ヒューマンインタフェース 44－7
（1992. 9. 10）

話速変換に伴う時間伸張を吸収するための一方法

池沢 龍　中村 章　清山 信正　都木 徹　宮坂 栄一

*NHK*放送技術研究所

話速変換システムにおいて話速（話す速さ）を遅くする際，発話時間が伸張し，必然的に実時間に発声される時間との「ずれ」が問題となる．これを解決するために，文章間の無音区間を聴感上，違和感なく最短に短縮し，かつ，話速を固定ではなく，ピッチの大まかな変化に追随するよう，声立てと次の声立ての区間を単位にして，この区間の開始点では話速を遅くし，終了点に向かって徐々に話速を速める手法を開発した．これにより，発話時間を原音声の発話時間に保ったまま，聴きやすいゆっくりした音声に変換することが可能となった．

A METHOD OF ABSORBING TEMPORAL ENLARGEMENT OF SPEECH LENGTHS IN THE VOICE SPEED CONVERTING SYSTEM FOR ELDERLY

Ryou Ikezawa Akira Nakamura Nobumasa Seiyama Tohru Takagi Eiichi Miyasaka

***NHK* Science And Technical Research Laboratories**

1-10-11, Kinuta, Setagaya-ku, Tokyo 157, Japan

Only voice speed can be made slower than normal at a constant rate with other features(such as pitch, personality) held original in the voice speed converting system which has recently been developed by ***NHK***. This constant conversion in speed causes temporal enlargement of speech lengths. This paper presents a new algorithm to absorb temporal discrepancy in length between the converted speech and the original one, no matter how the processed speech can perceptually sound slower.

The algorithm has two characteristic features; (1) Every long pause between adjacent uttered sentences is shortened with perceptual naturalness. (2) Voice speed set to be slow at the onset of voicing is made faster step by step roughly along a curve of the envelope of pitch frequencies, resulting in a value of speed faster than normal.

1 まえがき

筆者らは高齢者にも良好な音声放送サービスをめざして，高品質リアルタイム話速変換システムを開発した(*1)．このシステムの特徴は，話速を遅くすることによって高齢者の識別速度の劣化を補償することにあるが，発話時間が伸張するため、必然的に実時間に発声される時間との「ずれ」が生じる．発声される音声が長時間にわたる時には，十分長い無音区間などが数多く存在するので，これを短縮することで吸収できるが，この方法では，ＴＶなど映像を伴う場合，その「ずれ」が気になることも考えられる．今回，この時間の「ずれ」をできるだけ短時間内に吸収するため，文章間の長い「ま」を短縮し，かつ，話速を一定の規則で変化させる方法を検討し，心理実験によってその有効性を確認した(*2)．

2 本方法の特徴

高品質リアルタイム話速変換システムでは，入力された音声を無音，無声，有声の３つの区間に識別し，無音区間，有声区間をそれぞれ独立に制御（伸張，短縮）することができる．

本方法の特徴は，

(1) 有声区間における大まかなピッチ変化に基づいた話速制御

(2) 文章間の長い「ま」（無音区間）の制御

の２つの方法を組み合わせることにより，発話時間の伸張時間を最小限に抑えることを可能にしている．

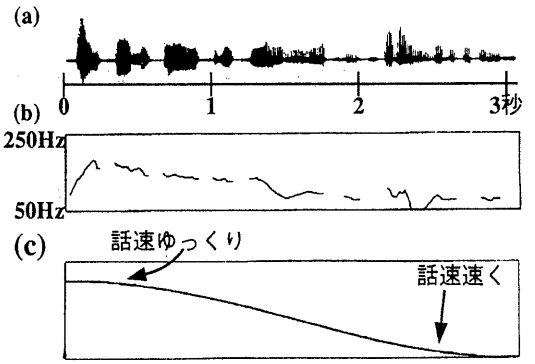


図1　あるニュース文の(a)波形,(b)ピッチ変化と(c)ピッチ変化に基づいた話速制御

2.1 有声区間における大まかなピッチ変化に基づいた話速制御

一般に意味上の重点がある箇所をゆっくり話すことによって聴きやすさが向上することが考えられる．また，意味上の連続と区切りは息つぎなどの有無の他に，ピッチの変化で，表現される場合が多いことが報告されている(*3)．

図1(a)(b)に，あるニュース文の波形とその

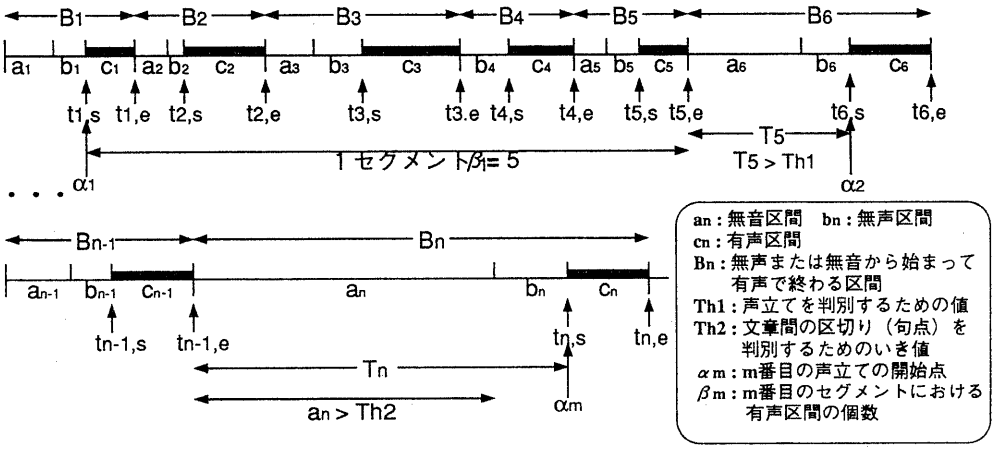


図2　処理に基づく音声データのセグメンテーション

ピッチ変化を示す．息つぎ直後は，ピッチが高く，次第にピッチが下降する．同時に音声のパワーもはじめ大きく，次第に小さくなる．そこで図1(c)に示すように，ピッチの大まかな変化に話速が追随するよう，声立てと次の声立ての区間を１単位（セグメント）として，息つぎ直後のピッチの高い区間をできるだけゆっくりと話速変換し，後半のピッチが低く，パワーも小さい部分の話速を速める．この方法により，以下の効果が期待できる．

(1) ピッチの高い部分は，比較的意味の重要性が高いと考えられるので，ある程度，意味上の重要箇所の話速をゆっくりすることができる．

(2) 息つぎ直後の話し始めをゆっくりさせることにより，受聴者にゆとりを持って聴く「手がかり」が与えられる．

(3) 一般に通常のニュース音声などの場合も，一息で発声される区間内で，話速は始めゆっくりで後半，速くなる傾向がある．従って一律に話速を遅くすると一本調子になることが多いのに対し，本方法を用いると，テンポにゆらぎを持たせることになるため，聴感上聴きやすくなる．

次に，図2に従って具体的な処理方法を説明する．

(1) 無声または無音から始まって有声で終わる区間を１ブロック（B_n：n=1,2,…）とする．このブロック内では、無音区間（a_n），無声区間（b_n），有声区間（c_n）の３つに識別される．b_nとc_nの境界，すなわちｎ番目の有声区間の開始点を$t_{n,s}$，その終了点を$t_{n,e}$

$$t_{n,e} = t_{n,s} + c_n$$

と表現し，m番目の声立ての開始点をα_mとする．

(2) ｎ番目の有声区間の開始点（$t_{n,s}$）と１つ前の有声区間の終了点（$t_{n-1,e}$）との間の時間間隔

$$T_n = t_{n,s} - t_{n-1,e}$$

を算出する．

(3) T_nがある閾値Th1を越えた場合には，$t_{n,s}$の時点を声立てα_mと判断する．１つ前の声立てα_{m-1}と１つ前の有声区間の終了点$t_{n-1,e}$の範囲を一つのセグメントとする．図2の例では，

$$T_5 = t_{6,s} - t_{5,e} > Th1$$

とすると，$t_{6,s}$の時点が声立てα_2，$t_{1,s}$から$t_{5,e}$までの区間が１セグメントとなる．この時，m番目のセグメントにおける有声区間の個数をβ_mとするとβ_1=5となる．

(4) 開始点の有声区間の伸張倍率r_sを，$r_s \geq 1.0$の範囲内の値に設定する．この伸張倍率をセグメント終了点に向かって徐々に小さくし，終了点

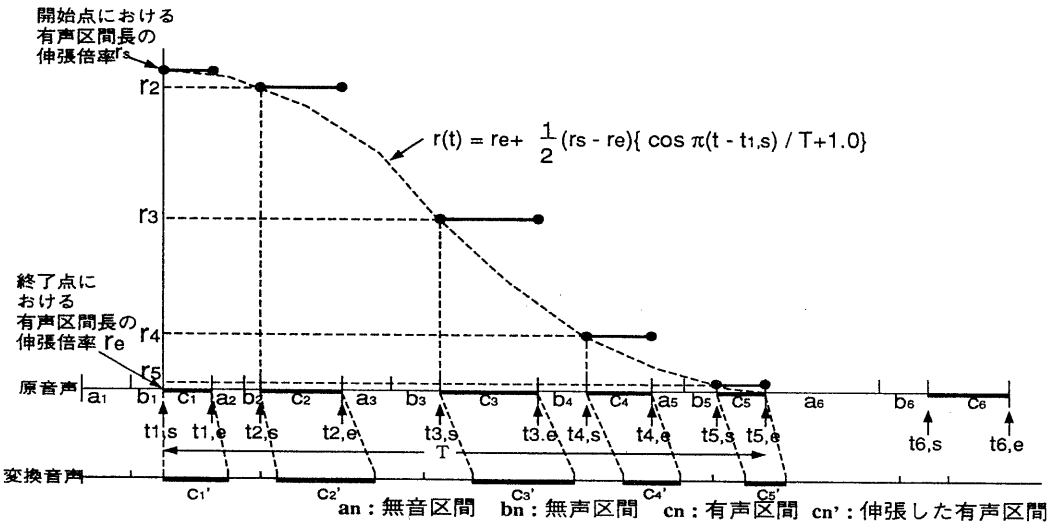


図3　有声区間の伸張倍率の求め方

の有声区間の伸張倍率r_eが$r_e \leq 1.0$となるようにする．図3に図2のセグメント1に属する有声区間の伸張倍率の求め方を示す．セグメント開始点の有声区間c_1は伸張されて $c_1' = r_s \times c_1$ 次の有声区間c2は，$c_2' = r_2 \times c_2$ となる．セグメント終了点における有声区間c_5は，$c_5' = r_5 \times c_5$ となるが，r_eは$r_e \leq 1.0$であるから，$r_e < 1.0$では原波形より短縮される．有声区間の伸張倍率の変化は，開始点から終了点に向かってなだらかに単調減少する以下の関数r(t)を用いる．

$$r(t) = r_s + 0.5 \times (r_s - r_e)\{\cos \pi (t - t_{n,s})/T + 1.0\}$$

但し、$t = t_{n,s} \sim t_{n+\beta m-1,e}$

なお，有声区間以外の無声区間a_n，無音区間b_nについては処理を施さないため，a_n,b_nの長さは不変である．

2.2 文章間の長い「ま」の制御

無音区間がある閾値Th2を越えた場合，この無音区間を文章と文章の区切り（句点），つまり，文章間の長い「ま」と判断し，この無音区間の一部を削除し，全体の長さを短縮する．この無音区間の時間長をa_n，無音区間内で削除する区間の時間長をd_n，削除後の無音区間の時間長をe_nとした場合，e_nは図4に示すように

$$e_n = a_n - d_n$$

となる．この際，分析時の無音区間の指定誤りから，無声区間までも長い無音の一部と識別してしまう可能性があるため，a_nの先頭から，d_nを削除するのではなく，a_nの中心点からd_n部分を削除する．また，d_nの両端には、10msのテーパーをかけてクリックの発生を防ぐ処理を施す．

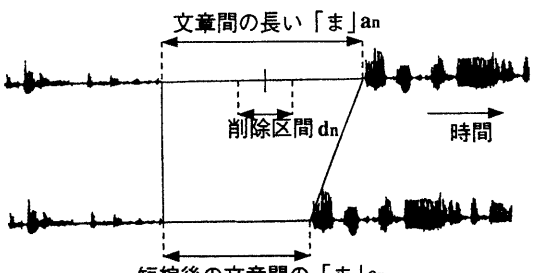


図4 文章間の長い「ま」の短縮方法

3 聴きやすさの許容範囲の測定

代表的なニュース文を素材にして話速の制御，長い「ま」の短縮の許容範囲を心理実験によって求めた．

3.1 実験1 有声区間における ピッチ変化に基づいた話速制御

2.1で提案した方法により，開始点の有声区間の伸張倍率r_s，終了点の有声区間の伸張倍率 r_eを変え，不自然に聴こえない伸張倍率の範囲を求める実験を行った．

ここでは，声立ての時間を決定するための閾値Th1を予備実験の結果から，

$$Th1 = 350ms$$

に設定した．海木[*4]らが，ポーズ時間長には単峰性，双峰性の２つの分布があることを報告しているが，今回，実験で用いた350msの長さは，この双峰性（２つのピークを持った）分布のうち，ポーズ長の長い方のピークにおけるポーズ長とほぼ一致している．

3.1.1 刺激音

付録に示した男性アナウンサーが発声したニュース音声

（長さ2分13秒，Air収録，7.9モーラ／秒）

3.1.2 実験方法

付録に示したニュース音声(1)-1~(6)-3について開始点における有声区間の伸張倍率r_sを1.4,1.3,1.2,1.1,1.0倍，終了点における有声区間の伸張倍率r_eを 1.0,0.9,0.8,0.7倍，開始点から終了点までの伸張倍率は関数r(t)によって変化させた．これら24の組み合わせがランダムに提示されるように刺激を作成し，表1(a)(b)に示す２通りの５段階の評価を行った．原音声と比較して処理音声の自然性，聴きやすさの劣化の程度を

知る必要から，被験者は健聴な20,30歳代の男女5人を用いた．実験は6回繰り返した．刺激音の提示は，スピーカー再生で提示音圧レベルは約70dB(A)である．

表1　実験1,2で用いた5段階評価

	1	2	3	4	5
(a)	不自然で非常に気になる	不自然で気になる	不自然でやや気になる	少し不自然だが気にならない	自然である
(b)	意味がわかりにくい	意味がややわかりにくい	どちらともいえない	意味がややわかりやすい	意味がわかりやすい

3.1.3 実験結果及び検討

図5に実験結果を示す．同図は、横軸が開始点の有声区間の伸張倍率r_s，縦軸が終了点の有声区間の伸張倍率r_e，□は『自然性』，○は『わかりやすさ』の評価の度合いを径の大きさで表したものである．また，$r_s = r_e$は、話速が一定であることを意味する．全体として以下の傾向が見られる．

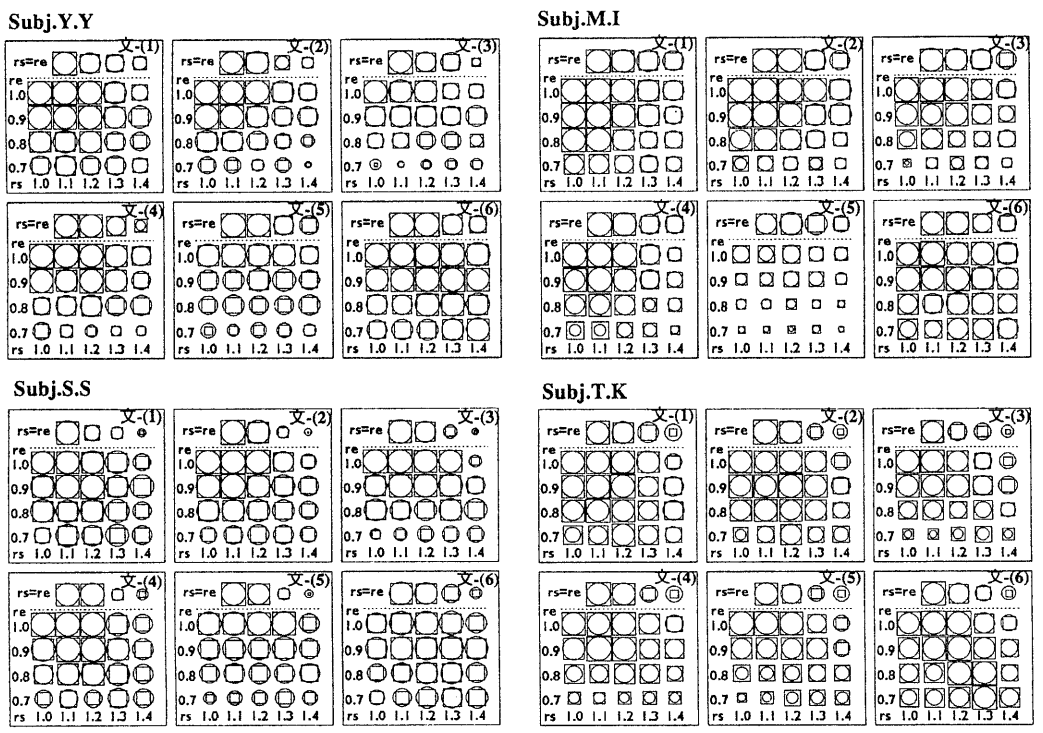
(1) どの文章の場合も，開始点の話速が1.0～1.3倍で終了点の話速が0.9～1.0の範囲では，『自然性』，『わかりやすさ』の両方について高い評価である．

(2) 話速が1.4倍で不変の時には，評価が低く，前回の結果[*5]と同様の結果が得られた．しかし，終了点の話速を1.0～0.9倍に速めると評価が上がり，さらに話速を速めて0.7倍になると，再び評価が落ちる．このように，開始点の話速が1.4倍の時には適度に話速を速めていくことによって聴感上聴きやすくなる．

(3) 表2に示したように『自然性』と『わかりやすさ』の間には比較的高い相関がある．

表2　実験1における『自然性』と『わかりやすさ』の相関

Subj(age)	Y.Y (28)	M.I (24)	S.S (28)	T.K (23)	R.I (33)
相関	0.67	0.80	0.77	0.81	0.83



横軸：開始点の伸張倍率　縦軸：終了点の伸張倍率
点線の上部は，話速が一定の時の評価を表す．
□：『自然性』の評価　○：『わかりやすさ』の評価

評価　5　4　3　2　1
(径が小さくなるにつれて評価が下がる)

図5　話速を変化させた時の『自然性』と『わかりやすさ』の評価

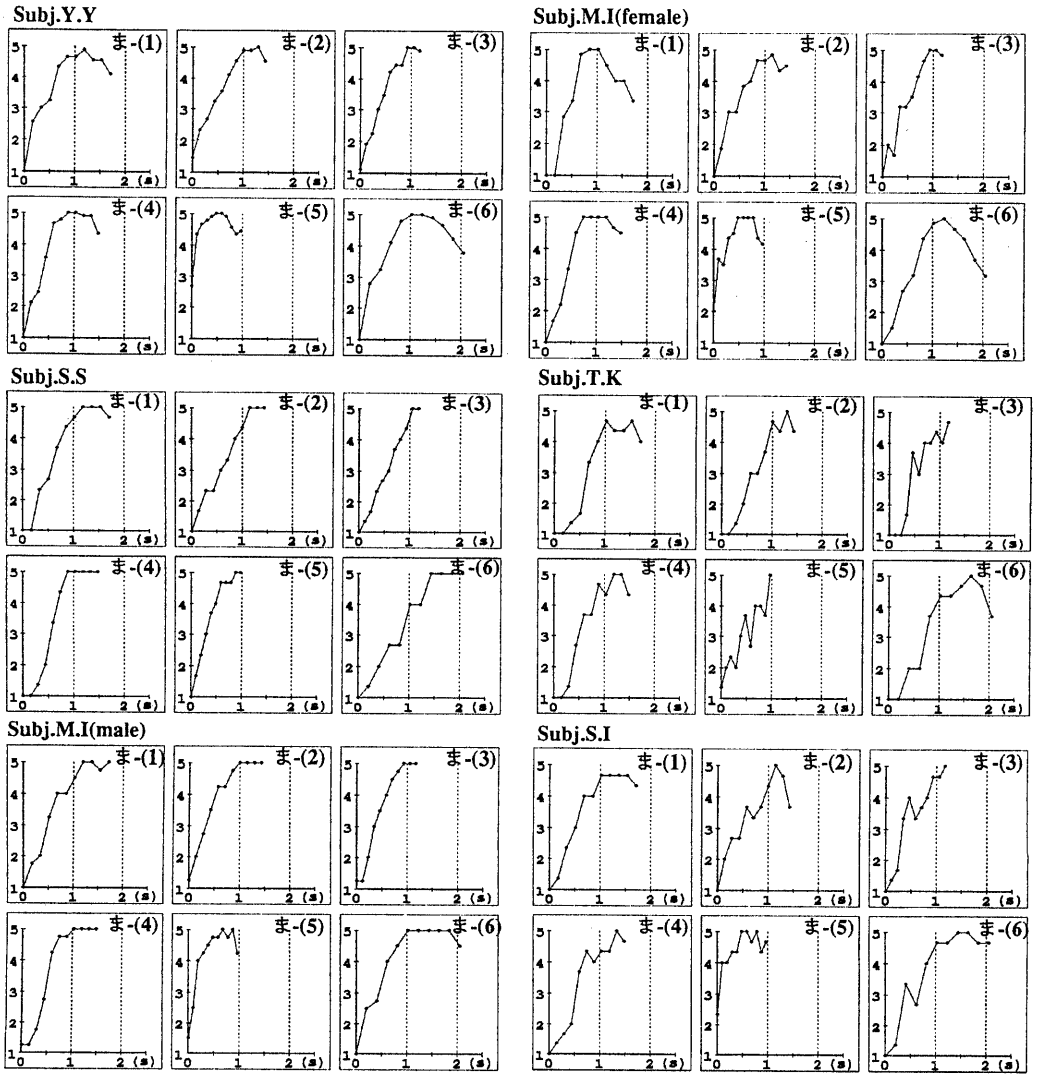


図6　「ま」を短縮した時の『自然性』の評価　（横軸：文章間の「ま」の長さ　縦軸：評価値）

表3　評価4と5の中間値を閾値とした場合の「ま」の長さ(ms)

Subj(age)	ま-(1)	ま-(2)	ま-(3)	ま-(4)	ま-(5)	ま-(6)	平均*
Y.Y(28)	680	860	850	580	100	750	744
M.I(24)	540	820	800	600	400	900	732
S.S(28)	930	1050	950	790	560	1200	984
T.K(23)	980	980	1020	950	900	1320	1050
R.I(33)	620	750	750	580	250	950	730
A.N(31)	840	950	800	850	430	960	880
M.I(67)	1000	800	700	670	400	820	798
S.I(59)	980	1060	900	1000	410	960	980
平均	821	909	846	753	431	983	862

*ま-(5)は、除く

3.2 実験2

文章間の「ま」（無音部分）の短縮

文章間の「ま」の長さは，ほぼ1秒以上あることが，予備実験より得られたので，文章間の「ま」と判断する閾値Th2をTh2=1000msに設定し，この文章間の「ま」を短縮しても不自然に聴こえない範囲を求める実験を行った．

3.2.1 刺激音

実験1と同一のニュース音声

3.2.2 実験方法

付録に示したニュース音声における6つの長い「ま」（ま-(5)のみ読点，それ以外は句点）の各々について，短縮率0%（原音声）から10%毎に100%（「ま」の削除）までの11通りの「ま」を有するニュース音声を作成し，これをランダムに提示して『自然性』についての5段階の評価を行った．実験は3～6回繰り返した．被験者は20歳代から60歳代の男女8人である．他の条件は実験1と同様である．

3.2.3 実験結果及び検討

図6は，6人の実験結果で，評価の平均値を「ま」の絶対時間の関数として表したものである．また，表3は評価4と5の中間値を自然に聴こえる閾値とした場合の「ま」の時間長である．これらから，以下の傾向が見られる．

(1) 約1000msまで「ま」の長さを短縮しても聴こえの自然性は保たれる．

(2) 被験者によっては「ま」が長すぎるとかえって不自然に聴こえる場合もあり，原刺激の「ま」を短縮することによってより自然性が増加することがある．

(3) 読点（ま-(5)）の場合は他の句点に比べて「ま」の長さが短くても不自然に聴こえない．

4 本方法による時間伸張吸収の効果

ここでは，本手法による時間伸張吸収の効果について検討する．まず，話速制御のみによる時間伸張の吸収率を表4(a)に示す．なお，吸収率は，次式で定義される．

$$\text{吸収率}=\frac{(\text{一律の場合の伸張時間})-(\text{変化させた場合の伸張時間})}{\text{一律の場合の伸張時間}}\times 100\ (\%)$$

原音声136秒に対し，有声区間のみを一律に1.2倍時間伸張させると146.4秒となり，約10秒ほど長くなる．これを本方式に従って，話速の伸張倍率を1.2倍から，0.9倍に可変速にすると，全体の伸張時間は，2.6秒におさまり，一律の場合の10.4秒に比べると，伸張した分を75%程度吸収したことになる．

次に話速制御と長い「ま」の短縮の両方の操作を行った時の時間伸張の吸収率を表4(b)に示す．同表から自然性，わかりやすさを保った状態で伸張を吸収する代表的な2例について検討する．

表4(a)　話速制御のみによる時間伸張の吸収率

時間伸張倍率の変化	時間長	伸張時間	吸収率
ｵﾘｼﾞﾅﾙ（原音声）	136秒	———	———
一律に1.2倍	146.4秒	10.4秒	———
1.2倍 →0.9倍	138.6秒	2.6秒	74.7%
1.2倍 →0.8倍	136.3秒	0.3秒	97.3%
一律に1.3倍	152.1秒	16.1秒	———
1.3倍 →1.0倍	144.7秒	8.7秒	55.9%
1.3倍 →0.9倍	142.1秒	6.1秒	62.4%
1.3倍 →0.8倍	139.7秒	3.7秒	77.3%

表4(b)　本手法による時間伸張の吸収率

時間伸張倍率の変化	短縮後の「ま」の時間長	伸張時間	吸収率
1.2倍 →0.9倍	1.4秒	0.4秒	97.5%
	1.3秒	-0.1秒	100.6%
	1.2秒	-0.6秒	103.7%
1.3倍 →0.9倍	1.2秒	2.8秒	82.6%
	1.1秒	2.3秒	85.7%
	1.0秒	1.8秒	88.8%

(1) 開始点の有声区間の伸張倍率r_s=1.2倍の場合

終了点の有声区間の伸張倍率r_eを0.95倍，文章間の「ま」を1200msまで短縮した時に，吸収率は100%となる．

(2) r_s=1.3倍の場合

r_e=0.9倍，文章間の「ま」=1000msまで短縮した時に，吸収率は89%となる．この時，136秒の文に対し，1.8秒の時間伸張となり，完全な吸収はできない．しかし，話題が異なる時の3,4秒程度のより長い無音区間を短縮することで吸収は可能になる．

5 結論

話速変換に伴う時間伸張を短時間内に効果的に吸収する方法を提案した．今後，以下に述べる課題を解決する必要がある．

(1) 高齢な被験者を対象に，種々の音声に対して本手法の効果を検証する．

(2) 任意の話速の音声についても適応可能となるように話速変化の許容範囲を求める．

(3) 本手法では，短時間内に時間伸張を吸収することが可能なため，音声と映像の「ずれ」は少ないと考えられる．また，積山ら(*6)によって，日本人はアメリカ人より読唇依存性が低いことが報告されていることから，「ずれ」による不自然性は少ないと予想されるが，本手法によるこれらの「ずれ」の評価実験を行う．

(4) 今回は，一息で発声される区間を前もって抽出し，その区間情報をもとに話速を変化させているが，話速変換装置に組み込むにはリアルタイム化が必要で，リアルタイム吸収方法について検討する．

参考文献

*1)中村 章ほか, 日本音響学会春季大会,p.329-330 (1992)
*2)池沢 龍ほか, 日本音響学会春季大会,p.331-332 (1992)
*3)杉藤 美代子, 樟蔭国文学 第23号, p. 92-110 (1986)
*4)海本 延佳ほか, 日本音響学会春季大会,p.251-252(1992)
*5)中村 章ほか, 日本音響学会秋季大会,p.381-382 (1991)
*6)積山 薫ほか, 日本音響学会秋季大会,p.401-402 (1991)

付録

実験1,で用いたニュース文

文(1)-1 神奈川県の大和市と綾瀬市にまたがる
厚木基地の周辺住民が、
-2 基地を使用する米軍機などの騒音によって
-3 精神的、肉体的に大きな被害を受けたとして
-4 国を相手におこしていた第２次厚木基地訴訟
の最終弁論が、
-5 今日、横浜地方裁判所で行われ、
-6 提訴以来７（ヒヂ）年にわたって争われていた
裁判が結審しました。

ま-(1) (1709ms)

文(2)-1 この厚木基地の騒音をめぐる訴訟は
-2 第一次の訴訟をおこし、
-3 一審で損害賠償については一部認められた
ものの、
-4 飛行の差し止めについては却下され、
-5 二審では損害賠償についても退けられたため、
-6 最高裁判所で現在争われています。

ま-(2) (1436ms)

文(3)-1 今日の第二次訴訟は、一審で飛行の差し止めが
認められなかったことから
-2 厚木基地周辺の６つの市に住む住民１５６が、
-3 基地を管理する国を相手取って
-4 総額６億７千万円の損害賠償と
-5 夜８時以降の飛行の停止などを求めて
-6 新たに訴えをおこしていたものです。

ま-(3) (1169ms)

文(4)-1 今日は午前１０時半から横浜地方裁判所で
-2 まず原告側の最終弁論が行われ、
-3 この中で住民側は、
-4 アメリカの空母ミッドウエイとそのあとを引き
継いだインディペンデンスの
-5 夜間発着訓練によって
-6 夜間の騒音が飛躍的に増え、
-7 ビルの工事現場に相当する激しい騒音などが
繰り返されて、
-8 難聴などの被害がでており、
-9 著しく犯されていると改めて訴えました。

ま-(4) (1487ms)

文(5)-1 これに対して被告の国側は、続いて行われた
最終弁論の中で、
-2 厚木基地の管理や運営は、
-3 自衛隊法や日米安全補償条約に基づいて
行われており、
-4 高度の政策的判断を必要とする統治行為で
-5 原告の訴えは法律上成り立たない、

ま-(5) (980ms)

文(5)-6 また、基地の騒音自体もがまんできる限度
-7 受認限度を越えていないなどと述べて
-8 訴えを却下するよう求めました。

ま-(6) (2040ms)

文(6)-1 提訴以来、７年に渡って争われていたこの裁判
は、今日の原告被告双方の最終弁論で結審し、
-2 判決は来年秋ごろまでには言い渡されるものと
見られていますが、
-3 どこまで認めるか、注目されています。