ＴＲ－ＳＬＴ－００２３

中国語 TTS 音声合成のための接続環境代替と韻律制御

Substitution of Chinese Syllable and prosodic control for Concatenative Speech Synthesis of Chinese

陸　金林
Jinlin　Lu

2002 年 9 月 30 日

概要

近年のＴＴＳ音声合成において、高品質な自然音声を得ることが最大な目標である。本研究報告では、自然性に大きく影響する単位接続について、単位接続コストを減らすために、中国語音節の音素と声調の代替接続による自然性劣化の知覚的評価実験を行った。音素のタイプ、声調の種類による単位接続時の自然性劣化の度合いを表にまとめた。また、自然な韻律を生成するために、名工大らの HMM に基づく音声合成方式で、韻律生成法を中国語に適応し、その試みの結果について述べる。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目 2 番地 2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

目次

1　はじめに

近年、規則による音声合成方式が実用化され、製品やサービスに徐々に利用されている。機械から音声を出力する音声合成は、ヴォイスメール、人と機械の対話など、テキストから音声を合成するテキスト音声合成システム（TTS）は更なる多くの利用が見込まれている。音声合成では、自然性に影響を与える主な要素として、単位接続による音質劣化と韻律制御の不適切による違和感があげられる。自然な合成音を得るために、音声合成における音声単位の選択に関する研究が報告されている[1-3]。波形接続型の音声合成では、接続単位が抽出時と使用時の音素環境の不一致に起因する接続歪みは日本語などについて検討されている[2-3]。その結果、VCV 音声の環境代替のときで、音素によって自然性の評価値が異なる。しかし、中国語についてはこのような研究例はあまり見ない。

一方、現在、テキストから音声を合成する音声合成方式が数多くあるが、HMM に基づく音声合成についてもいくつか提案されている[1,2]。その中に、スペクトル、ピッチ、継続長を HMM の枠組みで統一的にモデル化する手法[2]が提案され、スペクトル、ピッチ、継続長モデルとしてそれぞれ連続分布 HMM、多空間確率分布 HMM、多次元ガウス分布を用い、音素環境、アクセント、品詞などのコンテキストを考慮した決定木に基づくコンテキスト依存モデルを構築することによって、良好な韻律制御が行なえることを日本語音声合成によって確認されたことが報告されている。中国語音声についても、同様な手続きで行なうことで韻律を制御することを試みる。

本報告はこれらについて実験と検討を行なった結果を以下に詳しく述べる。

2　中国語音節の音素と声調の代替接続による自然性劣化の知覚的評価

2．1　はじめに

中国語音声合成において、音節の接続では、前の漢字の“韻母”(母音、Final)と後続する漢字の“声母”(子音、Initial)との接続を考慮するだけでなく、各環境においての声調(Tone)のつながりについても考慮する必要がある。中国語の音節には、Final が 38 種、Initial が 22 種あり、２音節単語においての声調組み合わせでは、前後音節の声調変化も 4×4=16 種類以上ある。これらの接続に関係する要因だけを考慮しても、すべての組み合わせを実験することは膨大な量が必要であるため無理である。そこで、本報告では、音声合成の単位を音節と想定したとき、単位選択のための基礎実験として、その一部について検討する。予備実験では、Final と Initial と声調の適当な組み合わせによって、環境代替の全体の様相を見る。次に、同じ声調の音節間の Final と Initial の環境が代替されるとき、それに、接続する Final と Initial を一定にし、声調が異なるときの声調代替が、接続音声の自然性に対する評価を聴覚実験によって調べる。

2．2実験方法

本報告での代替接続は波形接続を用いている。2音節単語があったとき、その接続の1例を図1に示す。

図1の例では、2つの2音節単語（W1とW2）から新たな2音節単語（W3）を作り出す例である。各単語の第1音節をS1、第2音節をS2に記する。単語W1とW2の第1音節S1、第2音節S2をそれぞれ音節S11,S12とS21,S22としたとき、単語W3は音節S11とS22との接続からなる。音節を子音、母音、声調の順で表わすと、本例では、S11＝di4, S22＝yi3 の場合である。今回の実験では、波形を直接接続することで、各音節間の基本周波数の調整は考慮していない。なお、音節の構成部分の組み合わせ条件によって後述の3つの実験を行なう。

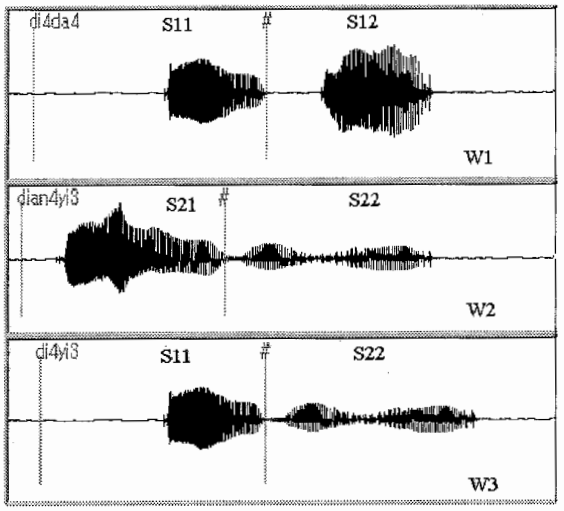

図1 波形接続による音節の代替接続

（W1、W2： 元の単語； W3： 接続された単語）

この例のW3は、W1のS12の子音 d がW2の y によって代替された。また、W2のS21の母音 ian がW1の i によって代替されたといえる。

以上のように接続合成された音声のつながり部分の良さについて、7段階(3: 非常に良い ～ -3: 非常に悪い)の聴覚評定を行なった。刺激音声は実験ごとにランダムして防音室でヘッドホンによる両耳受聴する。毎回の本番実験の前に、練習として、30サンプルを試聴する。実験結果は各被験者の得点を平均0、分散が1になるように規格化する。

2．3 実験内容と実験結果

2．3．1 実験1：予備実験

実験1は予備実験として、まず全体的な様相を見るために、第2音節S2のInitial(I2)をできるだけ /d/ にし、第1音節S1のFinal(F1)を /i/ にするほかはあまり制限を与えていない。

中国語音声合成データベース CoSS-1 (Corpus of Speech Synthesis -1)の 4 音節単語から先頭の 2 音節を単語試料とする。S1 の Final の発声方法と S2 の Initial の調音様式などを考慮して選択したものは表 1 に示すように 11 単語である。声調については、特に限定していない。音声は男性 2 名、女性 1 名の計 3 名が発声したものを用いる。すべての単語について、音節 S1 と自分自身の単語の S2、それに他の単語の S2 を波形接続して、つなぎ合わせて新しい単語（環境代替）を作成する。このように作成した 121 単語/人×3 人 =363 単語をランダムにして 1 セットとし、4 名の被験者がそれぞれ 4 セットを試聴した。

表1. 2音節無意味単語試料

S1			S2		
I1	F1	T1	I2	F2	T2
d	a	4	sh	i	1
d	ian	4	z	i	3
d	ao	3	t	i	3
d	ing	1	y	i	1
d	eng	1	sh	i	1
d	a	3	s	i	3
t	a	3	j	i	2
d	ian	3	q	i	1
d	i	4	zh	i	1
d	ian	4	j	i	2
t	ong	3	y	i	1

後続 Initial(I2)が代替されるときの自然性の評価値を図 2 に示す。半母音/y/と摩擦音 /s, sh/ の劣化が大きい。これらはバウンダリの切り出しが難しいものであり、前後の音節に大きな相互影響を与えるケースであると考えられる。一方、閉鎖部を持つ破裂音など境界が相対的にはっきりしているので、よい音質が得られている。

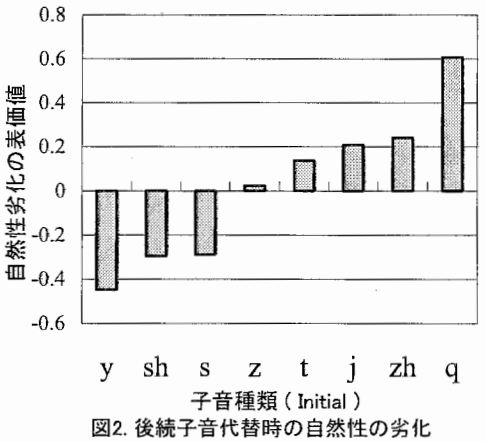

図2. 後続子音代替時の自然性の劣化

一方、先行母音(F1)が代替されるときの実験結果を図3に示す。この7種類のFinalについては、全体的に自然性の変化は少ない。F1がingのとき、相対的に自然性の劣化が生じやすいが、Initial の代替に比べて小さい。

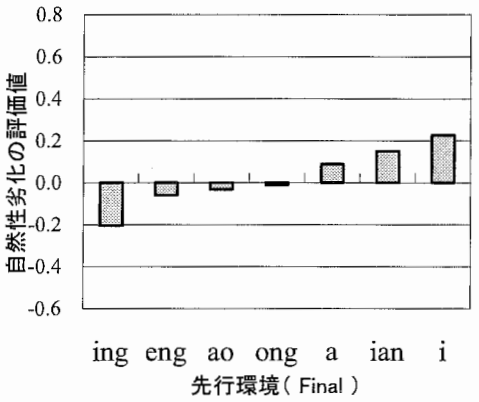

図3. 先行母音代替時の自然性の劣化

これらの実験結果より、先行母音(F1)の代替よりも、後続Initial (I2) が代替されるとき、Initialの種類によって自然性の劣化が大きいものがある。このように、声調については考慮されていないものの、音節の代替で自然性の劣化が生じ、音声合成で良い合成単位を選択するときには、考慮すべきものである。そこで、以下にもっと詳しい実験を行ない、今後合成単位選択時の一つの参考基準になるようにする。

2.3.2 実験2：後続Initial (I2) の代替

実験2ではより詳しい後続子音環境を考慮する。すべての声調を第1声にする。子音の発声方法と発声部位を考慮して、第2音節の子音I2 を/d,t,z,c,zh,ch,j,f,h,m,l,r,w,y/ とし、また、ゼロ子音としての母音連鎖で、/a,e/を用いて、合計16単語である。ゼロ子音の/a,e/の音節には、その子音を今後0a, 0eと呼ぶ。これらの単語に対応する第1音節の子音I1は音節zi(c), zhe, zhi(ch)以外は/d/に固定する。F2は発声位置や構成成分などを考慮して、以下のような16種類を用いる。

/i,ai,ic,ie,e,ai,au,ao,ich,iao,uo,an,ian,ang,eng/

一方、第2音節の母音F2 はできるだけ/a/に統一するように、4つの音節のjiao, yi, e, reng以外は全部/a/にする。これら合わせて256単語に接続できる。

実験2の音声は成人女性1名がATR防音室でコンデンサマイクロホンを用いて収録した。音声は48kHzでサンプリングしてから16kHzにダウンサンプリングした。被験者は3名である。3人とも各刺激音を4回聴取する。

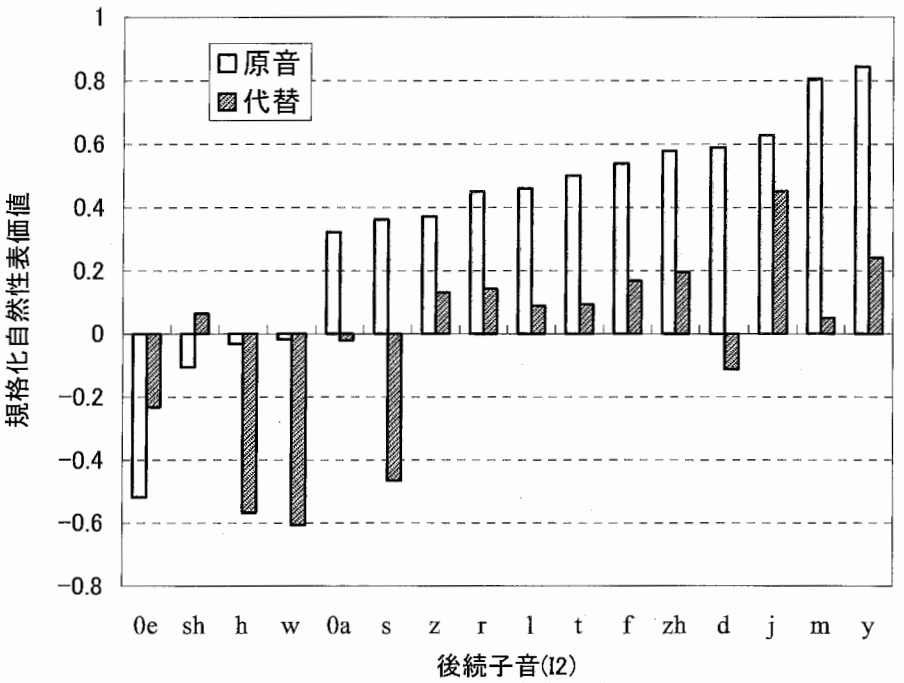

図4 後続子音が代替時の自然性劣化

図4には、同単語間の接続（原音）と異単語間の接続（代替）による単語の自然性評価値を示す。一部の原音の自然性評価値が低いが、これは今回用いる自然発話の音質と波形接続に起因するものである。後続子音の種類によって自然性の変化が大きいことが分かる。

表2には、自然性の劣化程度を下のように設定したとき、各子音が代替されるときの自然性の劣化を示すものである。

劣化の自然性評価値 ＝ 原音の自然性評価値 — 代替音の自然性評価値

表2 後続子音の劣化

	劣化	原音
s	0.83	0.36
m	0.76	0.80
d	0.70	0.59
y	0.60	0.84
w	0.59	-0.02
h	0.54	-0.03
t	0.41	0.50
zh	0.38	0.58
l	0.37	0.46
f	0.37	0.54
0a	0.34	0.32
r	0.31	0.45
z	0.24	0.37
j	0.18	0.63
sh	-0.17	-0.11
0e	-0.29	-0.52

表2より、音声バウンダリを容易に定められない音素の代替による劣化は大きい。特に子音 s, m, y, w の劣化が大きい。劣化が少ない、或いは逆転されるものは原音の音質が良くない sh と 0e だけである。

2.3.3 実験2：後続 Initial (I2) の代替

前節は後続子音環境について考慮したが、ここでは先行母音について考える。実験は前節のもので、先行母音の代替による自然性の変化をみる。

3.2節と同様に、先行母音が代替されるときの自然性の変化をそれぞれ図5と表3に示す。

先行母音が代替されるとき、原音の音質は後続子音と一対一の関係で、同順位であるが、母音によって代替した後の自然性の劣化程度は異なる。子音が代替したときの自然性の評価値が3つも -0.4 以下であるが、母音の代替では1個だけである。また、どちらも原音の音質が良くないときは、代替した後も自然性が悪い。図4と図5を比較してみると、原音の音質が高い単語は、代替によっても自然性がよいが、元の音質からは相当落ちている。

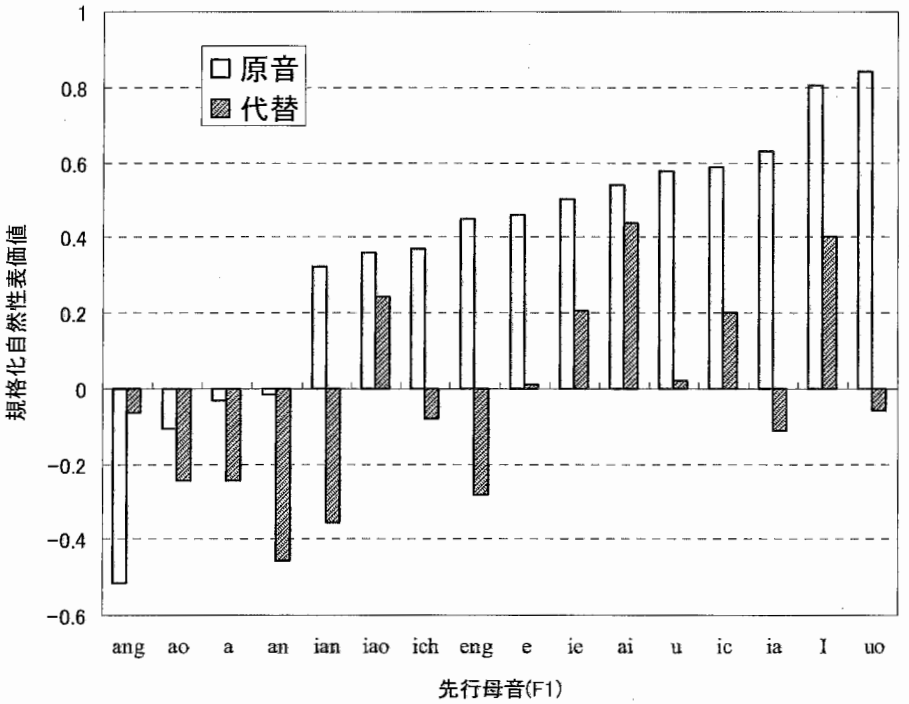

図5 先行母音が代替時の自然性劣化

表3 母音代替の劣化

	劣化	原音
uo	0.90	0.84
ia	0.74	0.63
eng	0.73	0.45
ian	0.68	0.32
u	0.56	0.58
e	0.45	0.46
ich	0.45	0.37
an	0.44	-0.02
I	0.40	0.80
ic	0.39	0.59
ie	0.30	0.50
a	0.21	-0.03
ao	0.14	-0.11
iao	0.12	0.36
ai	0.10	0.54
ang	-0.46	-0.52

2.3.4 実験3:声調の代替

実験3では、声調の異なる環境から代替接続するときの影響を調べるものである。2音節単語がそれぞれの音節各4つの声調の組み合わせで、最大16種類の声調組み合わせの単語ができる。2種類の単語を用いることで、16×16 ×2 = 512単語になる。原音も入れると、544サンプルになる。たとえば、2つの単語の声調がそれぞれT1T2 = 12（第1, 第2音節）とT1T2 = 34としたとき、声調が12 と34 のままで単語を構成できる(原音)以外に、音節の位置を変えないで、14 と32 の声調を持つ単語も接続して合成できる(代替)。実験に用いる単語の音節はそれぞれ、S1 = da, S2 = ma と S1 = di, S2 = da である。今回用いる原音の音質が、T1T2 = 21, 33, 43 の単語のとき、とても悪いことから、それらの結果は取り除いている。

実験3の音声収録は実験2と同じである。被験者も同様の3人である。

表4には、実験3の異なる単語からの異声調の接続実験の平均結果である。T1T2 =11, 41, 44のとき、最も良い代替が行なえることが分かる。また、第3声の音節はピッチ周波数の変化が単純ではないので、波形接続後の劣化が大きい。

表5には、更なる詳しい評価値を示している。第一音節が第1声、或いは第2声のときの同士代替は高い自然性が得られる。一方の第一音節の声調が第4声のとき、もう一方の単語の声調が24, 31, 34, 41, 44のいずれの場合もよい状況である。第3声は全体的に自然性劣化が大きい。

表4 声調の代替による自然性の劣化

		第2音節			
	声調	1	2	3	4
第一音節	1	0.3	−0.2	0.2	0.1
	2	0.2	−0.3	0.1	0.1
	3	−0.2	−0.1	−0.5	−0.2
	4	0.5	0.1	−0.3	0.3

表5. 声調が代替されるときの自然性劣化の評価

		第2音節が接続される単語の声調													
	声調	11	12	14	22	23	24	31	32	33	34	41	42	44	平均
第一音節が接続に用いられる単語の声調	11	1.0	0.7	0.4	0.3	0.3	0.5	0.2	−0.7	−0.1	−0.2	0.1	−0.6	−0.3	0.1
	12	1.0	0.4	0.8	0.1	0.3	0.3	0.1	−0.8	0.1	−0.5	−0.0	−0.3	−0.1	0.1
	14	0.6	0.7	0.3	−0.3	0.3	0.2	−0.0	−0.9	0.2	−0.1	−0.2	−0.8	−0.2	−0.0
	22	0.7	0.1	0.3	0.1	0.1	0.5	0.2	−0.6	0.4	0.2	0.0	−0.8	−0.2	0.1
	23	0.3	−0.1	0.2	0.0	0.1	0.2	0.0	−0.8	0.0	−0.2	−0.3	−0.5	−0.5	−0.1
	24	0.5	0.1	0.2	0.1	−0.1	0.3	−0.0	−0.7	0.2	0.1	−0.0	−0.8	−0.4	−0.1
	31	−0.6	0.2	−0.6	−0.1	−0.8	−0.1	0.2	0.4	−0.5	−0.1	0.1	0.4	0.2	−0.1
	32	−0.9	−0.4	−0.9	−1.0	−0.8	−0.7	−0.2	0.0	−0.7	−0.2	−0.2	−0.1	−0.2	−0.5
	33	0.2	0.5	0.1	0.4	0.5	0.2	0.2	0.1	0.1	0.2	−0.0	−0.3	−0.1	0.2
	34	−0.5	−0.6	−0.4	−0.6	−0.9	−0.4	0.0	0.2	−0.5	−0.1	0.1	0.0	0.1	−0.3
	41	0.1	0.1	−0.1	−0.1	−0.5	0.3	0.3	0.0	−0.2	0.1	0.8	−0.1	0.5	0.1
	42	−0.1	0.1	−0.2	−0.0	−0.3	0.3	0.9	0.2	−0.1	0.4	0.9	0.2	0.7	0.2
	44	0.1	0.2	−0.2	−0.1	−0.3	0.2	0.8	0.3	−0.2	0.4	1.0	−0.1	0.6	0.2
	平均	0.2	0.1	−0.0	−0.1	−0.2	0.1	0.2	−0.2	−0.1	−0.0	0.2	−0.3	0.0	

本実験では、基本周波数の相違による生じる接続歪みについては考慮していないので、音声単位選択時にこれらの結果を用いる時は、その因素を除外する必要がある。

2.5 まとめ

中国語音節の代替接続について、音素と声調の両方から検討を行なった。先行 Final を代替することよりも、後続 Initial を代替したときに引き起こす自然性の劣化の大きいものが多い。また、音声バンダリを容易に決める音素でないときの劣化も大きい。声調については、第1声と第2声から始まる単語同士の代替は比較的にスムーズに行なえることが分かった。今後はさらに多くの母音連鎖についての結果を検討し、合成システムに反映する方法を検討する必要がある。

3 中国語音声合成における HMM に基づく韻律生成について

3.1 はじめに

声調言語である中国語は、音節ごとに声調が付与されていることで、日本語より韻律の変化がはるかに激しいことがこれまでの研究からも周知である。波形接続型の中国語テキスト音声合成システムを作成する際に、データベースの拡充、接続歪みの削減方法の検討が必要である一方、いかに適切な韻律の制御ができるかも、合成音声の自然性を上げる重要なポイントである。中国語のピッチ周波数のモデル化については、いくつかの研究が発表されているが[6]、テキスト音声合成に応用して、良好な自然性を得るためにはまだ改善の余地がある。

本報告は、以上のことを踏まえて、名古屋工業大学などで開発される統一モデル化音声合成手法を中国語に応用し、中国語音声の韻律を制御することを試みる。音声合成システムからの韻律生成結果を波形接続型中国語音声合成システムの韻律制御に応用する。

3.2 コンテキストの種類

本報告では、1 つの HMM は 1 つの半音節に対応する。半音節を最小単位として、音節、形態素(単語)、呼気段落、文の順で階層的にコンテキスト情報を取り扱う。以下のコンテキストを考慮した。

・文の長さ
・当該呼気段落の位置
・{先行、当該、後続} 呼気段落の長さ
・当該形態素の位置、前後のポーズの有無
・{先行、当該、後続} 形態素の品詞 (28 種類)、長さ
・当該音節の位置
・{先行、当該、後続} 音節の声調(5 種類)、声調変化
・{先行、当該、後続} 半音節(60 種類)

ここで、長さ、位置の単位は音節である。声調変化とは、音節の前後環境によるレキシコン声調の変形であり、例えば、多くの連続した第 3 声と第 3 声の音節が第 2 声と第 3 声の音節の発声に変わることを指す。

なお、文頭、文末の無音、文中のポーズも半音節として扱っている。

3.3 HMM 音声合成システムの構築

まずは HMM 音声合成システムを構築する。HMM の学習データとして、現在 ATR の防音室で収録している中国語音声データベース[4]の一部の 469 文(約 1.2 時間、音韻バランス平叙文、女性ナレータが発声)のうち、420 文を学習データとして用いた。サンプリング周波数は 16kHz、分析周期は 5ms とした。メルケプストラム分析を行い、パワー情報を含む 0～24 次のメルケプストラム係数を求めた。特徴ベクトルとして、メルケプストラム係数及

びその動的特徴量であるデルタ、デルタデルタ、とピッチ及びそのデルタ、デルタデルタを合わせた全 78 次元のベクトルとした。

図 1 に非学習データについて生成されたピッチパタンを示す。生成されたピッチパタンは継続長の調整(応用)が現時点で未完成であり、文単位全体での時間の線形伸縮を行っている。一方、学習データの量による違いを見るため、学習文をそれより小さいほうの 170 文、340 についてもピッチパタンを生成した。

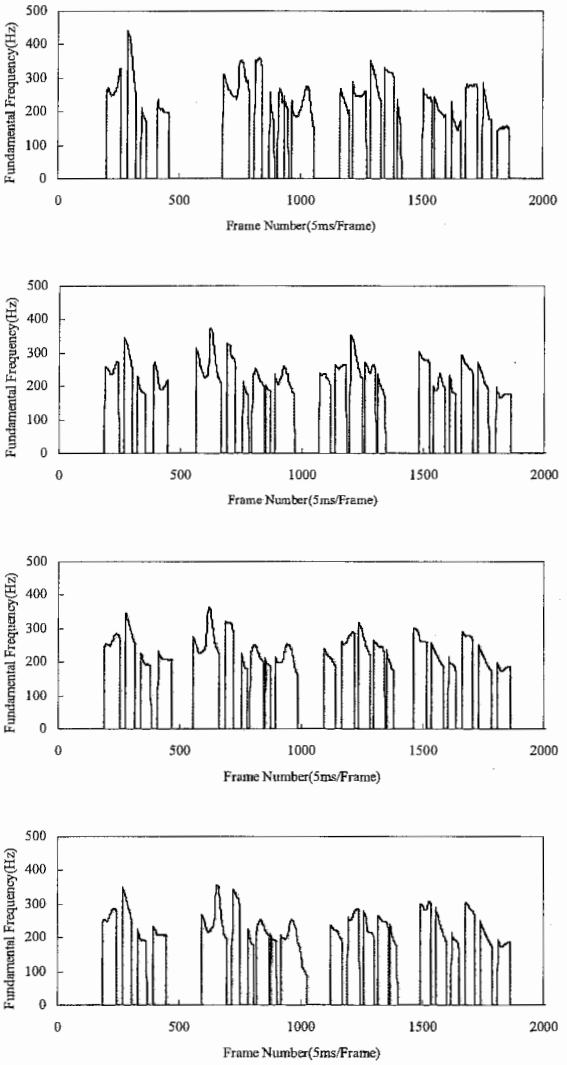

図1　生成されたピッチパタン(上から順に　原音声、170 文、340 文、420 文の学習)

図1のように、動的特徴量を用いても、生成されるピッチパタンは学習データの量による違いがある。学習データがどの程度あれば満足できることは現段階のこの結果ではまだ言えないが、生成されるピッチパタンが、半音節をモデル単位に利用していることで、声

調が必ずしも目標の声調と一致していないことが分かる。しかし、音節内のピッチパタンは滑らかであることが言える。

非学習音声10文章を、以上の3種類の学習データにより生成されたピッチパタン、継続長及びメルケプストラムでHMM合成を行ったところ、その音を3人の試聴者が非公式にそれぞれ2回試聴した。その音質は学習文の数が多くなるにつれて、7段階のMOS値の平均でそれぞれ3.81, 4.2, 4.43である。この生成方法を中国語に応用する現段階において、学習データの量を増加する必要があると見られる。これは今後改良すべき点でもある。

3.4 まとめ

本章は、日本語HMM音声合成システムを中国語に応用し、中国語の韻律生成を試みた。コンテキストを考慮した決定木に基づくコンテキスト依存モデルを用いることによって、中国語の韻律制御にも効果があることが分かった。

今後、コンテキストの構成として、声調が形態素に対する影響なども考慮して検討する必要があると見られる。また、学習データの増加についても検討する必要がある。

4 まとめ

本報告は波形接続型の音声合成では、接続単位が抽出時と使用時の音素環境の不一致に起因する接続歪みは、音素によって自然性の評価値が異なることから、中国語について調べ、注意すべき音素などが分かった。

一方、現在、テキストから音声を合成するために、HMMに基づく音声合成についても、スペクトル、ピッチ、継続長をHMMの枠組みで統一的にモデル化する手法を用い、スペクトル、ピッチ、継続長モデルとしてそれぞれ連続分布HMM、多空間確率分布HMM、多次元ガウス分布を用い、音素環境、アクセント、品詞などのコンテキストを考慮した決定木に基づくコンテキスト依存モデルを構築することによって、良好な韻律制御が行えることを中国語音声合成に適応して確認した。

参考文献

[1] 舛田 剛志、他、“単位接続型音声合成における音素環境代替による自然性劣化の知覚的評価”、音講論、2-6-11, pp. 303-304, 2001.3.

[2] 河井 剛、他、“波形素片接続時の音素環境代替による自然性劣化の知覚的評価”、信学技法、SP 2001-22 (2001).

[3] 岩橋直人、他、“音響的尺度に基づく複合音声単位選択法”、信学論、Vol.J72-D-II, PP.1174-1179 (1989).

[4] R.E. Donovan and E.M. Eide, “The IBM trainable speech synthesis system,”, Proc. ICSLP, Vol.5, pp.1703-1706 (1998).

[5] 吉村、徳田、益子、小林、北村、“HMMに基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化”、信学論(D-II), 2099-2107(2000).

[6] Lu, S.N.;Yang,Y.F.;Cao,J.F.;Huang,J.C. “Prosodic control in Chinese TTS system.” ICSLP 03065(2000).

[7] J. Ni, H. Kawai, “Phonetically and prosodically balanced text corpus for mandarin speech synthesis,” 音講論、3-2-8, pp. 319-320, (2001.9).