

Internal Use Only (非公開)

TR-SLT-0042

動画翻訳におけるノンバーバル情報の制御
Control of Nonverbal Behavior
for Multi-modal Translation System

足立 吉広 森島 繁生 中村 哲
Yoshihiro ADACHI Shigeo MORISHIMA Satoshi NAKAMURA

2003 年 6 月 24 日

概要

近年映画等において、音声翻訳による「吹き替え」がよく行われている。これらは一般的に人によって翻訳され、声優によって発音された音声を元の動画の音声と入れ替える手法がとられている。この手法は映画中の俳優の口の動きと翻訳音声との間にずれが生じている。このような問題に対して、過去に緒方らによって動画翻訳システムが実現されている。この翻訳システムは音声を自動で翻訳し、かつ映画中の俳優に翻訳音声に同期した口形を合成することにより、あたかも映画中の俳優が翻訳音声を話しているように見せるシステムである。しかしこのシステムは発話と密接に関係して起こる「うなずき」や「目の周りの動き」等のノンバーバル情報について考慮されていなかった。その為、翻訳動画では音声とノンバーバル情報の関係は崩れてしまっていた。

そこで翻訳音声とノンバーバル情報との関係を、原動画の音声とノンバーバル情報との時間軸上の相対的な関係を保存する手法を用いて、翻訳動画におけるノンバーバル情報の時間的位置の補正を行った。

(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

©2003 (株) 国際電気通信基礎技術研究所
©2003 Advanced Telecommunication Research Institute International

第1章 研究概要

近年映画等において、音声翻訳による「吹き替え」がよく行われている。これらは一般的に人によって翻訳され、声優によって発音された音声を元の動画の音声と入れ替える手法がとられている。この手法は映画中の俳優の口の動きと翻訳音声との間にずれが生じている。このような問題に対して、過去に緒方らによって動画翻訳システム^[1]が実現されている。この翻訳システムは音声を自動で翻訳し、かつ映画中の俳優に翻訳音声に同期した口形を合成することにより、あたかも映画中の俳優が翻訳音声を話しているように見せるシステムである。しかしこのシステムは発話と密接に関係して起こる「うなずき」や「目の周りの動き」等のノンバーバル情報について考慮されていなかった。その為、翻訳動画では音声とノンバーバル情報の関係は崩れてしまっていた。

そこで翻訳音声とノンバーバル情報との関係を、原動画の音声とノンバーバル情報との時間軸上の相対的な関係を保存する手法を用いて、翻訳動画におけるノンバーバル情報の時間的位置の補正を行った。

第2章 発話とうなずき

発話とうなずきの関係の補正システムの全体像を図1に示す。本システムは原動画分析部と翻訳動画補正部により構成されている。まず原動画に対して発話区間の検出、うなずき区間の検出をおこない、それらの時間軸上の相対的位置関係を抽出する。次に音声翻訳を行い、音声翻訳された動画に対しても発話区間の抽出を行う。この発話区間の時間的位置情報を元に、翻訳動画におけるうなずき区間のあるべき位置を推定し、動画伸縮によりうなずき区間を時間軸上でシフトさせ、翻訳動画における音声とうなずきの時間的位置を補正して出力する。

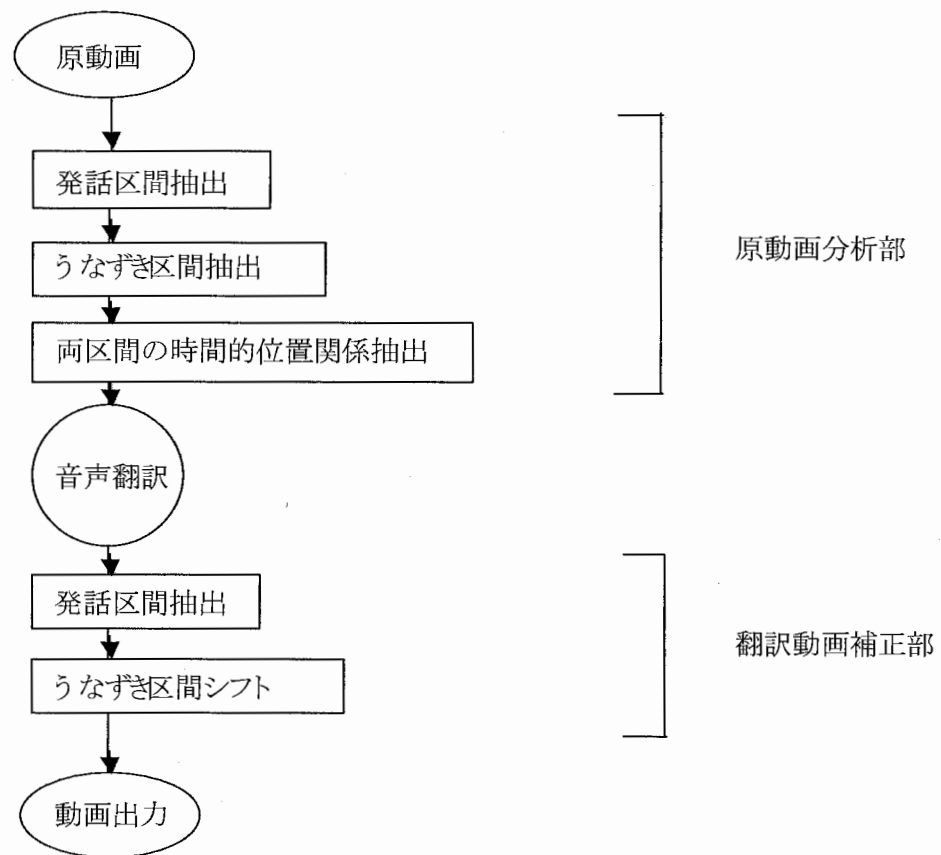


図1 発話とうなずき関係の補正システム

2.1 発話区間抽出

動画に対する発話区間の抽出は音声のパワー情報を用いて行った。パワーというのは音量の強弱のことであり、音声波形で振幅の大きさとして現れる。本研究で用いた動画は図2のように被験者にカメラの前に立ってもらい、人に道を尋ねる場面を想定して、「ちょっとお聞きしたいことがあるのですが、えー博物館にはどのようにしていけばいいですか」と発話してもらったものを用いた。この動画の発話の音声波形を図3の赤線で示す。一般的に発話している区間ではパワー(図3緑線)が大きいので、このことを利用して、閾値(図3黄線)以上のパワーを持つ区間を発話区間として抽出した。

しかし、本システムでは音声翻訳を文単位で行うため、この発話区間抽出も文単位で行う必要がある。その為、図3中の丸印のように発話区間でないと判定されてしまった区間についても、その継続長が短い場合に対しては発話区間とする平滑化を行い、発話抽出を行った(図3)。



図2 本研究で用い動画

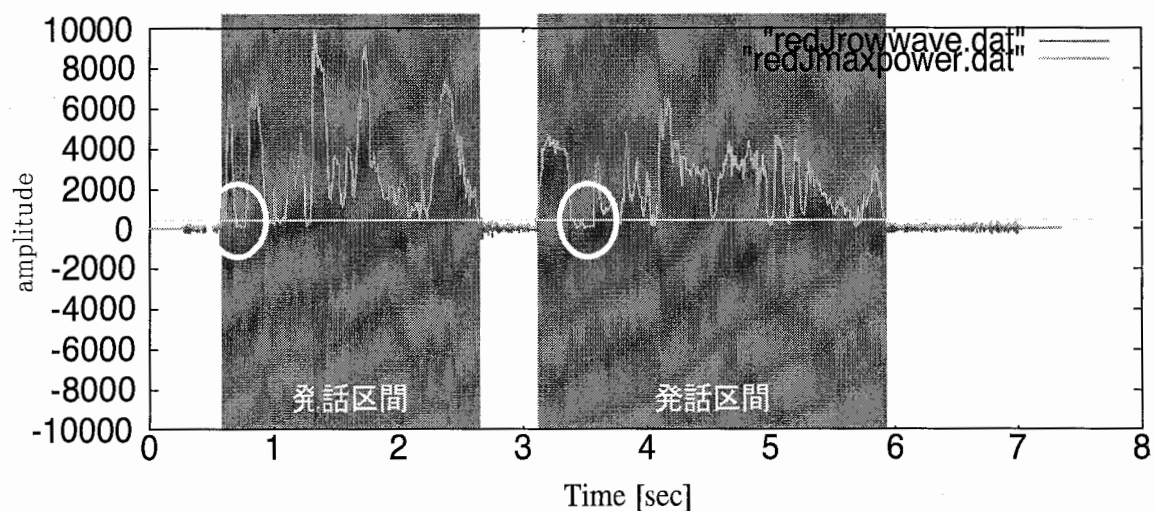


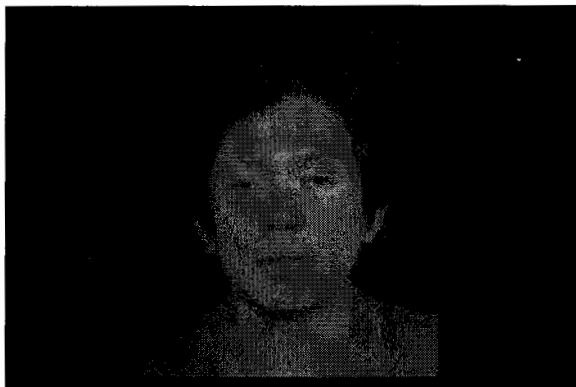
図3 発話区間抽出

2.2 うなずき区間抽出

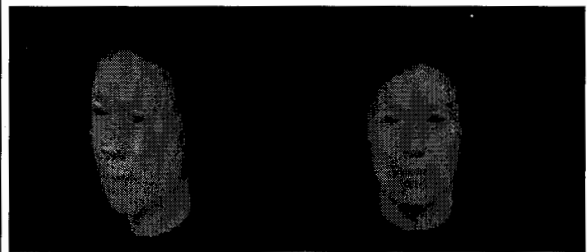
本研究では動画中の人物の頭の動きをテンプレートマッチングにより推定し、その縦振り回転の遷移を抽出し、回転速度とその継続長が閾値以上の区間をうなずき区間と定義し、その抽出を行った。

2.2.1. 首の縦振り回転の抽出

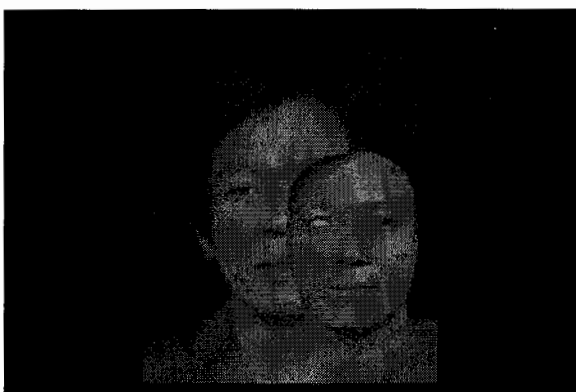
首の縦振り回転を抽出するために、2次元動画中の人物の頭の動きから3次元空間での頭の動きを推定する方法として、本研究ではテンプレートマッチングを行った。この手法は2次元動画を用意し(図4 a)、あらかじめ動画上の人物の3次元モデルを作成する(図4 b)。そのモデルを2次元に投影し動画の各フレームの画像とピクセルごとにRGB値の差をエラー値として計算する(図4 c)。この値が小さくなるようにモデルを移動、回転させ、最小となったときのモデルの移動量、回転量をテンプレートマッチング結果として出力する(図4 d)。この手法により首の縦振り回転の遷移を抽出することができる。



a) 2次元動画



b) 3次元モデル



c) 3次元モデルを投影してエラー計算



d) エラー最小

図4 2次元動画中の人物の頭の動きの推定のためのテンプレートマッチング

2.2.2. うなずき区間の選出

本研究ではうなずいている間は首の縦振り回転の速度が早く、かつうなずくにはある程度の時間がかかるということに着目した。テンプレート マッチングにより首の縦振り回転の角度の遷移(図5青線)が得られているのでその値から差分をとることにより、首の縦振り回転の速度が抽出される(図5桃色線)。うなずきにはある程度時間がかかるというのは頭のぶれのような動きに対して、うなずきと判定されないようにするための考え方である。ぶれは単発的に起こり、うなずきは時間を掛けて起こるため、時間幅をもったフレーム内の速度の合計を求めることで区別する。ぶれの場合は単発的に速い速度が出て、フレーム内の合計値への影響は少ないが、うなずくと速い速度が継続的に出るためフレームの合計値は大きくなる。これを実装するためにフレーム長 0.5[sec]のフレームを 33[msec]ごとにシフトさせ、それぞれのフレーム内の回転の差分の合計をそのフレームの首の縦振り回転の速度(図5黄線)とし、この値に対して閾値(図5 赤線)を用いてうなずき区間を検出した(図5)。

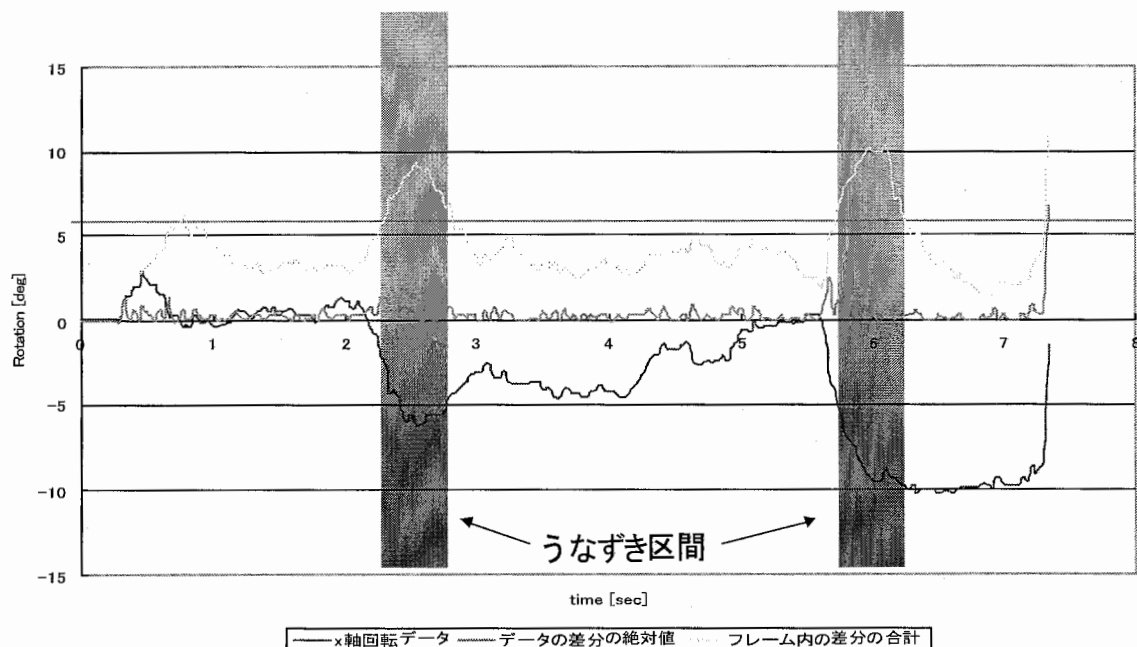


図5 うなずき区間抽出

2.3. 音声翻訳

本システムの音声翻訳部は緒方らによって確立された動画翻訳システムの音声翻訳部をそのまま用いる。そのシステム像を図6に示す。この音声翻訳システムはATR音声翻訳通信研究所で開発されたATR-MATRIX^[2h]と呼ばれているものである。ATR-MATRIXはSPRECと呼ばれる音声認識部とTDMTと呼ばれるテキストに変換された言語情報を翻訳する音声翻訳部、CHATRと呼ばれるテキストから合成音声を生成する音声合成部からなっている。この自動音声翻訳システムを用いることにより、動画中の音声を自動で翻訳することが出来る。

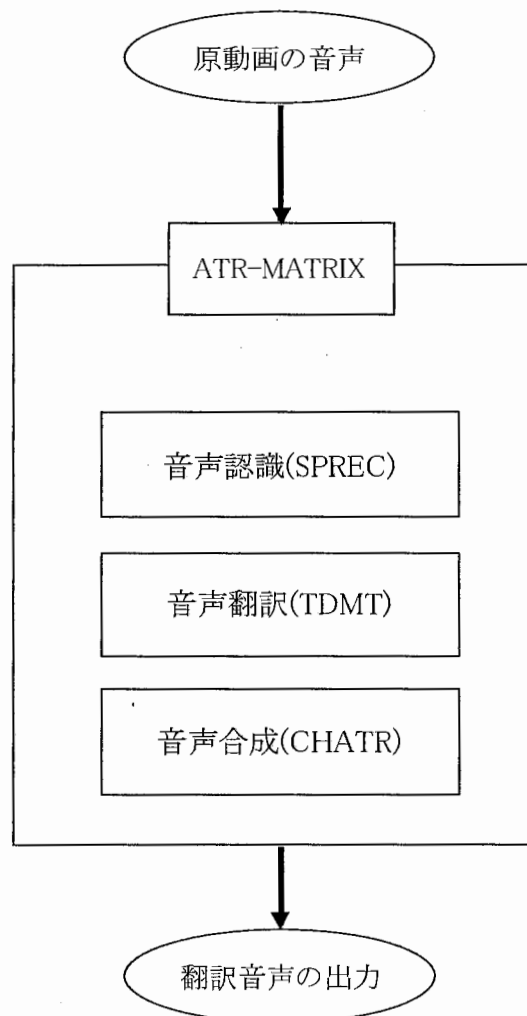


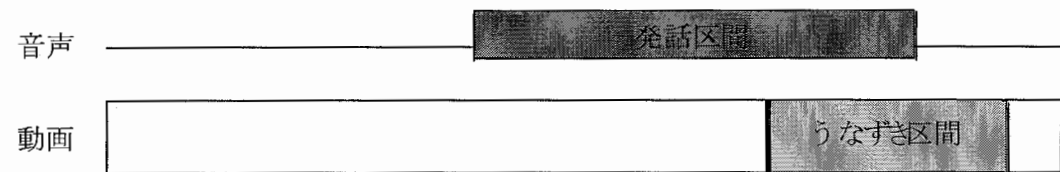
図6 音声翻訳部のシステム像

2.4. 翻訳動画に対する発話区間抽出

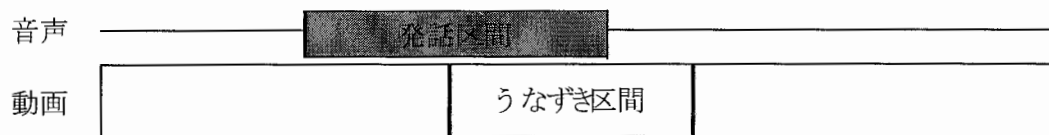
本システムでは音声とうなずきの時間軸上の相対的關係を補正する際に、音声を基準にうなずきの時間軸上の位置補正を行う。その為に翻訳動画に対しても発話区間抽出を行う。

2.5. うなずき区間のシフト

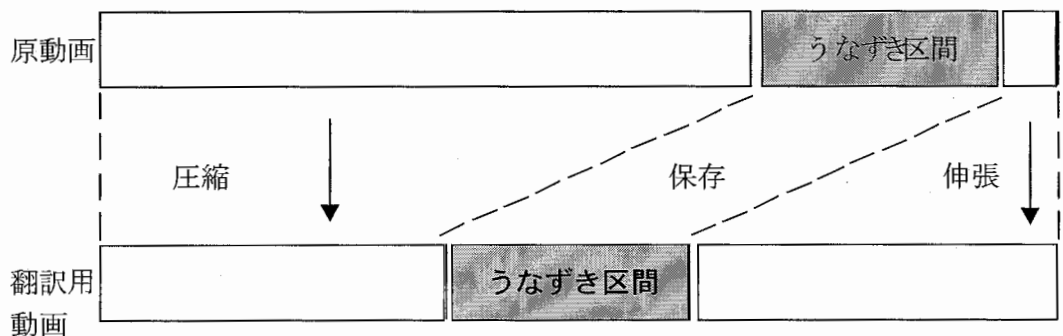
うなずき区間シフト法を図7に示す。a)は音声翻訳前に抽出された発話とうなずきの関係の一例である。b)は翻訳動画の発話区間とうなずき区間のあるべき位置である。この「あるべき位置」は原動画の両者の相対的關係から発話区間を基準として推定される。この位置にうなずき区間をシフトさせるために原動画をc)のようにうなずき区間とそうでない区間とに区切る。そしてうなずき区間の両隣の画像を時間軸上で伸縮させることにより、うなずき区間の動画をシフトさせる。うなずき区間を伸縮させずに保存するのは、この区間では動画中の人物の頭の動きが激しいため、時間伸縮するとその影響が目立ってしまうためである。



a) 動動画の発話区間とうなずき区間の関係



b) 翻訳動画における発話区間とうなずき区間のあるべき位置



c) 動画を区切り画像伸縮

時間

図7 画像伸縮によるうなずき区間のシフト

第3章 発話と目周り

発話に伴うノンバーバル情報はうなずきだけではなく、目の周りも発話と密接に関係した動きが発生していると考えられる。例えば発話と同時にまゆを上げる、目を見開く等である。これらもうなずきと同様に音声翻訳することによりそれらの関係が崩れてしまっている。そのため翻訳動画の音声と目周りの動きの関係を補正する必要がある。

目周りの補正システム像を図6に示す。まずは原動画の各フレーム画像(図8a)とテンプレートマッチングにより得られた位置に移動した目の周りの3次元テンプレート(図8b)を用いて、重ね合わせて3次元テンプレートにテクスチャを貼り付け(図8c)、目周りの3次元モデルを各フレームごとに作成する(図8d)。次に音声翻訳を行う。ここで発話の長さの違いから発話と目周りの動きの関係が崩れると考えられる。その後、目周りの3次元モデルのアニメーションを時間伸縮して、翻訳動画中の人物の頭の角度にあるように、3次元モデルを時間伸縮前のフレームに割り当てられていた位置に移動、1回転させ、目周りに重ね合わせることで、翻訳音声と目周りの動きの関係が補正された動画を得ることができる。

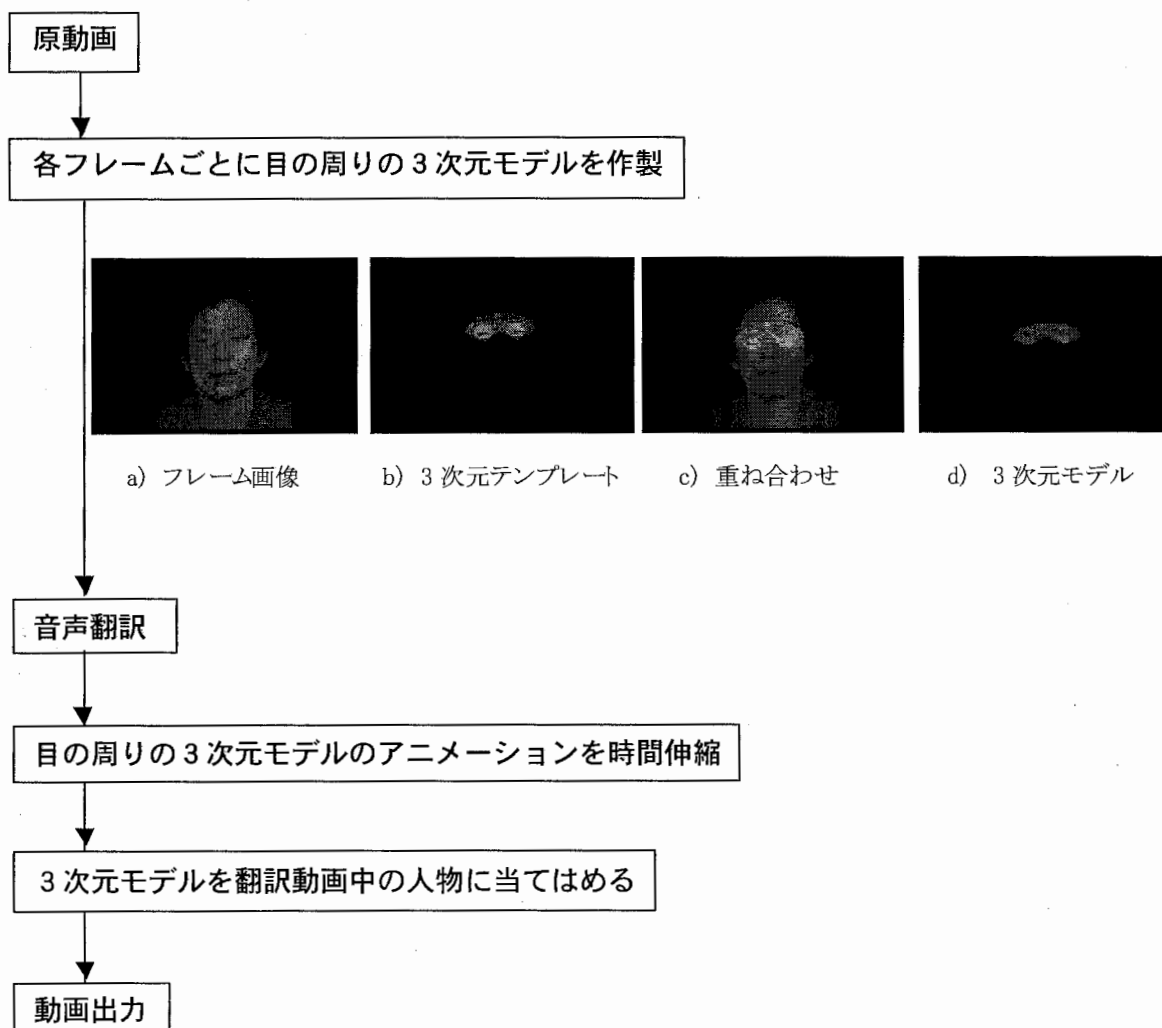


図8 目の周り補正システム像

第4章 まとめ

翻訳動画において、崩れてしまった音声とうなずき等のノンバーバル情報との時間軸上の関係を、元の動画における相対的關係を維持するという手法を用いて、その補正を行った。

今後の課題としては、本手法により音声とノンバーバル情報を補正された動画が元の翻訳動画に比べてどういう印象を受けるのか、また本研究で用いた、相対的位置を保存するという補正法が正しいのかどうかを調査する必要がある。

うなずき等のノンバーバル情報というのは会話中においてコミュニケーションの補助的な役割を持つが、これは必ず起こるわけではなく、またその定義が難しい^[3]。したがって人間の会話に密着したノンバーバル情報についても調査する必要がある。

参考文献

[1] 緒方信、中村哲、森島繁生, “ビデオ翻訳システム - 自動翻訳合成音声とのモデルベースリップシンクの実現 -”, 情報処理学会シンポジウム、インタラクション2001論文集, Vol.2001, No.5, pp.203-210, March 2001.

[2] 菅谷、竹沢、横尾、山本「日英双方向音声翻訳システム(ATR-MATRIX)の対話実験 日本音響学会 1999 年春季研究発表会講演論文集, pp 107-108, 1999

[3] 渡辺富雄、夏井武雄「ヒューマン・インタフェースへの音声対話時の引き込み現象の応用に関する研究:うなずき反応を視覚的に模擬する音声反応システムの開発」昭和63年度厚生省心身障害研究