社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59 (34)
2005／12／22

音声/非音声判別法を用いた時間圧縮音声再生法

竹内 伸一†, 杉山 雅英†

†会津大学 大学院 コンピュータ理工学研究科
〒965-8580 会津若松市一箕町
E-Mail: †{d8051105,sugiyama}@u-aizu.ac.jp

あらまし

近年 HDD レコーダや DVD レコーダ等の映像録画機器の普及に伴い，大量の映像・音声に関するマルチメディアの蓄積が容易になっている．データの蓄積は専用機械によって自動的に行われる一方データの視聴は人間が行わざるを得ないため，大量のデータを視聴する場合には視聴速度の向上が求められる．本報告では元となるマルチメディアデータの音声データに着目し，データ中の音声区間の再生を主とした時間圧縮音声再生法について提案する．提案手法は我々がこれまで提案してきた特徴量 Block Cepstrum Flux を用いた音声/非音声判別手法を用いて対象となる音声区間を検出した後に定常部削減を行う，音声/非音声のパターン認識に基づく時間圧縮音声再生法である．元となるデータから非音声区間を取り除くことにより音声区間を残すことになるため，一律に圧縮した場合よりも話速が遅くなり，より聞き取り易い圧縮音声を生成することが可能となる．圧縮音声を視聴した主観評価実験の結果はデータを一律に圧縮する PICOLA 法を単独で用いた場合よりも良好な結果を得ることができ，提案手法と PICOLA 法を併用することも可能であることを示した．

キーワード: 音響特徴量，区間検出，早聞き，非線型時間圧縮

Time-compressed speech playing method using voice/non-voice classification

Shin'ichi TAKEUCHI†, Masahide SUGIYAMA†

†Graduate School of Computer Science and Engineering, The Univ. of Aizu
Ikki-machi, Aizu-Wakamatsu, Fukushima, 965-8580, Japan
E-Mail: †{d8051105,sugiyama}@u-aizu.ac.jp

Abstract

Recently, the effectiveness of audio-visual machine improves and they becomes tc be able to storage many multimedia data. Although to storing data is done by machine automatically, to watch data is done by human and it is bottle-neck for improvement of multimedia data processing. This report attentions to sound part of multimedia data and proposes time-compressed speech playing method. The proposed method uses feature parameter Block Cepstrum Flux we have proposed and it can classify voice and non-voice section in sound data. The Proposed method picks out voice section and deletes continuous part. By to remove non-voice section from sound data, it can create compressed data with more slowly voice than the compressed data with constant compress rate. The experimental result for compressed sound listening test shows that the proposed method brings more better impression than constant compress rate.

Keyword: feature parameter, section detection, fast listening, non-linear time compress

1 まえがき

近年 HDD レコーダや DVD レコーダ等の映像録画機器の普及に伴い，大量のマルチメディアデータの蓄積が容易になっている．データの蓄積は専用機械によって自動的に行われる一方，データの視聴は人間が行わざるを得ないため，大量のデータを視聴する場合には視聴速度の向上が求められる．

視聴速度の向上策として，映像に関しては一定間隔でデータを間引く簡単な手法で実速以上での処理が可能である．一方音声に関しては同様に一定間隔でデータを間引く手法の他に，音声のピッチ周期性に着目して波形を編集する TDHS 法 [1] や PICOLA 法 [2]，音声波形に含まれる無音等の定常部分を削減することでデータ量の圧縮を図る MFCD 法 [3] 等が提案されている．データ中の音声区間に着目した場合，[1, 2] は波形編集に処理時間を必要とし，[3] は定常部の削減のみに着目している．また両手法とも音声/非音声を一律に圧縮するため音声区間の話速は速くなるので，一定以上の圧縮率を得ることは困難である．

本報告では，元となるマルチメディアデータの音声データに着目し，データ中の音声発話区間の再生に着目した時間圧縮音声再生法について提案する．我々は音声/ 音楽区間やそれを拡張した音声/非音声区間の判別を目的とする研究 [4] を行っており，これまでに音声/音楽の判別に効果的な特徴量 Block Cepsrum Flux (BCF) を提案し [5]，その有効性を示してきた．提案する時間圧縮音声再生法はこれまで提案してきた BCF を用いた音声/非音声判別手法を用いて対象となる音声区間を検出し，非音声区間を除外した後に定常部削減を行う，音声/非音声のパターン認識に基づく時間圧縮音声再生法である．本報告は以下の様に構成されている．2 節で提案手法について述べる．3 節で手法の評価実験を行い，4 節でまとめを述べる．

2 時間圧縮音声再生法

この節では時間圧縮音声再生法と，その元となる音声/非音声の判別手法について述べる．提案手法は予め非音声区間を削除することで，求める圧縮音声中の音声部分の圧縮率を緩和することが可能である．

2.1 Block Cepstrum Flux

式 (1) は音響データ内の J フレーム間のケプストラムベクトルを基準フレームから相互に比較する Cepstrum Flux と呼ばれている．Block Cepstrum Flux (BCF) は Cepstrum Flux を W フレームで平均化したものであり，式 (2) で定義される．

$$D_n(J) = \frac{1}{J}\sum_{j=1}^{J} d^2(\boldsymbol{c}_n, \boldsymbol{c}_{n-j}) \tag{1}$$

$$B_n(W) = \frac{1}{W}\sum_{i=0}^{W-1} D_{n-i}(J) \tag{2}$$

J は時間軸での幅 (フレーム数)，$d^2(\boldsymbol{c}_n, \boldsymbol{c}_{n-j})$ は n フレーム目と $n-j$ フレーム目のケプストラムベクトル間の距離をそれぞれ表す．BCF は 2 点間での対数スペクトルの距離を平均化したものであるため，スペクトルの変動が大きい区間 (音声区間) では値が大きく，変動が小さい区間 (音楽区間等) では値が小さくなる．従って適切な閾値を設定することで，BCF の値が閾値よりも大きい時に音声，小さい時に非音声区間と判別することができる．$J = 3, W = 4$ の時の Cepstrum Flux，BCF の概念を 図 1 に示し，図 2 に 3.1 で述べるデータ C の音声及び非音声区間から計算された BCF の値の分布を示す．

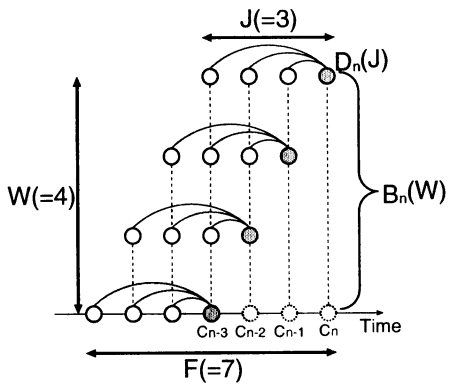


図 1: Cepstrum Flux と Block Cepstrum Flux

$B_n(W)$ はフレーム毎に決定することを目的としているが，一方式 (3) で定義される区間毎に音声/ 非音声を決定するための区間 F 毎の BCF を提案した [4]．

$$B^*_m(F) = \frac{1}{F-J}\sum_{i=mF+J}^{(m+1)F-1} D_i(J) \tag{3}$$

式 (2) の BCF は各フレーム毎に計算されるので計算用区間のオーバーラップがあるのに対して 式 (3) の BCF は窓のオーバーラップが無い．両者の区別の為に 式 (2) を Frame BCF，式 (3) を Segmental BCF と呼ぶ．

2.2 再生区間検出法

これまでに Segmental BCF を用いた音声/非音声の判別法 [4] を提案した．本報告ではその判別法を条件を変化

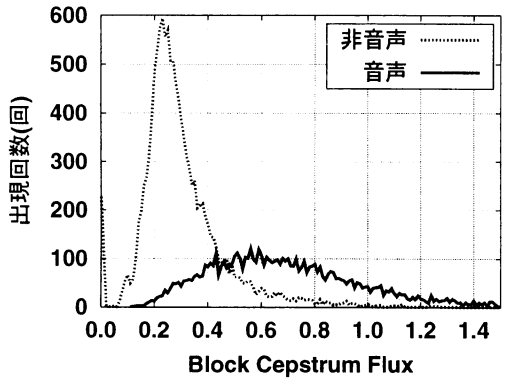


図 2: 音声/非音声区間の BCF の分布

させた二段階で適用し，圧縮音声を作成する．図 3 に全体の流れ図を示す．

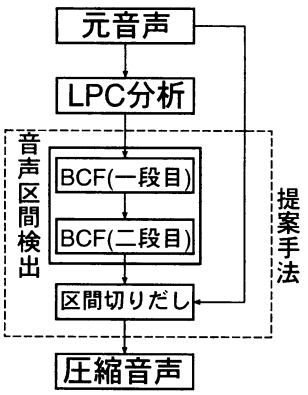


図 3: 提案手法の処理の流れ

LPC 分析の後，一段目として Segmental BCF で求められる音声/非音声の判別結果を用い，式 (4) 前半で示される閾値 T_1 との比較を行い条件を満たす区間を次段に渡す．二段目としてその区間内の各点に対して式 (4) の後半にあるように Frame BCF と閾値 T_2 との大小比較で音声として再生する区間を決定する．

$$\begin{cases} B_m^*(F) \geq T_1 \\ B_n(W) \geq T_2 \end{cases} \tag{4}$$

BCF による音声/非音声の判別は，窓長が大きい場合には値が平均化されるために判別性能が向上する．音声部分の欠落を避けるために T_1 を低めに設定した 式 (4) を用いることで Frame BCF を単体で用いるよりも音声/非音声判別性能が向上することが分かっている [4]．また音声区間内でも母音は子音に比べてスペクトルの変化が小さいので BCF の値も小さくなることから，短時間での変動を求めるために窓長を小さく設定した Frame BCF と閾値 T_2 との大小比較を行うことで母音区間内の冗長な部分の間引きが可能となる．対象の音声データから非音声区間及び母音の冗長な部分の削減を行ってデータ量を圧縮することにより，実速以上での再生が可能となる．また提案手法は区間検出のみを行い波形変換を行わないため，極めて高速で動作することが可能である．

2.3 圧縮率による閾値 T_2 の自動決定

式 (4) 中で用いられる BCF に対する 2 つの閾値 T_1 及び T_2 は，これまで予備実験で求めた値を用いてきた [6]．圧縮率は 2 つの閾値の組合せで決定されることになる．ここで圧縮率は圧縮処理後の音声長を元の音声長で割った値である．実際に利用する場合を想定して指定された圧縮率となる T_2 の与え方を述べる．T_1 に関しては音声/非音声の大別に用いるので，従来の実験から安定動作すると考えられる値を用いることとし，T_2 を圧縮率から以下の手順で求める．

0. 圧縮率 $(0 \leq x \leq 1)$ を指定する
1. Segmental BCF を計算し，$B_m^*(F) \geq T_1$ を満たす区間を切り出す
2. 圧縮率 x を用いて切り出したフレーム数と目標とするフレーム数との比 a を決定する

   $a = x \times$元データのフレーム数/切り出したフレーム数
3. 切り出した各フレームについて Frame BCF を計算し，その値 b_n を度数とするヒストグラム h_n を作成する
4. h_n を低い度数から累積させた h_n^* を計算する

   $h_n^* = \frac{1}{N}\sum_{m=0}^{n} h_m \quad (\text{ただし } N = \sum_{m=0}^{\infty} h_m)$
5. $1 - h_n^* \leq a$ を満たす最小の b_n を T_2 とする

3 評価実験

評価実験として提案手法によって圧縮されたデータを 10 代から 20 代の男女 9 名の被験者が『音声の聞き取り易さ』に留意して試聴した後に表 1 に表された 5 段階で主観評価 (MOS) を行った．

評価の際には無圧縮時の音声を 5 として評価の基準とするよう指示してある．実験に使用したデータは閾値 T_1 の値を予備実験の結果から音声区間を誤って除去することの無い安全な値 $T_1 = 0.26$ と定め，圧縮後のデータ量

をそれぞれ 1/2 から 1/5 となるよう閾値 T_2 を決定し圧縮を行った.

実験では ① 提案手法, ② 一段目に Segmental BCF を用いて非音声区間の削除を行った後に PICOLA 法を用いて規定のデータ量になるまで圧縮した手法, ③ PICOLA 法の 3 種類の比較を行った. 手法 ② においては PICOLA 法の圧縮率パラメータを a とし, 手法 ③ においては圧縮率を x と指定することで圧縮音声を生成した.

表 1: 5 段階評価

5	細かなニュアンスまでわかる
4	内容が把握できる
3	雰囲気はわかる
2	長時間聴くのはつらい
1	聴くに耐えない

3.1 受聴評価用データ

評価には音声/非音声区間を含むデータとして発話形式の異なる 3 種類のデータベースを用いた. データ A は発話形式が朗読である「五体不満足」データベース, データ B は発話形式が出演者のセリフである会津若松市の観光案内ビデオの音声データ, データ C は発話形式がパーソナリティの会話および CM である CampusWave Database [7] である. それぞれの詳細を表 2 に示す.

表 2: 評価用データ

	データ A	データ B	データ C
データ長	47 分	19 分	60 分
音声区間の比率	78.5 %	31.3 %	40.0 %
平均継続時間長	4.1 秒	1.2 秒	2.4 秒
話者数	女性 1 名	男女 7 名	男女数名
主な発話形式	朗読	セリフ	会話, CM

試聴には得られた圧縮データのなかから聞き覚えによって評価が変化しないよう内容の重複が無いように切り出された 20 秒の区間を使用した. 各実験条件毎に用いた区間は 1 種類である.

表 3 に音響分析条件を示す. ケプストラム距離計算時に一次の差分フィルタ $(1-\alpha z^{-1})(\alpha=-0.8)$ を用いて低域周波数に重み付けを行った. Segmental BCF の窓長は $F=63$ (約 1 秒), Frame BCF の窓長は $J=1, W=1$ とした.

表 3: 音響分析条件

標本化周波数	16 kHz
窓長	256 点 (16ms)
更新周期	256 点 (16ms)
窓関数	Hamming 窓
高域強調	$(1-0.97z^{-1})$
特徴量	LPC ケプストラム
LPC 分析	14 次
ケプストラム分析	16 次

3.2 実験結果

図 4 にデータ A からデータ C までのそれぞれに対する圧縮率毎の MOS 値を示す. 図中の "BCF-BCF" は手法 ① を, "BCF-PICOLA" は手法 ② を, "PICOLA" は手法 ③ をそれぞれ示す.

データ A では手法間に顕著な違いは見られない. これは表 2 に示した通りデータ A に含まれる非音声区間の割合が音声区間にくらべて少ないため, 一段目の BCF による非音声区間の削除を行ったとしてもまだ大きな割合の音声が圧縮対象として残ってしまうためである. その場合は全体を一律に圧縮する PICOLA アルゴリズムとさほど変わらない圧縮データを生成することとなり, 手法間の違いが明確にならない.

データ B 及びデータ C では手法 ① 及び手法 ② が良好な評価を示した. これは表 2 に示した通りデータ B 及び C に含まれる非音声区間の割合が音声区間にくらべて多いため, 一段目の BCF による非音声の削除を行うことで, 二段目の圧縮の対象となるデータ量を削減することが可能となる. 圧縮対象となるデータ量は小さければ小さいほど a が大きくてすむ. a は大きいほど話速の上昇が小さなものになり話速は元音声に近くなることを意味する. 音声の再生速度は『音声の聞き取り易さ』にとって重要な要素であり, 元の音声に近いほど聞き取り易くなる. 提案手法による圧縮データは元音声を一律に圧縮する手法 ③ よりも話速が元音声に近く比較的遅くなるため, 同圧縮率の場合に評価が高くなると考えられる.

二段目の BCF と PICOLA 手法の比較としては, データ及び圧縮率によって評価がわかれた. 一段目で対象となるデータ量が削減されたことで, データを一律に圧縮する PICOLA アルゴリズムは単独で動作させる場合よりも話速を低速化させることができ, また波形の編集を行うことで自然な圧縮を実現できたため高評価になったと考えられる.

表 4 に各データに対する圧縮率 $a=1/3$ の時の各手法の計算時間を示す. 提案手法は対象データの波形編集等を行わず, 区間の切りだしのみを行うため, 処理速度は

19 分のデータ B に対して Pentium 4 (3.2 GHz)，メモリ 2.0 GB のマシンを用いて 2.21 秒と極めて高速に動作する．一方，区間の接続部にノイズが生じる欠点があり，評価実験の感想でも「接続部のノイズが気になる」といった意見があった．

表 4: 各手法の処理時間 (sec.)

手法	データ A	データ B	データ C
①	6.04	2.21	7.64
②	49.18	11.76	47.62
③	53.11	21.63	58.31

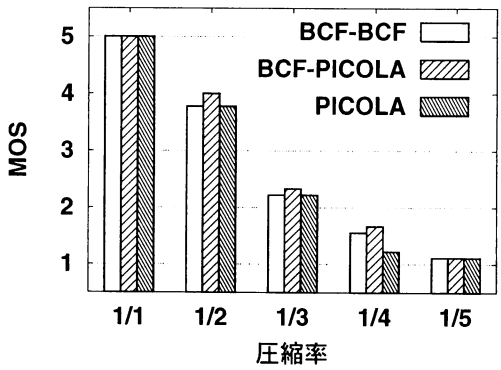
表 5 に手法 ① により生成されたデータ内の音声/非音声区間の合計区間長と，元データ内でのそれぞれの区間に対する比率を示す．各データの音声/非音声の項目に並記されているのは元データ中でのそれぞれの合計区間長である．例えばデータ A 中の音声区間は 2199.6 秒であり，1/2 に圧縮したデータ内での音声区間は 1285.1 秒となってこれは元の長さの 58.4% である．非音声区間よりも音声区間の割合が多く残っており，話速の高速化を緩めることに繋がっていることがわかる．

提案手法はヒストグラム h_n の作成に全てのフレームを必要とするので offline での動作となる．残存する非音声区間に対しては [4] で提案した音声/ 非音声モデルにより判別性能の向上が図れるので，さらに性能向上が期待できる．

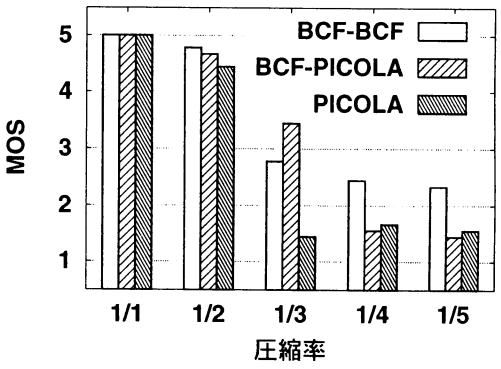
表 5: 圧縮データ内の音声/非音声区間

元データ (sec.)		圧縮率 x 1/2	1/3	1/4	1/5
A	音声	1285.1	887.5	681.6	547.9
	2199.6	(58.4%)	(40.3%)	(31.0%)	(24.9%)
	非音声	115.0	36.6	18.4	12.2
	602.4	(19.1%)	(6.1%)	(3.1%)	(2.0%)
B	音声	254.6	189.1	155.6	132.6
	353.2	(72.1%)	(53.5%)	(44.1%)	(37.5%)
	非音声	308.9	182.8	126.1	92.8
	774.4	(39.9%)	(23.6%)	(16.3%)	(12.0%)
C	音声	946.8	674.3	537.3	447.6
	1510.6	(62.7%)	(44.6%)	(35.6%)	(29.6%)
	非音声	891.3	538.9	381.7	287.6
	2166.3	(41.1%)	(24.9%)	(17.6%)	(13.3%)

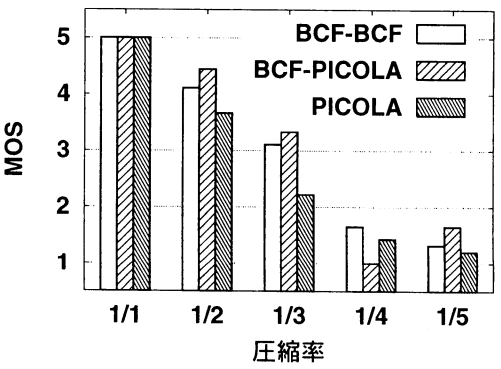
2.3 で述べた方法で決定された T_2 の値と適切と考えられる値を手動で決定した場合 [6] の主観評価結果との差異を，データ A からデータ C までそれぞれについて 図 5



(a) データ A



(b) データ B



(c) データ C

図 4: 3 つの手法の主観評価結果の比較

に示す.

データ A に関しては全体的に自動決定の方が良好な評価となり，データ B 及びデータ C に関しては評価にばらつきが見られる．データ A が最適値よりも良好な結果となった原因としては試聴区間が 1 区間のみであるので内容に依存した可能性も考えられるが，自動決定によっても評価が低下しない結果を示している．

4 むすび

本報告ではこれまで提案してきた BCF に基づく音声/非音声判別法を用いた時間圧縮音声再生法を提案した．元となるデータから非音声区間を取り除くことにより全体を一律に圧縮した場合よりも音声区間の圧縮率が低下するため，一律に圧縮した場合よりも話速の過度な高速化を防ぐことが可能となり，より聞き取り易い圧縮音声を生成することが可能となった．

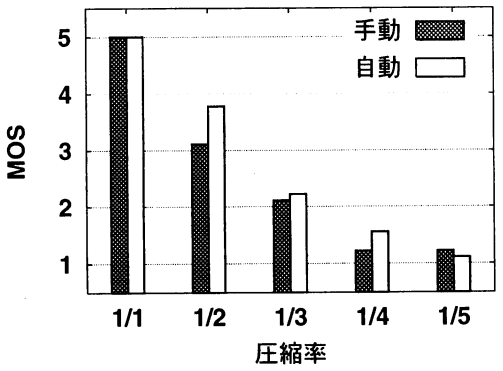
提案手法によって生成された圧縮音声を視聴した主観評価実験の結果は，話速の高速化を防ぐことで発話内容の聞き取り易さの点でデータを一律に圧縮する PICOLA 法を単独で用いた場合よりも良好な結果を得ることができ，またその手法を提案手法と併用することで双方の特徴を活用することで良好な結果が得られた．

今後の検討課題としてはリアルタイムでのデータ処理，offline での性能向上，接続部のノイズに対する処理，評価実験の大規模化がある．

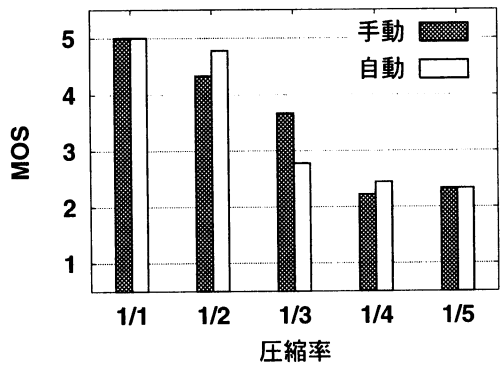
謝辞 日頃有益な討論を頂き，評価実験に協力してくださった本学ヒューマンインタフェース学講座の諸氏に感謝します．

参考文献

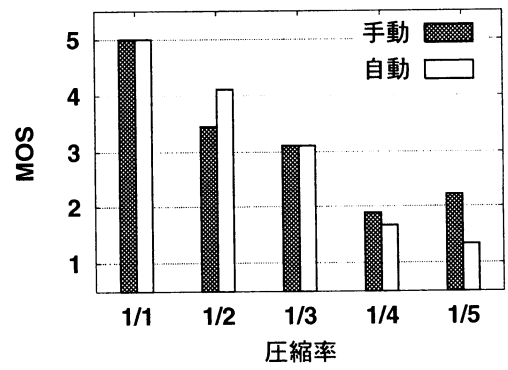
[1] 古井，“音響・音声工学”，近代科学社，1992.
[2] 森田，板倉，自己相関法による音声の時間軸での伸縮方法とその評価，信学技報，EA86-5，1986.
[3] 小西，他，映像速覧のための音声のノンリニア時間圧縮再生方法に関する検討，信学技報，SP2004-148，pp.19-24，2005.
[4] 竹内，杉山，音声/非音声区間検出のための自動モデル学習法の評価，信学技報，SP2005-29，pp.13-18，2005.
[5] 浅野，山下，杉山，Cepstrum Flux を用いた音声区間の検出，音学講論，3-Q-2，pp.121-122，1999.
[6] 竹内，杉山，音声/非音声判別法を用いた時間圧縮音声再生法，音楽講論，1-Q-24，pp.421-422，2005.
[7] 内田，杉山，CampusWave 音声データベースの作成，電気関係学会東北支部連合大会，2A-6，2000.



(a) データ A



(b) データ B



(c) データ C

図 5: 閾値の自動/手動決定による評価差