単位選択における音素境界の不連続感の検出及び音声合成への応用*

⊙丁　文、ニック キャンベル (ATR 音声翻訳通信研究所)

1　はじめに

音素単位の波形接続型合成においては単位接続部の不連続感がしばしば感じられる。我々は以前の研究 [1] で信号処理に基づいて合成音声の韻律制御への試みを行った。そこで、信号処理を最小限に用いるため、知覚される不連続部を自動的に検出し、信号処理を適応する方針を示した。また、合成音声に韻律部分の不自然感と接続歪みによる不連続感があると思われる。今回は後者に対しての予備実験として、日本語母音の接続歪みに関する聴覚実験を行い、聴覚スコアは triphone コンテキスト及び音響パラメータとの関係について調べ、不連続感に対する有効な検出特徴量について検討してきた。また、音響パラメータを用いて不連続感の主観尺度 MOS を予測する回帰木モデルを構築し、これらの結果について報告する。

2　実験資料

波形接続型合成では母音連鎖の部分の接続歪みによる不連続感が起こりやすいと考えられる。このような不連続感の主観尺度とコンテキスト、母音接続部の音響パラメータとの関係を調べるために、聴覚実験を行った。

単位選択は triphone コンテキストに基づいて行うため、まず二つの triphone 接続を対象として実験データを作成した。例えば、図 1 に示すように、母音結合を含む目的音声 /s-a-i-d/ は、1 つ目の triphone の前 2 つの音素と 2 つ目の triphone の後ろ 2 つの音素を接続して作成したものである。そこで、2 番目の triphone の先行母音の種類によって、/s-a:i+d/ の接続歪みが違う。この場合、当該音素と先行母音の種類及び接続点の音響的特徴に付与する接続歪みと聴覚スコアの関係は不連続感の検出や単位選択に対しては重要である。

今回の実験では ATR データベース男性話者 MHT の 503 文の音声を使用した、2 番目 triphone の先行母音の種類を考慮したすべての母音結合を調べるのは困難である。今回は、母音結合のフォルマント遷移パターン及びデータベースにある triphone コンテキストの数を考慮した上、母音結合を含む目的音声を /x-a-a-y/、/x-a-i-y/, /x-a-e-y/, /x-i-a-y/, /x-i-i-y/, /x-o-i-y/, /x-u-e-y/, /x-u-u-y/ とした。ここで、x, y は他の音素である。

これらの目的音声は図 1 のようにデータベース中の二つの triphone を使って合成したものである。例えば、目的音声 /s-a-i-d/ の場合、データベースにあるすべての s-a+i の候補と x-i+d の候補を使い、/s-a:i+d/ のようにすべての組み合わせの音声サンプルを作成した。ここの x は a,i,u,e,o のいずれかを表す。二つの音声波形を接続部のピッチマークに合わせて接続した。ほかの目的音声も同様に作成した。各種のサンプルから総数 120 のサンプルを実験データとした。

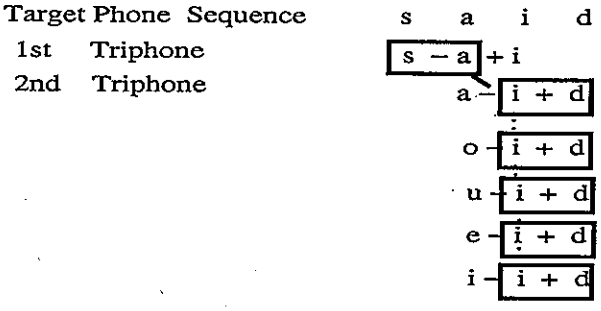

Figure 1: Description of triphone concatenation.

音声信号はサンプリング周波数 12kHz、16-bit である。防音室でヘッドフォンによる聴取実験を行い、5 人の健聴者から 120 サンプルの接続部の不連続感に対する MOS 値を収集した。

3　特徴ファクター

接続母音 $V^{(1)}$:$V^{(2)}$ に対して、接続境界において $V^{(1)}$ と $V^{(2)}$ の $F0$ の差、パワーの差、ケプストラム距離 CD、デルタケプストラムの距離 dCD を計算した。

ケプストラムは 14 次の LPC 係数から計算した 20 次ケプストラムの最初の 13 次の値である。$F0$ の差, パワーの差はマスキング効果を考慮して絶対値を取らずに、$F0.D = logF0^{(1)} - logF0^{(2)}$, $PWR.D = logPWR^{(1)} - logPWR^{(2)}$ である。ここで、$F0.D, PWR.D$ の値は負になる可能性がある。

4　実験結果及び考察

4.1　先行母音の種類と MOS の関係

評価音声に対する被験者の判断基準が個人によって違うことが十分考えられる。そこで、被験者の判断した MOS 値の平均を揃えるべく。5 人被験者毎の MOS 値をそれぞれの z-score で正規化し、正規化した値の平均値を最終的な MOS 値とした。

2 番目の triphone の先行母音と MOS 値の関係を表 1 に示す。目的音声 /s-a-i-d/ の場合、triphone は s-a+i と x-i+d である。ここで、先行母音 x は表中の行に示されている母音である。表より, 母音の調音結合を合成するとき、2 番目の候補 triphone の先行母音は 1 番目の triphone の当該母音の調音位置や調音様式と近い場合、接続歪みが小さかった。文献 [2] では、合成用のすべての音素の発音位置や発音様式などを幾つかのパターンを定義して単位選択に用いている。

*Detection of perceptual discontinuity between phoneme boundaries and its application to unit selection in speech synthesis
By Wen Ding and Nick Campbell, (ATR Interpreting Telecommunications Research Labs.)

Table 1: Relationship between vowel categories and MOS(row represents the preceding vowel of 2nd Triphone, column represents the current vowel of 1st triphone as shown in Fig. 1.

	a	i	u	e	o
a	0.13	-0.43	-0.17	-0.19	0.16
i	-0.32	0.35	0.02	0.38	-0.44
u	-1.18	-0.22	0.03	-0.23	-0.41
o	0.43	-0.39	0.91	0.09	0.82

また、MOS 値とそれぞれの特徴ファクターとの相関は以下のようである：

	CD	$F0.D$	$PWR.D$	dCD
MOS	-0.42	-0.18	-0.02	0.09

各々のファクターと MOS との相関が低いことから、不連続感の知覚に関して単独な特徴ファクターで判断を行うことは困難であると言える。

4.2 回帰木モデル

回帰木を使って MOS 値を予測するモデルを構築し、それぞれの予測ファクターの重要性について調べた。

回帰木の構築にサンプルの outliers の影響を取り除くため、MOS 値の標準偏差が 1.1 より大きい 25 個のサンプルを取り除いた 95 個のサンプルを用いた。構成した回帰木の特性は以下である。

式：MOS $\sim (CD + F0.D + PWR.D)$

ノード数: 15

Residual mean deviance (RMD): 0.25 (19.99 / 80)

残差の分布:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.32	-0.26	0.01	0	0.29	1.18

予測残差の分布はガウス分布に近い。この回帰木は prune 又は shrink の処理を行っていないものである。

回帰木の全体を図 2 に示す。ここで、縦方向の間隔は各ノード分割の重要度に対応している。接続点における各特徴ファクターの重要性も示されている。また、回帰木モデルの予測ファクターに dCD を加えても僅かの効果しか得られなかったため、モデルに考慮しなかった。

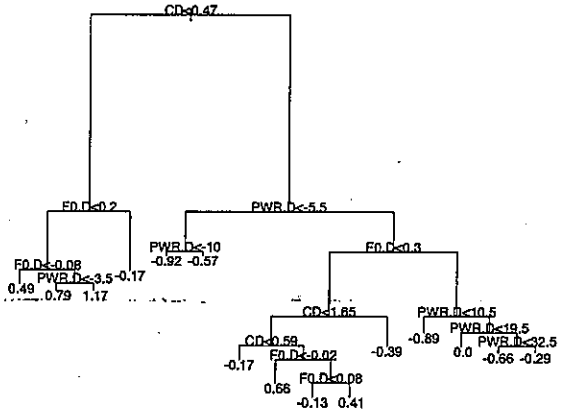

Figure 2: Dendrogram of the regression tree for prediction of MOS (node value).

次に、この回帰木により学習データを使った時の不連続感の予測値と MOS 値の関係を図 3 に示す。相関係数は 0.79 であった。

さらに、95 個のデータから 10 個をテストデータとして、残りの 85 個をトレーリングデータとして回帰木モデルを作成した。このように 9 回繰り返し、9 個の回帰木モデルとそれぞれに対応するテストデータを作成し、オープンテストを行った。その結果を図 4 に示す。相関係数は 0.58 であった。図 3、4 ともに、予測誤差の分布は零付近に多いことより、比較的安定な予測であると言える。

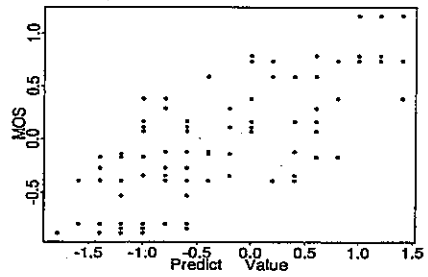

Figure 3: Relationship between MOS and the predicted value by the regression tree for the train data.

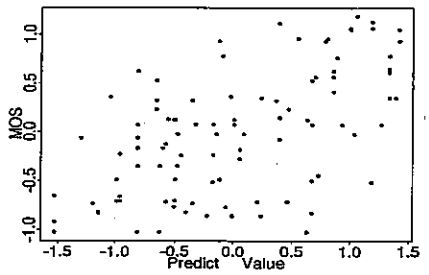

Figure 4: Relationship between MOS and the predicted value by the regression tree for the test data.

5 まとめ

音素接続型合成における母音部の接続歪みに対する不連続感を自動的に検出するため、日本語母音の接続歪みと triphone コンテキストに含まれる母音の種類、音響パラメータとの関係について調べた。接続歪みは母音の発話位置及び様式と深い関係があることが分かった。接続歪みを予測する回帰木モデルを構築し、各特徴ファクターの寄与及び MOS 値の予測について検討した。CD, $F0.D$, $PWR.D$ は有効な予測ファクターであることを示し、学習データ及びテストデータに対して安定な予測率が得られた。

今後は今回の実験のもとに、音素の自動クラスタリング及び単位選択への応用、部分的信号処理を用いた時の接続歪みの削減について検討する予定である。

謝辞

日頃ご指導頂く山本社長、本研究を進める上で多くのご指導を頂いた匂坂室長に感謝します。また、ご討論頂いた ATR 人間情報通信研究所の加藤 宏明氏に感謝します。

References

[1] 丁 文、藤澤 謙、ニック キャンベル、樋口 宜男、音響学会講演論文集, pp.211-212, (1997-09).

[2] ニック キャンベル、アラン ブラック、 信学技報、 SP96-7, pp.45-52, (1996-05).