ＴＲ－ＳＬＴ－００２８

複数人話者の会話シーンの画像翻訳

Image Translation Of Multi Speaker' s Conversation Scene

前島 謙宣†　　森島 繁生　　中村 哲
Akinobu Maejima　　Shigeo Morishima　　Satoshi Nakamura

2002年11月11日

概要

本論文では、複数人の話者の会話シーンにおける画像翻訳の手法について述べる。会話シーンにおいて、ビデオ映像中の人物の顔の動きを推定し、映像中に存在する各話者について発話判定を行う。発話が検出された話者の口領域を、別に用意された音声に同期して合成された口唇映像で置き換えることにより、他言語もしくは変換された発話内容へのリップシンクを 実現する。

（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
†成蹊大学大学院工学研究科
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

複数人話者の会話シーンの画像翻訳

前島 謙宣†‡　森島 繁生†‡　中村 哲‡
† 成蹊大学大学院工学研究科
‡（株）ATR 音声言語コミュニケーション研究所

1．はじめに

近年、海外の映画やTVの日本語への吹き替えが行われているが、それらを見ると口の動きと発話内容の不一致に気が付く。このような問題に対して、従来からビデオ翻訳の研究が進められている[1]。ビデオ翻訳とは、音声の翻訳のみではなく、話者の調音器官の画像も翻訳しようとするである。

従来のビデオ翻訳システムでは、話者一人が会話しているシーンに対してのみ適用可能であった。しかしながら、実際の映画やTVへの応用を考えると、複数人の話者が会話しているシーンに対しても適用できることが望ましい。

そこで、本研究では、従来のビデオ翻訳システムを、複数人の話者が会話しているシーンに対しても適用できるように拡張することを狙い、その初期段階として、複数人の話者の会話シーンについて画像の翻訳を行う手法を提案する。

2．3次元顔モデルの生成

本章では画像翻訳のための最初の工程として、自動顔トラッキングと個人用の口領域モデルに利用する為の3次元頭部モデルの生成法について説明する。

2-1．3次元頭部モデル

人間の顔は基本的な形状や構造は同じといってよいが、目、鼻、口等の各要素を構成する形状や位置は、個人によって異なる。このためCG(Computer Graphic)によって自然な表情を合成するには、対象人物の顔に、より忠実かつ演算量の少ない3次元頭部モデルを構成する必要がある。

そこで本研究では、当研究室で開発されている3次元頭部モデル[図1]を用いて、口領域の3次元モデルの生成を試みた。この3次元頭部モデルは約1500ポリゴンの三角形パッチより構成されており、格子点数は約800から構成されている。また、本研究では、顔領域全体に渡って整合を行っている。整合については次節で説明する。

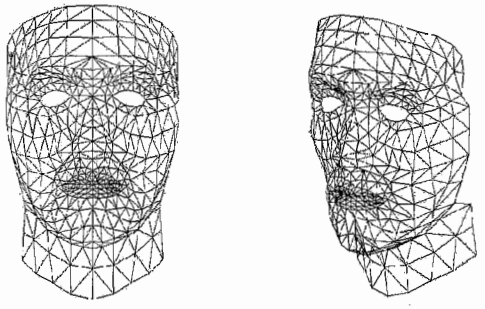

図1．標準3次元顔モデル

2-2．モデルの整合

本研究では、標準顔モデルと個人用モデルの3次元形状の整合手法として、物体の正確な3次元形状が測定可能なレンジファインダを使用し、対象人物の頭部の3次元形状とテクスチャを取得する方法を用いた。レンジファインダを用いることにより、取得したテクスチャの色情報を手がかりに、テクスチャ座標に対応するモデルの3次元形状を取得することで、標準モデルに対象人物の正確な形状情報を付加することが可能となる。

3次元形状を取得した後、ビデオフレーム画像を、作成した個人用のモデルに反映させるために、専用のGUI[2]を用いてさらに2次元への整合を行う。図2に生成された3次元口領域モデルと顔追跡に用いるための、生成されたテンプレートを示す。また口内は、あらかじめ用意したモデルを個人用のモデルに適合して用いた。

3．撮影条件

本研究ではビデオ映像として、以下のような制約の基で撮影された映像を使用した。

1．カメラと被写体が十分に離れているものとして、映像空間は正射影として近似
2．頭の動きは、普段の会話時に発生する程度のもの
3．口は探索に使用しない為、運動しても可
4．極端に影が表出しない光源の設定
5．人物の表情は、極度に皺のできるものは想定しない。
6．顔の大きさが極端に変化しないもの
7．音声は1ch
8．一人が発話している間は、他の話者は発話していないものとする。すなわち、音声の重複はないものとする。

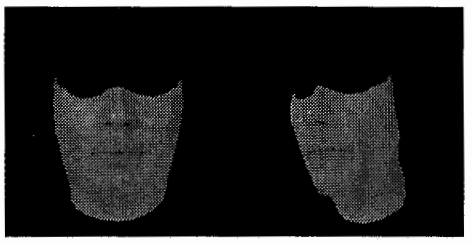

図2.a 生成された口領域モデル

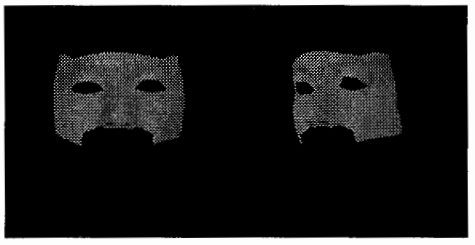

図2.b 生成されたテンプレートモデル

4．3次元テンプレートを用いた自動顔トラッキング[1][2]

本章では、ビデオを映像中の対象人物の顔追跡手法である、前章で導入した2次元の画像テンプレートを用いた手法について述べる。

4-1．画像テンプレートの作成

画像テンプレートは、前章で述べた3次元顔モデルにビデオ映像中の初期の1フレームに対してテクスチャを貼り、そのモデルのx,y方向への平行移動と、x,y,z軸を中心とした回転運動に対して、それぞれを少しづつ変化させ、2次元に投影することで作成する。原動画映像中の対象人物の口周辺は、発話中絶えず運動していることを考慮して、マッチング結果への影響を減らすために、テンプレートには口周辺領域を削除したものを用いた。

4-2．テンプレートマッチング

モデルを投影したテンプレートモデルを作成する共に、ビデオフレームとのコストを削除するために、テンプレートモデルをブルーバック上に投影することで作成され、青色領域を除いた顔部分のみでマッチングを行った。この為、テンプレートごとにマッチングするピクセルが異なる式（1），(2)のように誤差を正規化してから比較を行った。自動顔追跡アルゴリズムの流れ図を図4に示す。

$$Error = |R_V - R_T| + |G_V - G_T| + |B_V - B_T| \quad \cdots(1)$$

$$Normalized\_Error = \frac{Error}{Number\_of\_Pixels} \quad \cdots(2)$$

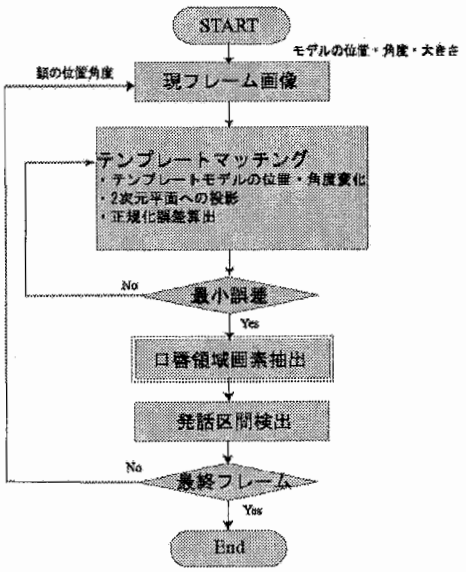

図4．顔自動トラッキングシステムの流れ図

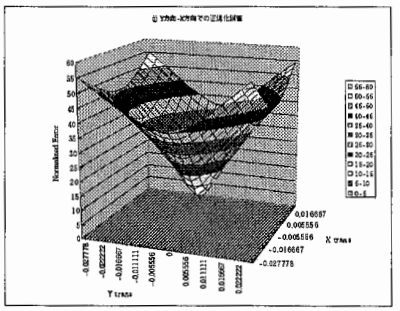

図5．x、y軸方向の変化に対する正規化誤差

4-3，最小誤差探索法

ー全周囲探索法

最も単純なテンプレートマッチングによる最小の誤差探索法には、必要な探索範囲全て（本研究ではX,Y方向への平行移動とX,Y,Z軸中心の回転運動5次元で表される最大移動量及び回転各）の正規化誤差を求めた結果、最小誤差を1つ決めるものがある。探索範囲は誤差のローカルミニマムは1つであると仮定し、前フレームでの位置・角度を起点として、起点の3^n-1近傍（nは次元数；5次元である時242近傍）の正規化誤差を調べ、誤差が最小となる位置・角度に起点を順次移していき、その起点が誤差最小となったときの値をそのフレームでの位置・角度とした。図5に各々の方向への移動及び各々の軸で回転させたときの正規化誤差のグラフを示す。全周囲探索法と急降下探索アルゴリズムを併用した手法[1][2]等があるが、本研究では、トラッキングの誤差が発話判定に影響を及ぼすため、今回はこの全周囲探索法のみを用いた。

5．口唇領域画素の抽出

本研究では、複数人の話者が会話しているシーンに対して、口領域モデルを重ね合わせて、合成動画像を作成するため、各音声区間において誰が発話しているのかを特定し、その結果を各々の口領域にモデルに対して反映する必要がある。そこで、4-2節で説明したテンプレートマッチングにおいて、正規化誤差が最小になった位置・角度における、モデルの唇上端点から下端点への画素列を抽出する。その後抽出した画素列をL*a*b表色系へと変換する。そして、計算対象フレームと前フレームとのL*a*b色座標系の距離をとることで、フレーム毎の画素の変化量（以下ではフレーム間エネルギーとする）を算出する。ここで、Δabは L*a*b 色座標上の Euaclid 距離を表す。また、話者特定する際に、話者によって唇の大きさが異なることから、抽出した画素の総量で正規化する。自動顔トラッキングでの誤差も考慮し、x・y方向に5画素分のオフセットを設け、それぞれの変化量が最小となる値をそのフレームにおける、フレーム間エネルギーとした。図6に算出されたフレーム間エネルギーを示す。

$$e(f, offset) = \frac{1}{Number\_of\_pixels} \sum_{j \in number\_of\_pixels} \Delta ab\{Lab(f,j) - Lab(f-1, j \pm offset)\} \quad \cdots(3)$$

$$E(f) = \min(e(f, offset)) \quad \cdots(4)$$

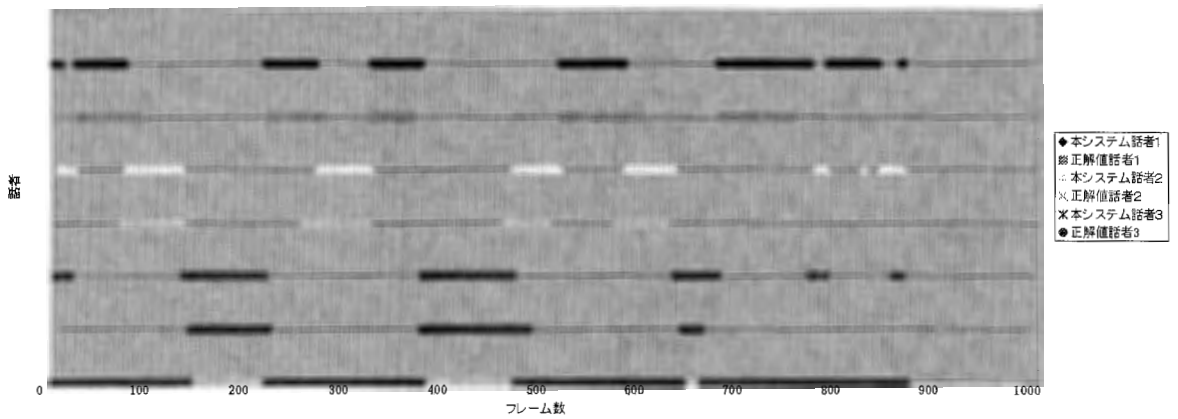

図9. 本システムとハンドラベルによるによる発話区間、話者判定の比較

6. 発話区間検出および話者判定

前章で抽出されたフレーム間エネルギーは、ある程度発話区間を検出しているが、このままの状態で、話者間のフレーム間エネルギーのみを比較し発話判定を行うと、合成動画像を作成した際にフレーム毎に話者の入れ替わりがおこり、結果として正確な発話を表現することができない。そこで、5章で抽出したレーム間エネルギーに対して、式5を用いて平滑化を行った。これは、対象フレームの前後のフレーム分の平均値をとるものである。本研究ではこの値を15フレームとした。この平滑化を行うことにより、話者間のフレーム間エネルギーを比較し、フレーム間エネルギーが最も高い話者を、対象フレームにおける話者であるとした。図7に平滑化後のフレーム間エネルギーを、図8に各話者間のフレーム間エネルギーの比較の様子を示す。

$$E'(f)=\frac{1}{2N+1}\sum_{i=-N}^{N}E(f+i) \qquad \cdots(5)$$

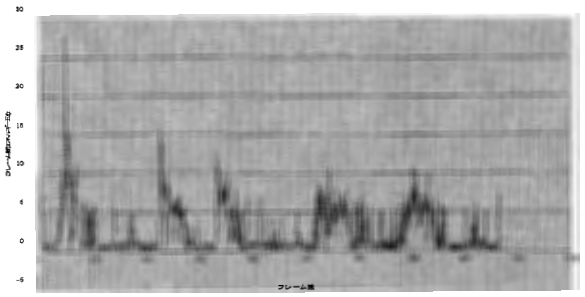

図6. フレーム間エネルギー

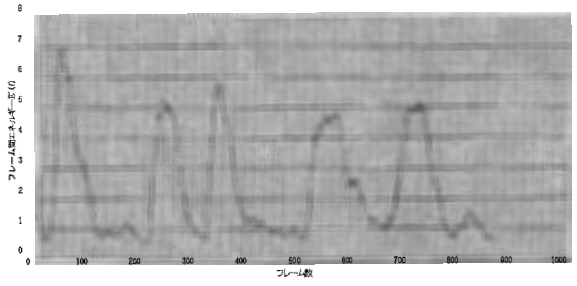

図7. 平滑化されたフレーム間エネルギー

7. 比較

ここでは本システムとハンドラベリングにより行った各話者の発話区間と話者判定を比較する[図9]。図9では、上から順に本システムによる話者の判定結果、ハンドラベリングによる話者判定の結果(正解値)を示している。本システムでは、フレーム間エネルギーを対象フレームの前後15フレーム渡り平滑化を行っているが、平滑化の影響により本来の発話区間よりも長く発話区間が検出されてしまっていることが図から読み取れる。本来の発話区間よりも長く発話区間が検出されると、本システムでは他の話者の発話区間に影響が及び、他の話者の発話区間が実際の発話よりも短くなってしまうことがある。本研究では4章で説明したように、一人が発話している際に他の話者は発話しないという制約をつけて撮影したビデオ画像を使用したが、調音器官の動作は、実際の音声が発せられている時間よりも長いため、画像上で話者の重複が生じているため、このような現象が起こると考えられる。本研究では、画像のフレーム間エネルギーのみを用いて判定を行っているので、今後音声のフレームパワーを用いて、発話区間を検出し、その結果と照らし合わせて判定を行うことにより、このような問題は解決できると考える。

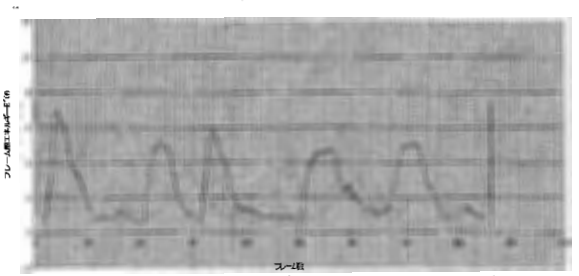

図8. 話者間のフレーム間パワーの比較

8. 発話口形の作成

人間が会話をする際、動作の大きい部分として、唇、顎、などが挙げられる。特に唇の動きは音韻と密接な関係がある為、正確な制御が要求される。しかし、本研究では、話者への依存性を削減する目的から次章の手法を用いた。

8-1. 標準口形状データの設定

8.1.1 口形パラメータによる口形の作成

口領域の動きを定量的に表現するために、文献[2]では口領域の制御点として、11点を定めている[図10]。各々の制御点はワイヤーフレームモデルの格子点と対応しており、骨格と筋肉の運動に基づく3次元の移動法則が定められている。本研究では、表現する音韻を表している正面と側面の2方向から撮影された、口形状の参照画像を用意し、上記の制御点を移動することで、ワイヤーフレームモデルをその参照画像に近づけられるよう変形させた時に得られる、格子点のベクトル移動量を基本口形のデータベースとしている。このデータは、口領域の大きさで正規化したベクトル移動量としているため、一度基本口形を用意すれば、すべての話者に対して適用することが可能である。このように本研究では話者に依存しない小規模なデータベースしか必要としないシステムを実現している。

8.1.2 レンジファインダを用いた口形の作成

口形パラメータには、唇、顎に関しては運動法則が定義されているが、実際の発話時に表出する頬や鼻に代表される口周辺の運動は再現することができない。しかし、3章で述べたレンジファインダを用いて発話時の口形と無表情時の3次元形状データを計測して、前項と同じく対応する座標の差分よりベクトル移動量を算出することで、口形パラメータで表現することのできなかった口領域の運動をモデルに反映することが可能となる。本研究では、人物に特化したデータにならないように6人の被験者のデータの平均を母音「あ」「い」「う」「え」「お」についてそれぞれ作成した。

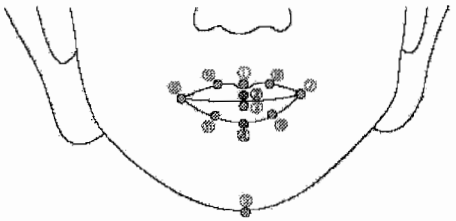

図 10. 制御点の位置

表1. 口形パラメータ

制御点	口形パラメータ	動作対象
1	1	上唇上側の縦方向の動きに対応するパラメータ
	2	上唇上側の奥行き方向の動きに対応するパラメータ
2	3	上唇下側の縦方向の動きに対応するパラメータ
	4	上唇下側の奥行き方向の動きに対応するパラメータ
3	5	下唇上側の縦方向の動きに対応するパラメータ
	6	下唇上側の奥行き方向の動きに対応するパラメータ
4	7	下唇下側の縦方向の動きに対応するパラメータ
	8	下唇下側の奥行き方向の動きに対応するパラメータ
5	9	顎の縦方向の動きに対応するパラメータ
6	10	唇端点の縦の動きに対応するパラメータ
	11	唇端点の横の動きに対応するパラメータ
	12	唇端点の奥行き方向の動きに対応するパラメータ
7	13	唇の横方向の開き具合に対応するパラメータ
8	14	上唇上側の制御点付近以外の動きに対応するパラメータ
9	15	上唇下側の制御点付近以外の動きに対応するパラメータ
10	16	下唇上側の制御点付近以外の動きに対応するパラメータ
11	17	下唇下側の制御点付近以外の動きに対応するパラメータ

8.1.3 舌モデルの導入

口内モデルに新たに舌モデルを導入した。本研究で導入した舌モデルは、実測を基にモデリングしたものではなく、手作業により作りこんだ作成したものを用いた。舌の運動には、舌自身の筋肉の運動によるものと外部の筋肉によって影響を受けるもの等多数あり、骨格が存在しないの分、その運動のすべてを定量化することは困難であると考え、6つのパラメータを定義した。表2に示す。

表2. 舌パラメータ

舌パラメータ	パラメータが制御する動き
1	舌の高さを調整する
2	舌を前後に移動する
3	舌の先端付近を曲げる
4	舌の中央付近を曲げる
5	舌の厚みを調整する
6	舌の幅を調整する

表3. 音素とVISEME対応表

VISEME No.	音素表記
1	/ae/
2	/ah/, /ax/
3	/A/
4	/aa/
5	/er/, /ah r/
6	/iy/, /ih/
7	/uh/
8	/uw/
9	/eh/
10	/oh/, /ao/
11	/ax r/
12	/l/
13	/r/
14	/b/, /p/, /m/
15	/t/
16	/d/, /n/
17	/k/, /g/, /hh/, /ng/
18	/f/, /v/
19	/s/, /z/, /sh/, /zh/, /ts/, /dz/, /ch/, /jh/
20	/th/, /dh/
21	/y/
22	/w/
0	/#/ 無音

VISEME No.	音素表記
1	/a/
2	/i/, /y/
3	/u/
4	/e/
5	/o/
6	/r/, /ry/
7	/b/, /p/, /m/, /by/, /py/, /my/
8	/t/
9	/d/, /n/, /ny/
10	/g/, /k/, /N/, /hy/, /gy/, /ky/
11	/f/
12	/j/, /s/, /z/, /ch/, /dy/, /sh/, /ts/
0	/#/ 無音

音素表記	VISEME No.
/aa r/	4+2
/ia/	6+5
/ia r/	6+5
/ua r/	8+11
/ea r/	9+11
/aw/	4+8
/ey/	9+6
/oy/	5+6
/ow/	5+8
/ao r/	5+2
/ay/	4+6

8-2. VISEMEにより音韻分類

8.2.1VISEMEの定義

VISEMEとは、音素である”phoneme”から作られた造語であり、音声学的に異なった音であっても同一言語の中で同一音とみなされる最小の単位の意である。また、phonemeが決まればVISEMEは一義に求まる関係にある。本研究では、発音記号をVISEMEに基づき。英語については22種類の音素表記に、日本語については13種類に分類し、さらに無音区間を加えた計36種類の基本口形をデータベースとして用いた。本来VISEMEは、発音記号[au][ei]等に現れる口唇運動の情報まで定義されるのであるが、本研究においてはそのようなVISEMEは複合VISEMEとしてさらに分類し、運動情報を所持しない、形状の情報のみで分類した。表3 に本研究で分類したVISEMEと定義した英語発音表記の対応を、図11に3次元頭部モデルで表現される音素表記/ae/の基本口形を、例として示す。

8-3. 口形状の補間

システムの口形状データベースには36種類の基本口形があることは前章でも触れたが、ある基本口形から次の基本口形に移行するまでの口形データは存在しない。

本章では、もう一つの音声パラメータである、音素継続長情報より、基本口形間の補間を行う手法について説明する。

発声された音素が継続している間は基本口形の要素を持ったベクトル移動量の情報がワイヤーフレーム上に存在しなければならない。本研究において、音素が発声される開始時は、基本口形状を構成していると定義した。

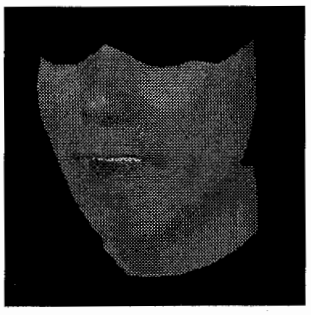

図11. /ae/の口領域モデル

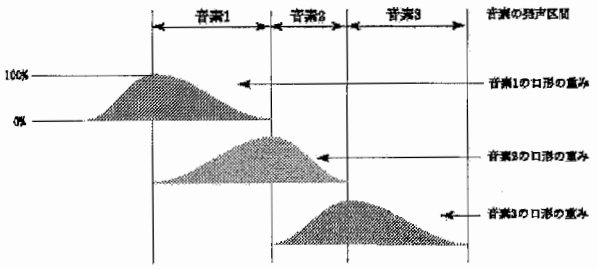

図12. 正弦波補間の概念図

従って、図12に示すように、音素継続時間の始点における、基本口形状を構成する格子点のベクトル移動量を100%とするとき、音素継続時間の終点では0%になるように正弦波波形を用いて補間を行っている。同様に、現時点で扱っている音素の次に現れる音素についても、現音素の継続時間長を基に、格子点ベクトルの移動量を0%から100%に補間する。こうして得られる時系列上の2つのベクトル移動量を加算した値が基本口形間におけるワイヤーフレームモデルを変形するためのベクトル移動量となる。すなわち、データベースに存在しない口形状も算出することが可能となる。本手法は人間の発話時における、骨格や筋肉の運動に直接的には結びついていないが、口唇運動を近似的に再現できていると考える。

9. 出力合成動画像の生成

9-1. 透過率を変化させたモデル画像の作成

3次元モデルによる口領域合成画像をビデオフレーム画像に重ね合わせる際、通常の画像の置き換えだけでは画像境界が発生する。これを避けるために、まず入れ替えるモデルと同型のマスク画像（輝度値のみの画像）を用意し、モデルの境界判定を行う。境界と判定された場合、モデルテクスチャの透過率を徐々に変化させた合成画像を作成する、これをビデオフレーム画像に重ね合わせることで画像境界の無い、自然な合成画像を得ることができる[図13]。

9-2. 発話判定結果を用いた合成動画像の作成

複数人の話者のモデルに対して、5-5節で説明した発話判定の結果を利用し、画像フレーム毎に口形状の補間法を用いて口形状画像を生成する。このとき、発話が認められていない話者のモデルに関しては、無音状態が続いていると定義し、無音の口形を生成する。また音声の無音区間についても同様の処理を行う。そして最終的に得られた合成画像を30[frame/sec]で出力する。以上により、複数人の話者が会話しているようなシーンに対しても、画像の翻訳が可能となる[図14]。また、入力音声において、発話判定の結果を本来ならば発話していない話者の口形状モデルに対して適用し、あたかも話者の声を入れ替えたようなシーンも作成することが可能である。http://www.ee.seikei.ac.jp/~akinobu/video_translation/demo/に本システムで作成した合成動画像があるので、そちらを参照していただきたい。

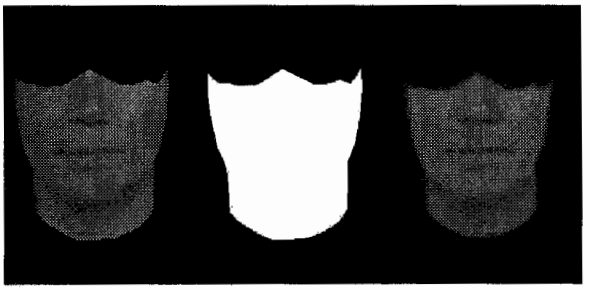

図13. 透過率を変化させたモデル画像

10. 考察

ここでは、本研究で用いた画像翻訳システムに対する考察を行う。まず挙げられるのが、口領域以外の調音器官の動きである。従来のビデオ翻訳システムでは、話者が一人であったがために、発話内容を翻訳し、口領域の合成画像を置き換えても不自然さがあまり目に付かなかったものが、話者が複数人になったことにより、そのような不自然さがより顕著になった部分がある。これは、実際にモデルを置き換えている部分は口領域だけであり、それ以外の部分は翻訳の対象になっていないために生じていると考えられる。例えば、話者Aが発話している区間を話者Bの口領域に反映させたときにより鮮明になる。つまり、翻訳の対象となる話者の発話時における、喉仏の動きや、瞬きなどの発話に関係する動きも考慮に入れなければいけないことを示している。これは今後の課題である。また、自動顔トラッキングに関しても、いまだに制約が多く、また処理時間も莫大なものであるため、精度の向上および高速化が求められる。また、発話判定も完全なものではなく、あるフレームにおいて話者のフレームパワーが、本来発話している話者のフレームパワーを越えてしまった際に、数フレームにわたって誤判定を起こし、異なる話者が発話を始めてしまうといった現象を起こしてしまう。現状のシステムでは、音声区間において発話しているものは一人と限定しているので、音声のフレームパワーを用いて、これが発話とみとめられるような時にのみ、画像側のフレーム間エネルギーが最大となる話者をその音声区間における話者として認めるようにすることで判定の誤りを防げると考える。また、唇領域の画素列の抽出法についても今後考慮していく必要がある。

11. 結論

本手法により、複数人が会話しているシーンに対しても画像翻訳が可能となった。これにより、従来のビデオ翻訳システムを複数人の会話に対しても適用することができ、本システムと、自動音声翻訳器とを組み合わせることにより、複数人の人物が会話しているようなビデオ翻映像に対する、日英双方向のビデオ翻訳を実現することできるだろう。

図14. 合成動画像の1フレーム

12. 参考文献

[1] 伊藤，三澤，武藤，森島.，「複数アングル画像からの3次元頭部モデルの作成と表情合成」.，電子情報通信学会技術報告，Vol99，No582，pp7-12，2000

[2] 緒方，三澤，森島，村井，中村.，「ビデオ翻訳システム－自動翻訳合成音声とモデルベースリップシンクの実現－」.，情報処理学会インタラクション2001，pp201，2001

[3] 村井，松井，Reiner，中村.，「口周囲画像による発話の検出」.，情報処理学会　2000年秋期全国大会予稿集

[4] http://www.w3.org/Graphics/Color/sRGB