2—Q—6

波形接続型音声合成における F_0 の傾きを考慮した接続コスト*

◎藤井 慶 (NAIST)　△柏岡秀紀 (NAIST/ATR)　ニック・キャンベル (NAIST/ATR/CREST)

1 はじめに

波形接続型音声合成では, 接続する音声単位間の接続点付近で不連続が生じ得る. この不連続には, ホルマントや声質といったスペクトル包絡上での不連続や韻律面での不連続がある. この問題への対策として, 不連続の起きにくい音声単位の検討, 単位選択部における接続コストの改良, 波形接続時における波形変形等が挙げられる.

著者らは韻律面での不連続のうち, F_0 の不連続をより効率良く抑制することを目的とし, 単位選択部における接続コストの改良を試みている. 本稿では接続点における F_0 の傾きのずれをコスト化する手法を提案し, その評価を行う.

2 従来法とその問題点

2.1 従来の F_0 接続コスト

i 番目音声単位 u_i とそれに後続する音声単位 u_{i+1} の, 接続点での F_0 のずれに関する従来コストは次式で表される [1].

$$C_{conv}(u_i, u_{i+1}) = |F_{0end}(u_i) - F_{0start}(u_{i+1})| \quad (1)$$

ここで $F_{0end}(u_i)$, $F_{0start}(u_{i+1})$ は各々の候補の終端の F_0 値と始端の F_0 値を表すものとする. なお本稿では, F_0 はメルスケール上で取り扱う.

2.2 従来法の問題点

従来法の問題点としては, 接続点における F_0 の傾きのずれからくる劣化を考慮していないことが挙げられる. 単純な例として図 1 のような F_0 曲線を持つ単位の組み合わせを挙げる. この場合接続点付近で品質劣化が予想されるが, (1) 式は値 0 をとるため, この劣化に対応出来ない.

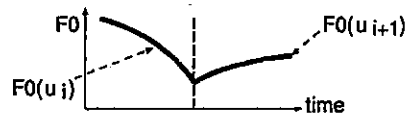

図 1: 従来の F_0 接続コストでカバー出来ない例

この例のような場合の単位列の選択は, 従来の F_0 ターゲットコスト [1] や F_0 傾きターゲットコスト [2] によって間接的に抑制されると考えられるが, これらのコストはあくまで合成音を予測ターゲットに近づけるためのものであり, この問題に関して十分とは限らない. なぜならば [3][4] で示されたこと (ある予測ターゲットに最も近い単位列が必ずしも最適ではなく, 若干離れた単位列の方がより自然な場合があり得るという主張) と同様の状況がこの場合にも起き得ると考えられるためである. 例として図 2 を挙げる. この場合 (u_i, u_b) の組み合わせの方が滑らかと考えられるが, 従来の F_0 ターゲットコストおよび F_0 傾きターゲットコストでは u_a の方がより低いコスト値を取るため, (u_i, u_a) が選ばれる可能性がある.

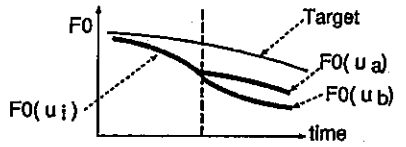

図 2: 従来のコストでカバー出来ない例

以上のことから次節以降で F_0 傾きの接続コストを提案し, その評価を行う.

3 F_0 傾き接続コスト

まず提案するコスト計算に必要な特徴抽出について説明し, 次に 2 種類のコスト計算法 (従来コスト ((1) 式) と独立に F_0 傾きの差をコスト化する手法と, 従来コストを考慮して F_0 傾き差をコスト化する手法) を説明する.

3.1 特徴抽出

提案法のコスト計算を行う際には各音声単位境界の F_0 の傾きが必要となる. ここで, 単位境界近傍に無声子音が存在する場合, 境界近傍の F_0 はマイクロプロソディの影響により値が不安定になりやすい. そこでまず成澤らの手法 [5] を適用してマイクロプロソディ除去, 補間, 平滑化を行った F_0 を用意する. その上で各単位境界を中心とする T $msec$ (本稿では $T = 100$) の F_0 を 3 次多項式で近似する ((2) 式).

$$F_{0sm}(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \quad (2)$$

ここで回帰する際には便宜上, 各々の単位境界上で時刻 t を 0 にする (よって $-T/2 \leq t \leq T/2$). (2) 式を微分し $t = 0$ を代入すると単位境界上の F_0 の傾き a_1 が得られる. 以降では u_i の始端および終端の F_0 傾きをそれぞれ $S_{start}(u_i), S_{end}(u_i)$ とおく.

3.2 コスト計算

ここでは提案コスト計算方法を 2 つ説明する.

3.2.1 提案コスト計算 1

*The concatenation cost considerating F0 change in concatenative speech synthesis. By Kei Fujii(NAIST), Hideki KASHIOKA(NAIST/ATR) and Nick Campbell(NAIST/ATR/CREST)

一つ目の提案法は, 接続点上における F_0 傾きの差をコストとすることであり, 次式で求められる.

$$C_{prop1}(u_i, u_{i+1}) = |S_{end}(u_i) - S_{start}(u_{i+1})| \quad (3)$$

3.2.2 提案コスト計算 2

一つ目の提案法は F_0 値の差に関わらず傾きの差をコストとしたが, ここでは F_0 差に応じたコスト計算を考える. これは, F_0 傾きのずれからくる劣化は F_0 差がある程度の範囲内に収まる場合に生じると予想したためである. そこで F_0 差が小さい場合は傾き差に沿った値を取り, F_0 差が広がるほど傾き差に依らず一定値に近付くようなコスト関数を試作した. 計算式を式 (4) に, 概形を図 3 に示す.

$$C_{prop2}(u_i, u_{i+1}) = -1 + 2/\{1 + \exp(-w_c C_{conv}(u_i, u_{i+1}) - w_p C_{prop1}(u_i, u_{i+1}))\} \quad (4)$$

ここで w_c, w_p はそれぞれ (1) 式, (3) 式の影響の度合を調整する定数であり, 本稿ではいずれも 0.1 とした. 式 (4) および図 3 より, F_0 差が大きい場合は一定値に近付くことが分かる.

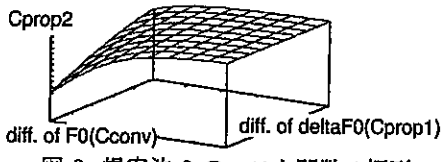

図 3: 提案法 2 のコスト関数の概形

4 評価実験

前節で提案したコストの導入により合成音の品質が向上するかどうかを評価するため, 聴取実験を行った.

4.1 実験条件

まず実験試料として 20 文の合成音を作成した. 用いたコーパスは男性話者による ATR503 文の朗読音声であり, 合成する文の発話の各音声単位をコーパスから外した上で, その発話の音韻および韻律をターゲットとし, 手法毎に単位選択を行った. ここで従来法で用いたコストは音韻環境, ターゲットコスト : (F_0, F_0 傾き [2], デュレーション, パワー), 接続コスト : (F_0((1) 式), パワー, MFCC) であり, 戸田らの提案する rms コスト [6] 化したものである. また基本音声単位には半音素を用いた. この従来法に (3) 式のコストを加えたものを提案法 1, (4) 式のコストを加えたものを提案法 2 とする.

次に, 従来法と提案法 1 の合成音を対にして比較する聴取実験 1, 従来法と提案法 2 の合成音を対にして比較する聴取実験 2 を行った. 被験者数はそれぞれ 5 名, 3 名である. 各被験者には, 2 種の合成音を防音室でヘッドホン受聴して, より人の発声に近い合成音を選んでもらった. その際に違いが分からない場合, 優劣を判断しかねる場合は "優劣なし" を選択出来るものとした. 各合成音の順序は被験者毎にランダムに入れ替えており, 判定を下すまで任意に聞き直せるものとした.

表 1: 実験 1 の結果

手法	被験者毎の選択率 [%]					
	A	B	C	D	E	全体
従来法	20	25	10	20	30	21
提案法 1	40	25	30	30	45	34
優劣なし	40	50	60	50	25	45

表 2: 実験 2 の結果

手法	被験者毎の選択率 [%]			
	A	B	C	全体
従来法	20	30	30	26.7
提案法 2	35	30	45	36.7
優劣なし	45	40	25	36.7

4.2 実験結果

各実験の結果を表 1, 2 に示す. 表より, いずれの実験でも提案法が従来法より多く選択されたことが分かる. また被験者毎に見ても, 提案法の選択率は従来法以上であった.

被験者の感想を聞くと, 手法毎の差はあまり大きくないとのことであり, 筆者も同様の印象を持っていた. そのため今回の評価では, MOS のような絶対評価ではなくプリファレンススコアを用いた. 今回の結果は, 提案法が合成音質を向上させることを示すものであるが, その向上の度合については未知数である. また 2 つの提案法の比較は今回行っていない. 今回の結果から両者の差位は小さいと推測されるが, この確認は今後の課題としたい.

5 まとめ

本稿では, 波形接続型音声合成における接続点での F_0 の傾きのずれに起因する合成音の劣化に着目し, 単位選択部に新たな接続コストを導入することを提案した. そして聴取実験を行い, 従来法に比べ提案法がより品質が高いとの結果を得た. 今後の課題としては品質向上の度合を調査することが挙げられる.

謝辞 本研究の一部は科学技術振興事業団戦略的基礎研究推進事業 (JST/CREST) の援助により行われた.

参考文献

[1] ニック・キャンベル, アラン・ブラック, 信学技報, SP96-7, pp.45-52 (1996/5).
[2] 藤澤謙, 平井俊男, 樋口宣男, 音講論, 2-7-2 (1997/3).
[3] T.Hirai, S.Tenpaku, K.Shikano, Proc. IEEE 2002 Workshop on Speech Synthesis (2002/9).
[4] 藤井慶, ニック・キャンベル, 音講論, 1-P-18 (2002/3).
[5] 成澤修一, 峯松信明, 広瀬啓吉, 藤崎博也, 情報処理学会論文誌, Vol.43, No.7, pp.2155-2168 (2002/7).
[6] 戸田智基, 河井恒, 津崎実, 鹿野清宏, 信学技報, SP2002-69, pp.19-24 (2002/8).