映像制作編集のため話速変換方式

6B−2

斉藤　一実　川口　尚久　飯島　泰裕

富士通研究所

1．はじめに

我々は数年前よりパソコンを使ったパーソナルな映像制作システムについて研究してきており，デジタル映像制作編集システムVideo Power Toolsを開発している．[1][2]

映像作品の制作では映像と音声を編集するが，映像編集を完了した後，映像の長さに合うよう音声を編集する．映像に合わせて音の長さを調節するとき，ＢＧＭであればフェードイン・アウトをかけたり，効果音であれば反復録音で長さ合わせを行う．しかしナレーションに関してはこのような手法が存在しない．このため，ナレーションは専門家であるナレーターを使い，録音の段階で長さを合わせるか，デジタル処理で長さを調整していた．

2．無音部圧伸方式

デジタル処理による話速変換方式としては，人声の母音部だけ削減する方式や，再生速度を変えた後，再サンプリングする方式など，音程が変らない方式が提案されている．しかし，前者は計算処理に非常に時間がかかり，また後者は聞き取り難くなるという問題があった．

そこで我々は，人が読む文章には，息継ぎや文章の句読点での間（無音部分）があることに注目して，図1のように無音区間のみを伸ばしたり縮めたりすることで全体の長さを変える方式を考案した．

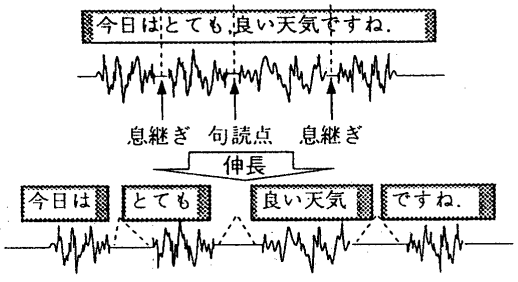

図1　無音部圧伸方式

この方式では音楽などは処理できず，人声に関してのみ利用できる．しかし，無音部分だけの変更なので計算処理が少なくてすみ，有音部の音声には変更を加えないので聞き取りやすいという特長を持つ．そして，パーソナルな映像制作において多用されるナレーションを，この方式で処理できる意義は大きい．

3．ナレーションにおける無音部の分布

本方式では最初に圧伸するための無音部を抽出しなければならない．そこで本方式の実現に当り，実際のナレーションでの無音部の分布を調査した．我々の開発したVideo PowerToolsシステムを使用してニュース番組のナレーションから無音部を取りだし，出現頻度をカウントした．内容は国際紛争のニュースでカナ２１７文字分，時間は３０秒弱である．

図2のグラフは解析結果で，横軸は無音部の長さ，縦軸は無音部の出現回数である．無音部は全部で約９００ヶ所あり，文字数の４倍以上ある．全体の８０％以上が５ｍｓ以下で非常に短い無音部が多数存在していることがわかる．

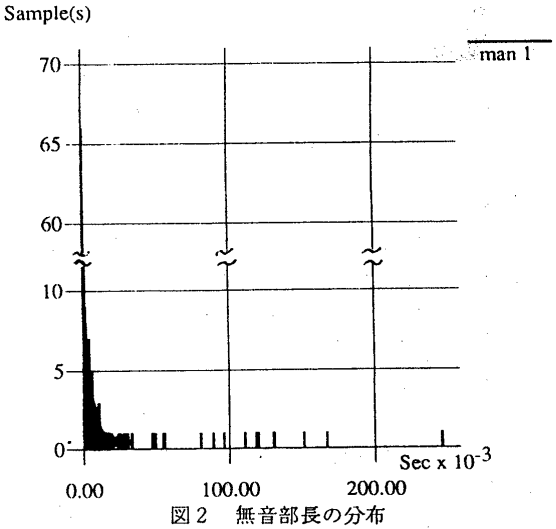

図2　無音部長の分布

これらは検出された無音部のほとんどが句読点や息継ぎなどの間ではないことを示している．本方式の実現ためには上のようにたくさんある無音部の中から，息継ぎ，句読点による無音部を正しく分別しなければならない．

そこで次に無音部の長さの分布について調査を行った．

[1]Speech speed conversion method for Post-Production System
Kazumi SAITO Naohisa KAWAGUCHI Yasuhiro IIJIMA
Fujitsu Laboratories LTD.

ナレーションの文章にはさまざまなものがあり，読み方には個人差が予想される．さらにアナウンサーなどナレーションの教育を受けた専門家と一般の素人でも読み方に違いがあると思われる．そこで次の３点について調査した．

1）文章による差異
2）個人差
3）専門家と素人による差異

4．文章による差異と個人差

まず，文章による差異と個人差がどのように現れるかを調べた．ＴＶニュース番組には文章の種類や話者が多数存在するので，これを利用した．政治，経済，文化など内容の異なるニュースから男性７件，女性３件の音声を収録して解析した．各音声は約３０秒程度の長さである．図３にその結果を示す（紙面の都合上３件のみ示す．データ数４件以上は割愛）．

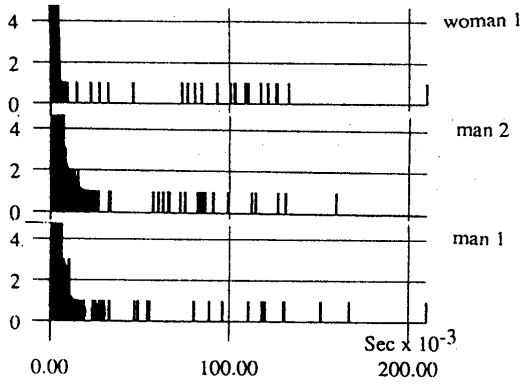

図３　無音部の長さと出現頻度（専門家の場合）

文章や話者によらず，無音部の９９％以上は２０ｍｓ以下である．また５０ｍｓ辺りからデータが疎らになる．最長の無音部の長さはサンプル毎にかなりのバラつきが見られた．実際の音声と解析結果を照らし合わせると，図３下man1の音声は句読点等で空けられた間が１０ヶ所あった．これを解析結果で見るとの長い方から１０個分の無音部は５０ｍｓよりも長く，６０～２５０ｍｓの範囲であった．

以上のことから文章や話者にはあまり関係無く，息継ぎ，句読点による間の長さは５０ｍｓ以上あると考えられる．

5．専門家と素人の差異

次に専門家と素人で差異が現れるか調査した．図４は図３下man1と同じ文章を，５人の素人が読んだ結果である．

５人中４人の被験者のデータは５０ｍｓ以降は無音部の個数が急激に減る．５０ｍｓを越える無音部は，１人が２６個と特に多い．他の４人は７～１５ヶ所と少なく，音声を聞いたときに，間を空けて読んでいる部分の個数に近い．２００ｍｓを越える無音部は専門家に比べると多い．素人は間を空ける時に専門家より長めにする傾向が見られた．

以上のことから，素人の場合でも専門家と同様に，息継ぎ，句読点による間の長さは５０ｍｓを越えている．

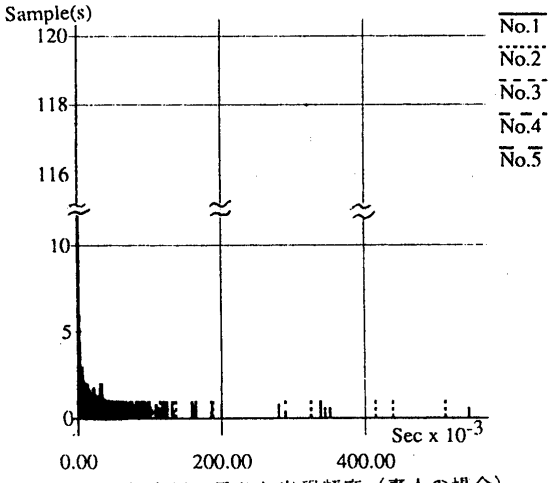

図４　無音部の長さと出現頻度（素人の場合）

6．まとめ

以上の結果に基づき，無音部として検出する長さの閾値として５０ｍｓを１つの候補とした．そして，ニュース番組，ビデオのナレーション等について，５０ｍｓ以上の無音部のみをのび縮みさせて話速変換を行い試聴した．結果は，文章が不用意に途切れたりせず，元のナレーションと同程度に聞き取りやすく，良好な話速変換結果を短時間で得られた．

問題点としては，無音部のみの圧伸のため，縮小率に限界が有ること，また伸長時の無音部の伸長分を完全無音にすると違和感がでやすいことがあげられる．

今後はこれらの問題点を解決し，より違和感のない話速変換を目指す．

［謝辞］

研究の機会を与えてくださったマルチメディア研究部藤田部長，山本部長代理，また開発にご協力頂いた東京システム技研小島氏に感謝いたします．

［参考文献］

[1] 飯島，川口，斉藤　”パーソナルなビデオ制作に向けて”，信学画像工学研究会 IE91-7，1991

[2] 斉藤，川口，飯島　”デジタル映像制作編集システム並列画像処理機構”，情報処理学会43回全国大会，1991