文発話を収録し，これらを合成音声が真似る手本とした．

3.3 合成音声の韻律パラメータの変更

HMM 音声合成によって，**3.2** で収録したシステム役の文発話と同じ内容の合成音声を生成する．合成器は，HMM によるパラメータ生成に HTS_engine [13] を利用し，合成には STRAIGHT [14] を用いている．

生成された合成音声に対して，以下に示す三つの韻律特徴を人の発話の韻律特徴に合わせた音声試料を用意する．また，これらの特徴の変化を同時に組み合わせた音声も作成した．

今後これらの韻律特徴を制御することを想定しているため，各韻律特徴は人の発話に完全に一致させるのではなくある程度制御可能な単位で変化させる．そのため，基本周波数は発話が区切れやすい文節毎に変更を行う．発話速度は自立語付属語の単位で変化させる．発話速度を自立語付属語の単位で変更するのは，人の発話は文節よりも小さい単位で速さが変化することが多いためである．例えば，継続を表す発話「日本代表がー」では，付属語である「が」のみが引き伸ばされ，自立語である「日本代表」はあまり変化しない．

3.3.1 文節毎の F_0 平均の変化

対話口調と読み上げ口調においては，強調したい単語や話節の末尾における継続などで，F_0 の高さは大きく変化する．そこで，合成音声の各文節の F_0 の平均値の変化が，人の発話における文節毎の F_0 の平均値の変化と同様の軌跡を描くように F_0 の値を変更する．

文節の認定には KNP[*1] を用いた．また F_0 の抽出には Wavesurfer[*2] の基本周波数抽出機能を用いた．各文節毎に新たな F_0 の平均値を求め，文節区間の F_0 系列が新たな平均値に合うように変更する．ある文節区間における合成音声の i サンプル目の F_0 値 p'_i は以下の式（1）によって求まる．ここで，μ，μ' は合成音声の文節区間における元の F_0 平均値，新しい F_0 平均値である．$\mu^{(h)}$ は人の発話における文節区間の F_0 平均値，$\boldsymbol{\mu}^{(h)}$,$\boldsymbol{\mu}$ はそれぞれ人，合成音声の文発話全体の F_0 の平均値である．元の合成音声の F_0 の軌跡を保ったまま，平均が μ' になるように文節全体の値を変化させ，STRAIGHT による再合成を行う．F_0 を変化させる際，値が極端に大きくなる，小さくなることによって音声そのものが壊れてしまわないよう上下限値を設定している．

$$\begin{aligned} p'_i &= p_i + (\mu' - \mu) \\ \mu' &= (\mu^{(h)} - \boldsymbol{\mu}^{(h)}) + \boldsymbol{\mu} \end{aligned} \tag{1}$$

3.3.2 文節毎の F_0 の分散

人の発話では，強調したい単語の F_0 の分散は大きく，接続詞や発話の末尾では分散が小さくなるといった特徴が見られる．このような特徴を合成音声に反映させるため，合成音声における文節毎の F_0 の分散値が，人の音声の F_0 の分散値に合うよう，元の合成音声の F_0 を変化させる．ある文節区間の F_0 系列における i 番目の $F_0 p'_i$ は以下の式（2）で新しい F_0 へと変換される．$\sigma^{(h)}$，σ はそれぞれ，人，合成音声の該当文節区間の F_0 の分散値である．μ は合成音声の文節の F_0 平均値である．**3.3.1** 同様 F_0 の値を文節毎に全て変更し，再合成を行う．

$$\begin{aligned} p'_i &= (p_i - \mu)k + \mu \\ k &= \sqrt{\sigma^{(h)}/\sigma} \end{aligned} \tag{2}$$

3.3.3 自立語付属語単位での話速

人の発話は，話している内容や話し相手とのインタラクションによって発話速度が大きく変化し，これによって会話特有のテンポを実現していると考えられる．そこで，合成音声における自立語・付属語ごとの話速を，対応する人の音声の話速になるよう音素継続長を変更する．

自立語，付属語の認定には **3.3.1** と同様に KNP を用いた．話速の定義は (モーラ数/継続時間) としている．実際にはモー

*1 日本語構文・格・照応解析システム KNP version 4.11, http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

*2 http://www.speech.kth.se/wavesurfer/

ラ数の同じ区間を比較するため，それぞれの区間における人の音声と合成音声の継続時間の比 r を計算し，合成音の母音部分の時間にその比率をかけた数値を新たな時間とした．しかし，人同士の自然な対話における話速は合成音と比較すると 2 倍以上速い箇所もあり，合成音声の話速を人と全く同じ速さにしてしまうと被験者が発話内容を聞き取れない可能性があった．そのため，人の発話における自立語・付属語毎の話速のバランスを保ったまま，話速が速くなっている箇所では変化が緩やかになるよう調整を行った．継続時間の比 r は以下の式（3）で表される．ここで D_h は一定区間における人の音声の継続時間，D_s は一定区間における合成音声の継続時間である．

$$r = \begin{cases} \frac{D_h}{D_s} & (D_h \geq D_s) \\ \frac{1}{2}(1 + \frac{D_h}{D_s}) & (D_s > D_h) \end{cases} \quad (3)$$

3.3.4 パラメータの組み合わせ

3.3.1 から **3.3.3** までのパラメータの変更を組み合わせた合成音声についても作成した．話速を変えた音声と他のパラメータ（F_0 の平均値，分散値）を組み合わせる場合には，話速変更後の合成音声と，人の発話音声を比較して上記の方法で変更を行う．F_0 の平均値，分散値を組み合わせる場合には，先に平均値を変化させ（この際に分散値は変わらない），その後に分散値の変更を行う．平均値と分散値を加算的に変更することによって F_0 の値が極端に大きくなるような場合には，**3.3.1**，**3.3.2** と同様に上下限値を設けてそれ以上は F_0 の値が変化しないようにしている．

4. 主観評価実験

3.3 で作成した 3 種類の韻律特徴，それらを組み合わせたパラメータの変更を行った合成音声を用いて，五つの一対評価実験を行った．変更させるパラメータの組み合わせを全て一対比較実験で評価しようとすると，膨大な数になってしまう．また，一度に実験を行うとどのような韻律特徴が効果的であるかが不明確になってしまう恐れがあるため，実験は複数に段階を分けて実施した．音声資料には **3.2** で収録した発話から 4 個の文発話を選択し，これらの韻律特徴を変更した合成音声を作成して実験に用いた．被験者は 20 代の男女である．いずれの実験でも，各被験者は順序効果を考慮して比較するパラメータの全ての対比較評価を行う．各音声対を比較して,「よりシステムの話を聞き続けようと思う」音声を選択させた．

4.1 実験の流れ

5 種類の実験を行う．実験 1 では **3.3** で述べた三つのパラメータ，F_0 平均，F_0 分散，発話速度をそれぞれ変更した 3 種類の合成音声について一対比較実験を行う．実験 2 では，実験 1 にて最も良いとされたパラメータ（X とする）を変更した音声，残り二つのパラメータ（Y，Z）を組み合わせた $X+Y$，$X+Z$ を用いて対比較実験を行う．実験 3 では，実験 2 で良いとされたパラメタ（$X+Y$ とする）と，三つ全てを組み合わせたパラメタ（$X+Y+Z$）の比較を行う．

実験 4 ではパラメータを変更しない通常の合成音声（N とする）と実験 1 で最も良かったパラメータ X を変更した音声の比較を行う．実験 5 では実験 1 で選ばれた X と実験 2 で選ばれた $X+Y$ の比較を行う．

4.2 実験 1

F_0 平均（M），F_0 分散（V），発話速度（S）をそれぞれ変更した 3 種類の合成音声について一対比較実験を行った．3 種類の音声に対して順序効果を考慮した音声対を作成するため，一つの音声につき 6 組の比較対ができる．文発話は 4 個あるため合計 24 組の比較対ができる．これらについて，20 代の男女 8 名で一対比較実験を行った．結果を表 1 に示す．

表 1: 実験 1 音声の選択率

	M	V	S
M	-	45.3%	21.9%
V	54.7%	-	25.0%
S	78.1%	75.0%	-

表 2: 実験 1 各パラメータの心理尺度上の評価点

	M	V	S
評価点	-2.333	-1.444	3.778

セル [M, V] の 45.3% という数字は「M と V を比較して，M の方を選択した割合」となる． 表から，選択率を比較すると $S > M$, $S > V$ であることがわかる．これらの差について，シェッフェの一対比較法（浦の変法）を用いて算出された心理尺度上の評価点を表 2 に示す．同様に算出したヤードスティックは $Y(0.05) = 0.970$，$Y(0.01) = 1.253$ である．M-S 間，V-S 間の差はこの $Y(0.01)$ よりも開いているため，M（F_0 平均）と S（話速），V（F_0 分散）と S（話速）の間には危険率 1%以下で有意な差があると言える．

4.3 実験 2

4.2 で最も良いとされた話速を変更させた合成音声に対して，F_0 平均，F_0 分散をそれぞれ変更させた 2 種類の音声を作成し，実験 1 と同様に一対比較実験を行った．被験者の人数は 8 人である．文発話の数は前回と同様，一つの音声につき 2 組の比較対ができるため，被験者一人あたりが評価する音声の比較対は 8 組となる．結果を表 3 に示す． M + S が F_0 平均（M）と話速（S）を変更した音声， V + S が F_0 分散（V）と話速（S））を変更した音声である．$M+S > V+S$ となっているが，この結果について，音声の選択数を用いて片側符号検定を行ったところ有意な差は見られなかった．

4.4 実験 3

4.3 で良いとされた合成音声 M + S と，F_0 平均，F_0 分散，話速を全て変更した音声の 2 種類の合成音声について，同様に一対比較実験を行った．実験条件は **4.3** と同様である．結果を表 4 に示す．$M+S > M+V+S$ となっており，片側符号検定において危険率 0.5%で有意な差が見られた．

4.5 実験 4

4.2 で最も良いとされた合成音声 S と，標準の合成音声 N の二つの音声について，同様に一対比較実験を行った．実験条件は **4.3** と同様である．結果を表 5 に示す．$S > N$ となっ

表 3: 実験 2 音声の選択率

	M + S	V + S
M + S	-	53.1%
V + S	46.9%	-

表 4: 実験 3 音声の選択率

	M + S	M + V + S
M + S	-	68.8%
M + V + S	31.2%	-

表 5: 実験 4 音声の選択率

	S	N
S	-	68.8%
N	31.2%	-

表 6: 実験 5 音声の選択率

	S	M + S
S	-	79.7%
M + S	20.3%	-

ており，片側符号検定において危険率 0.5%で有意な差が見られた．

4.6 実験 5

4.2 で最も良いとされた話速を変更させた合成音声と，**4.3** で良いとされた F_0 平均，話速を変更させた合成音声の 2 種類の音声を作成し，同様に一対比較実験を行った．実験条件は **4.3** と同様である．結果を表 6 に示す．$S > M + S$ となっており，この結果について音声の選択数を用いて片側符号検定を行ったところ危険率 0.5%で有意な差が見られた．

5. まとめ

情報伝達対話のための発話系列の韻律制御に有効な要素の調査を目的として，実際に収録した人同士の対話音声の韻律情報を HMM 合成音声の韻律情報に当てはめ，主観評価実験によってその有効性を検証した．韻律情報として発話におけるピッチ平均の変化やピッチの分散の変化，発話速度変化などを変更し，各音声を一対比較によって評価したところ，発話速度を変えることが複数の話節から成る発話系列において有効である可能性が示唆された．まとまった量の情報を一度に伝える際には，どこを丁寧に伝えてどこを軽く伝えるのかといった情報の粒度を調整する必要があり，話速の変更はその一要素であると考えられる．また，**4.** で述べた実験結果は，4 個の文発話の評価をまとめたものとなっており，各韻律情報の変化の度合いやバランスというのは文毎に異なっている．F_0 平均値や分散値の変更した音声については，文発話毎に見るとその有効性が示唆されている音声も存在する一方で，直接合成音声の F_0 値を指定したことによってピッチの軌跡に合成器が対応できていない箇所もあった．今後は，評価に効いている要素をより詳細に考察していくとともに，人の発話におけるピッチのバランスを保ちつつ滑らかな F_0 の軌跡を生成できるような仕組みを考案する必要がある．

参考文献

[1] Kazuhiko Iwata and Tetsunori Kobayashi. Expression of speaker's intentions through sentence-final particle/intonation combinations in japanese conversational speech synthesis. In *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[2] Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, and Osamu Yoshioka. Prosodic variation enhancement using unsupervised context labeling for hmm-based expressive speech synthesis. *Speech Communication*, Vol. 57, pp. 144–154, 2014.

[3] Shinya Fujie, Ishin Fukuoka, Asumi Mugita, Hiroaki Takatsu, Yoshihiko Hayashi, and Tetsunori Kobayashi. A spoken dialog system for coordinating information consumption and exploration. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pp. 253–256, New York, NY, USA, 2016. ACM.

[4] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則. 快適な情 報享受を可能とする音声対話システム. 言語処理学会第 22 回年次大会発表論文集, 2016.

[5] 佐藤虎男. 音節の持続時間と文法. 1991.

[6] 前川喜久雄, 北川智利. 言語コミュニケーションの科学に向けて音声はパラ言語情報をいかに伝えるか. 認知科学, Vol. 9, No. 1, pp. 46–66, 2002.

[7] Shinya Fujie, Yasushi Ejiri, Yosuke Matsusaka, Hideaki Kikuchi, and Tetsunori Kobayashi. Recognition of paralinguistic information and its application to spoken dialogue system. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 231–236. IEEE, 2003.

[8] 大須賀智子, 堀内靖雄, 西田昌史, 市川熹. 音声対話での話者交替/継続の予測における韻律情報の有効性. 人工知能学会論文誌, Vol. 21, pp. 1–8, 2006.

[9] 中嶋秀治, 匂坂芳典. 対話音声合成を目指した対話音声の韻律分析. GITS, GITI 紀要, Vol. 2008, pp. 134–139, 2008.

[10] Nick Campbell and Ya Li. Expressivity in interactive speech synthesis; some paralinguistic and nonlinguistic issues of speech prosody for conversational dialogue systems. In *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, pp. 97–107. Springer, 2015.

[11] 伊藤芳幸, 岩野公司, 古井貞熙ほか. 話し言葉音声合成の韻律制御に関する検討. 研究報告音声言語情報処理 (SLP), Vol. 2009, No. 23, pp. 1–8, 2009.

[12] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則. 会話によるニュース記事伝達のための口語化における述語の書き換え. 言語処理学会第 22 回年次大会発表論文集, 2016.

[13] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. The hmm-based speech synthesis system (hts) version 2.0. In *SSW*, pp. 294–299. Citeseer, 2007.

[14] 河原英紀. Vocoder のもう一つの可能性を探る: 音声分析変換合成システム straight の背景と展開. 日本音響学会誌, Vol. 63, No. 8, pp. 442–449, 2007.