

TR-I-0040

種々の合成素片を用いる規則合成システムにおける
韻律制御法の改良

Improvement of prosody controls
for speech synthesis by rule system

山崎岳人†, 武田 一哉, 桑原 尚夫

Takehito Yamazaki, Kazuya Takeda and Hisao Kuwabara

1988.9

内容更概

種々の複合音声単位を選択的に抽出使用する規則合成システムのための、韻律制御法の改良について検討する。なお本稿は、静岡大学工学部修士課程1年の山崎岳人君がATRにおいて行った学外実習の実習報告書をまとめたものである。

†静岡大学工学部

Department of Engineering, Shizuoka University

ATR 自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

目次

1. はじめに	1
2. 規則合成システム	1
3. 音韻継続長	1
3.1 音声事象に応じた継続長の設定	2
3.2 閉鎖部の欠落に伴う異音化の改善	2
4. ピッチ周波数	2
5. パワー	3
6. まとめ	3
Appendix	

1 はじめに

我々は、種々の複合音声単位を音声データベースから選択的に抽出・使用する、規則合成システムの構築を進めている[1]。この方式は、入力音韻系列に応じて合成素片を選択することにより、以下の特徴を有している。

- ① 合成素片抽出環境と使用環境との一致
- ② 様々な調音結合の合成
- ③ 単位接続歪の抑圧

これまで、本手法における合成単位選択基準尺度の検討が報告されている[2]。

しかし生成される合成音声が自然で聞きやすいものであるためには、韻律の制御が自然でなければならない。本稿では、システムの韻律制御アルゴリズムに加えた改良について述べる。

2 規則合成システム

本稿で述べる韻律制御アルゴリズムは、不均一な合成素片を用いる規則合成システムのために開発されたものである。本システムの概略を図1に示す。図において各処理モジュールは四角で、処理に伴うデータファイルは2重下線で、データの流れは矢印で、それぞれ表されている。

unit_selectで選択された合成素片は、mkunitにおいてデータベースから抽出される。一方、durationは入力文字列を構成する各音韻の継続長を生成する。抽出された合成単位は、durset及びmkcepによって最適な時間長に伸縮される。合成音声のピッチパターンは、mkpitchにより生成される。合成音声波形は、ピッチパターン及び継続長制御された合成素片のケプストラム係数より、lmaフィルタを用いて、生成される。

本稿ではこれらの処理モジュールのうち、durset、mkcepにおける継続時間長の制御の改良について3節で、mkpitchにおけるピッチパターン生成の改良について4節で述べる。さらに5節では、今回新たに作成したパワー正規化モジュールpowerについて述べる。

3 音韻継続長

合成単位を選択的に用いる本手法では、抽出用データベースが種々のモーラ長を持つ単語から構成されているため、音韻長の制御は不可欠である。これまでのシステムでは、データベースから抽出した規則により、各音韻の継続長を生成し[3]、均一な間隔によるフレームのくり返し(間引き)により音韻の伸縮を行ってきた。この方法は、無破裂音の伸縮に際し明らかに、以下の問題を持っている。

- 1) 間引きにより、音韻性を多くになう破裂点付近のフレームが、失なわれる可能性がある。
- 2) 閉鎖部の極端な短縮、あるいは消去により、異音(「き」→「ひ」、
「か」→「は」等)を生じる。

そこで今回上記の2点について、以下に述べる改良を行なった。

3.1 音声事象に応じた継続長の設定

無声破裂音の音韻特徴として、破裂部のスペクトル[4]及び、VOTの長さが指摘されている[5]。このため無声破裂音の伸縮は、閉鎖部及び破裂+気音部(VOF)の時間的特徴、破裂点付近の波形を極力保持して行なうことが必要となる。そこで今回、図2に示すアルゴリズムを開発した。すなわち、

- ① 合成用素片の継続長と、規則により生成された継続長とを比較し、必要なフレームくり返し(間引き)回数を決定する。
- ② データベースより得られた統計(表1)により求めた、閉鎖部及び気音部の時間長特性に基づき、音韻毎に定めた割合により必要なフレームくり返し(間引き)回数を、各音声事象にわり振る。
- ③ 両音声事象に対し、②で求めた回数のフレームくり返し(間引き)を行なう。ただし、閉鎖区間終了直後のフレームは、破裂部に対応するフレームと見なしそのまま使用する。

つぎに、継続時間長の伸縮の上限と下限を設定した。これまでのシステムでは、規則により生成された継続長に従い無制限に伸縮を行ってきた。しかしあまり極端な伸縮には音質の劣化が避けられない。そこで伸長する場合は最大でも元の音韻の時間長の1.5倍まで、短縮する場合は最小でも元の音韻の時間長の0.5倍までに、伸縮を制限した。

3.2 閉鎖部の欠落にともなう異音化の改善

単位選択の結果、語頭から抽出された無声破裂音が、語中に用いられる場合がある。この場合、単位抽出用データベースのラベルの性質から、無声破裂音の閉鎖部が欠落してしまい、「か」→「は」、「き」→「ひ」の様な異音が生じる。これに対応するため、図3に示すように、閉鎖区間の挿入を行なうアルゴリズムを加えた。即ち、伸長する場合にはフレーム数だけ先頭にポーズ(closureに相当する)を挿入し、縮小する場合には、5フレームのポーズを挿入後、VOTの部分を縮小する方式をとった。これにより、closure部分の欠落に伴う異音化が避けられるようになった。

4 ピッチ周波数

合成音の聴取により特定の有声音(特に母音)部分で、いわゆるブザー音的な音質劣化が明らかになった。これらはいずれも、ピッチがかなり低い部分であったことから、ピッチ周波数の調整を試みた。本システムでは、いわゆる藤崎モデル[6]を時間正規化し、大語彙単語に対しA-b-S法[7]を適用することで得られたモデルにより、ピッチ周波数を生成している。そこで、モデルのパラメータの調整により改善を試みた結果、ピッチ周波数が低い部分を数十Hz程度上げると、ブザー音は無くなることがわかった。一方あまり大きくピッチ周波数を上げると、イントネーションが変化し自然性が低下した。そこでピッチ周波数の下限を100Hzに制限する形でピッチを調整することとした。

5 パワー

合成音を聴取すると、パワーの不連続感による音質劣化が明らかになった。これを解決するためには、パワーの正規化が必要である。そこで図4に示す正規化を行った。すなわち、

- ① ピッチパターンの概形を強調し、文全体のパワーパターンを生成する。
- ② 生成されたパワーパターンに、音韻毎の平均パワー値をかけ、ターゲットパワーとする。
- ③ 音韻毎にターゲットパワーを中心に、適当な範囲を定め、各音韻のRMS値が、この範囲含まれるよう波形レベルで拡大、縮小を行なう。

表2にデータベースにおけるパワーの統計値を、表3にそれを基に決定したパワーの許容範囲をそれぞれ示す。パワーパターンの生成に関し、

- (1) ピッチパターンのルート
- (2) ピッチパターンそのまま
- (3) ピッチパターンの5乗

の3者を比較したところ、明確な差異は聞きとれなかった。前節で述べたように、低いピッチ部分で音質の劣化が大きいことから、本改良では、ピッチパターンの5乗をパワーパターンとして用いることとした。

6 まとめ

規則合成システムに対して、以下の韻律制御の改良を行なった。

- 1) 音声事象に応じた音韻継続長制御アルゴリズム
- 2) ピッチ周波数の極端な低下を回避する、ピッチ制御アルゴリズム
- 3) ピッチ周波数の概形にあわせこむパワー制御アルゴリズム

今回の改良により、自然な韻律の制御が実現された。今後は、単位選択及び接続手法について改良を加えて行く。

謝辞

今回の実習の機会を与え、御指導下さった樽松社長、鹿野室長をはじめとするATRの皆様に感謝いたします。

文献

- [1] Y.Sagisaka "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc. ICASSP, (1988.4)
- [2] 安部他 "入力音韻系列に応じた音声素片の選択法," 音講論集, 1-1-23, (1988.3)
- [3] 武田他 "文発声と単語発声における音韻継続長の分析," 音講論集, 2-1-6, (1988.3)
- [4] Zue V.W. "Notes on spectrogram reading," MIT Preliminary Draft (summer 1985)

- [5] 川端他 "日本語スペクトログラム特徴の英語との比較," 音講論集, 3-3-10,(1987.10)
- [6] 藤崎他 "日本語単語アクセントの基本周波数パターンとその生成機構のモデル," 音響学会誌 27(9),(1971)
- [7] 宮武他 "種々の発声様式における基本周波数パターンの制御モデルの検討," 音講論集, 1-1-16,(1988.3)

表1(1) 音韻"p"のイベント構成別継続長

イベント 構成	イベント 名	サンプル 数	平均 [ms]	標準偏差 [ms]
all	all	33	122.8788	23.5176
>clp	all	11	161.8182	29.8499
>clp	>	11	33.6364	27.0388
>clp	cl	11	109.3182	23.0388
>clp	>cl	11	142.9545	31.5115
>clp	p	11	18.8636	11.0969
clp	all	22	120.2273	23.3295
clp	cl	22	104.7727	19.6259
clp	p	22	15.4545	6.8522
all	>	11	33.6364	27.0388
all	cl	33	106.2879	20.9357
all	p	33	16.5909	8.6563
all	>cl	11	142.9545	31.5115

実際に用いた伸縮比率

cl:p = 90:10

表1(2) 音韻"t"のイベント構成別継続長

イベント 構成	イベント 名	サンプル 数	平均 [ms]	標準偏差 [ms]
all	all	469	93.8913	27.0552
>clt	all	308	109.7565	26.6360
>clt	>	308	20.1867	6.1126
>clt	cl	308	73.7500	26.6946
>clt	>cl	308	93.9367	26.4503
>clt	t	308	15.8198	4.9794
clt	all	161	102.1584	25.2280
clt	cl	161	85.9472	25.5938
clt	t	161	16.2112	5.5105
all	>	308	20.1867	6.1126
all	cl	469	77.9371	26.9515
all	t	469	15.9542	5.1712
all	>cl	308	93.9367	26.4503

実際に用いた伸縮比率

cl:t = 85:15

表1(3) 音韻"k"のイベント構成別継続長

イベント 構成	イベント 名	サンプル 数	平均 [ms]	標準偏差 [ms]
=====				
all	all	1427	115.7025	26.6700
>clk	all	1062	138.6982	27.8879
>clk	>	1062	22.5235	8.5205
>clk	cl	1062	75.1460	19.2528
>clk	>cl	1062	97.6695	20.3616
>clk	k	1062	41.0287	13.1626
clk	all	364	114.3132	26.4869
clk	cl	364	78.6401	20.7228
clk	k	364	35.6731	12.1005
k	all	1	120.0000	0.0000
k	k	1	120.0000	0.0000
all	>	1062	22.5235	8.5205
all	cl	1426	76.0379	19.6979
all	k	1427	39.7179	13.2761
all	>cl	1062	97.6695	20.3616

実際に用いた伸縮比率

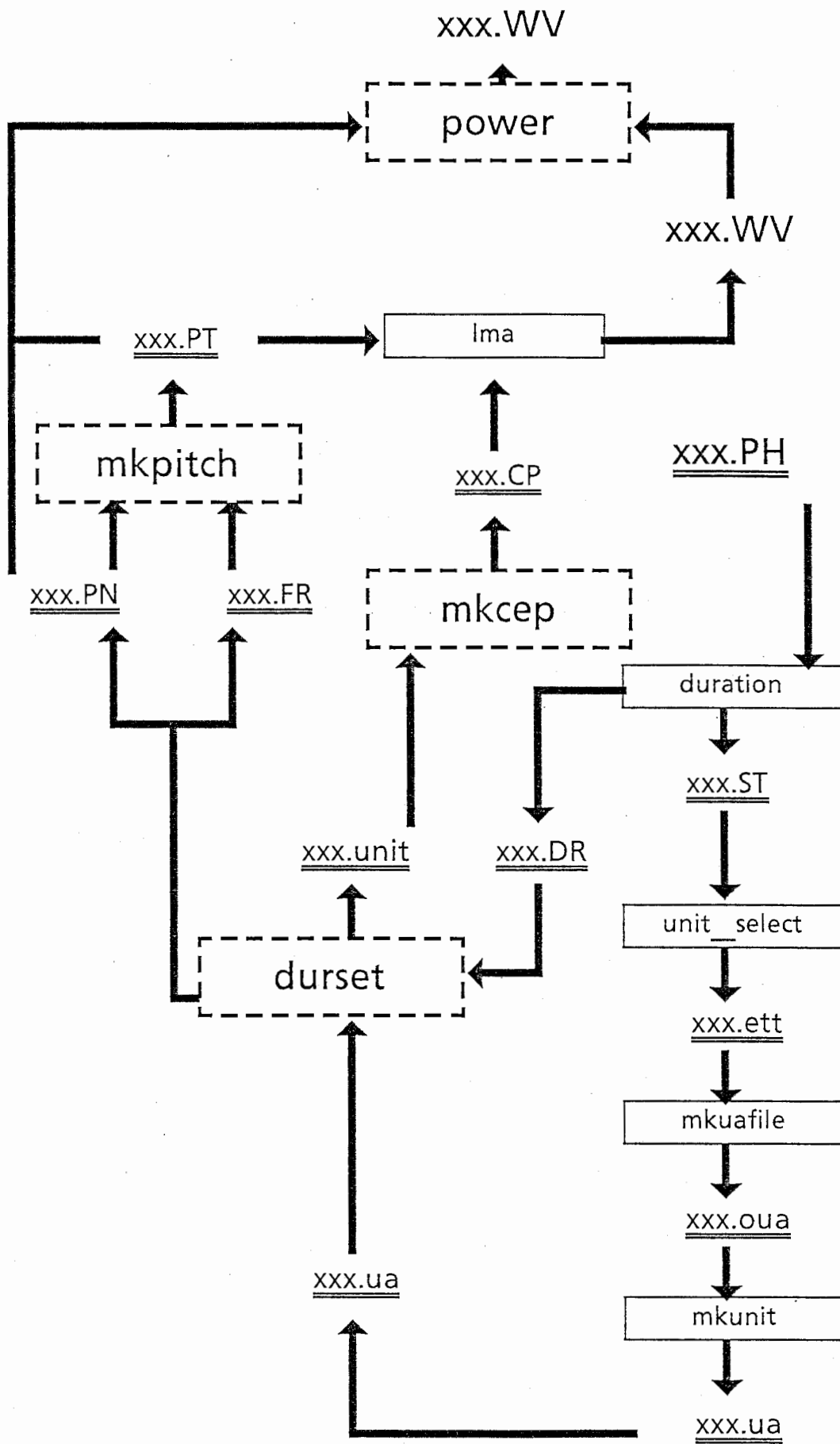
cl:k = 70:30

表3 データベースにおける有声音のパワーの平均と分散
(p-pRMS値)

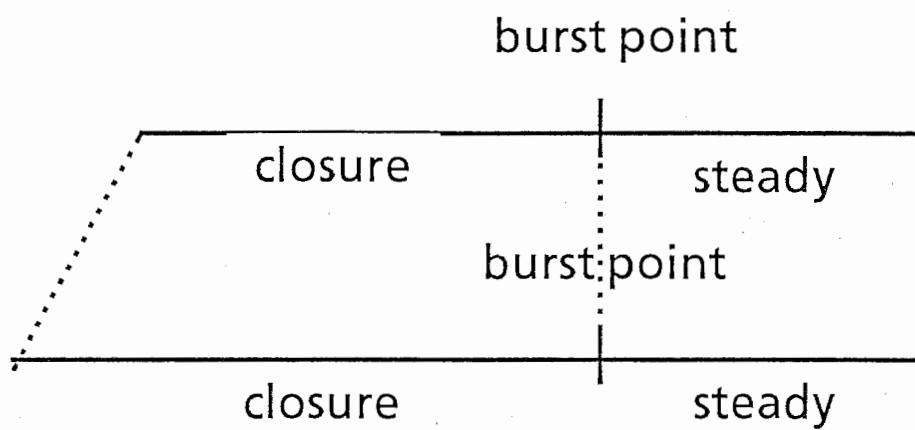
音韻	平均	標準偏差	サンプル数
#a	4231.29980	1440.25696	3473
#i	2408.72632	1184.99939	2522
#u	2089.82666	1233.65552	2917
#e	3890.01367	1204.17859	1328
#o	3686.51978	1390.82068	1740
#N	1886.42053	1105.26733	909
#b	962.40314	371.32996	415
#d	851.27594	367.10446	355
#g	2235.56934	1312.55725	513
#m	2420.22144	805.86566	958
#n	2375.85278	962.65698	525
#j	1307.11670	798.37451	285
#z	891.04468	438.68832	212
#y	3057.30884	1225.70630	335
#w	3554.06738	1391.89587	153
#r	1172.61877	532.81390	1484

表4 正規化に用いたパワーの範囲
(p-pRMS値)

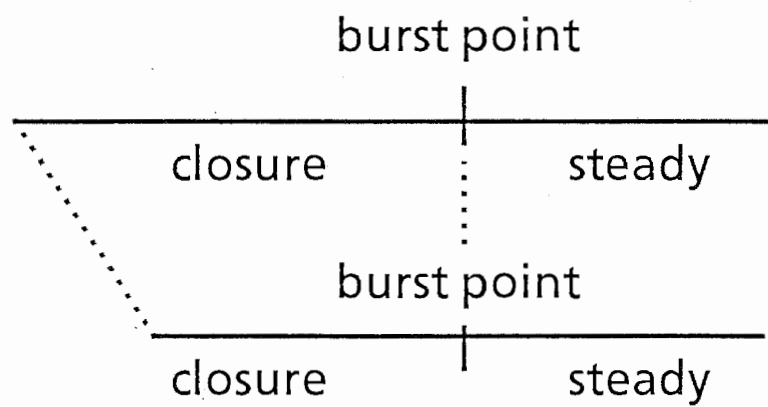
Phoneme	Max	Min
#a	3000	2000
#i	2000	1800
#u	2000	1800
#e	3000	2800
#o	2500	2300
#N	2000	1800
#b	1000	800
#d	1000	800
#g	2000	800
#m	2000	1800
#n	2000	1800
#j	1200	600
#z	800	600
#y	2000	1800
#w	2500	2300
#r	1200	1000



☒ 1 Data flows in the system



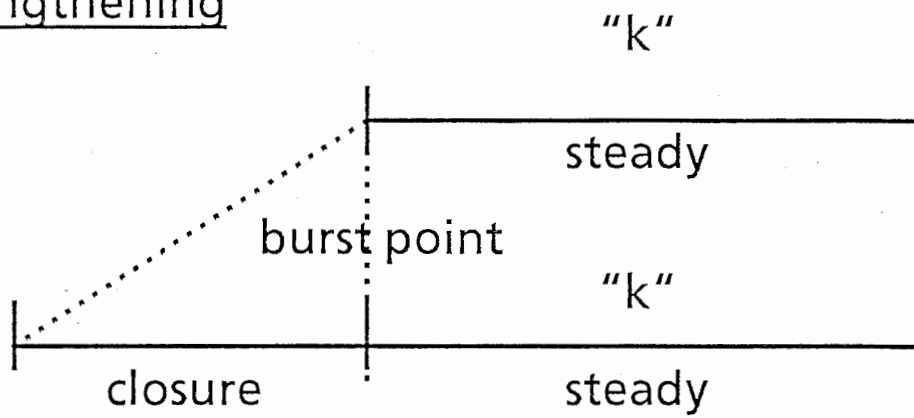
lengthening



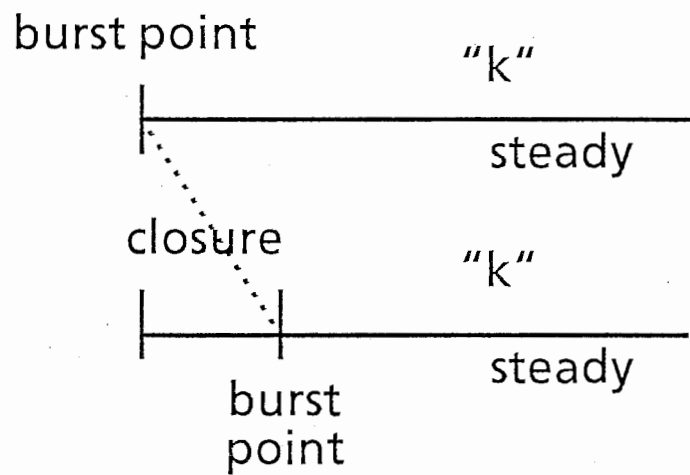
shortening

図2 Duration Control Algorithm

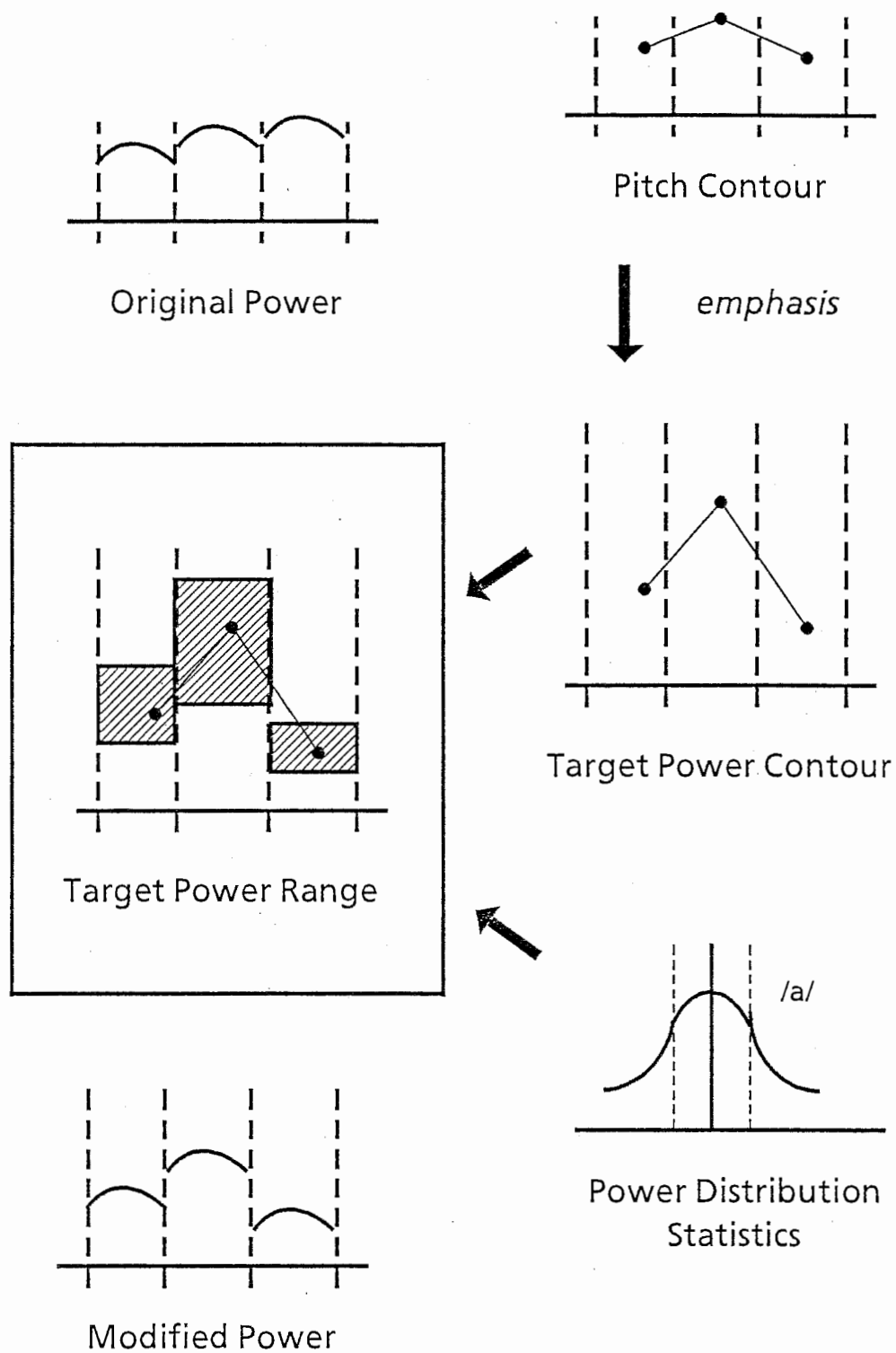
Lengthening



Shortening



☒ 3 Pause Insertion



☒ 4 Power normalization

APPENDIX

実験に使用した例文

例文番号	例文
001	ぶれいな くちを きく
002	ちょうだの れつが できた
003	さいごの つめが あまい
006	かかりの しじに したがう
008	なまえを おおきく かく
010	じけんの はいごを あらう
011	ひとりで しあんに くれた
013	ないぶの じじょうに あかるい
014	たんなる うわさに すぎない
015	じゅうしょを かくのを わすれた
018	さいごの きりふだを だす
019	やぶんの ほうもんを わびる
020	じけんの しんそうに せまる
021	らいげつは ぎょうじが おおい
022	かがいしゃが すがたを けした
023	そしきの かいかくに のりだす
024	じけんの てがかりを つかんだ