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 -

緒方　信[†‡]　　中村　哲[‡]　　森島　繁生[†]

†成蹊大学工学研究科　　‡ATR 音声言語通信研究所

あらまし　本稿では、従来より研究されてきた音声翻訳技術に加え画像をも翻訳する、日英双方向翻訳システムを紹介する。本手法は顔画像翻訳において、話者の表情を保つ為に口やその周囲の情報以外は原言語発話時の動画像をそのまま用い、口領域については任意の話者に適合可能な3次元ワイヤフレームモデルを用意し双方を合成することを試みた。この手法により、小規模なデータベースより顔画像の合成、翻訳が可能となった。

Multi-modal Translation System
- Model Based Lip Synchronization with Automatically Translated Synthetic Voice -

Shin OGATA[†‡]　　Satoshi NAKAMURA[‡]　　Shigeo MORISHIMA[†]

†Faculity of Engineering, SEIKEI Univ.　　‡ATR Spoken Language Translation Research Lab.

Abstract　In this paper, we introduce the multi-modal English to Japanese and Japanese to English translation system, which translates the speaking motion synchronized to the translated speech. To retain the speaker's facial expression, we substitute only speech organ's image with the synthesized one, which is made by a three-dimensional wire frame model that is adaptable to any speaker. Our approach enables the image synthesis and translation with extremely small database.

1．はじめに

音声翻訳の研究は、あらゆる言語間で、またさまざまな目的に応じて盛んに行われており、その発展は目覚しいものがある。

1993年に発足したATR音声翻訳通信研究所における研究の結果、話題の対象が限定されるなどの一定の条件下で、異なる言語での対話を翻訳するシステム(ATR-MATRIX[1])として利用可能であるという段階に達している。そしてこの分野の研究は、2000年に発足したATR音声言語通信研究所においても引き継がれ、より広いドメインにおける日常の自然な話し言葉に拡張されることが期待されている。

音声翻訳技術は韻律情報等を除けば、主に言語情報を扱う研究分野として発展してきた。しかし言語情報は、意思疎通の為のひとつの手段に過ぎず、Face-to-Faceのコミュニケーションにおいて、顔は言葉と共にさまざまなメッセージを伝えている。例えば映画などにおける吹き替えでは、音声のみを翻訳している為、口の動きと発話内容が一致しないという課題がある。また顔画像全体をコンピュータグラフィックスにより合成した場合、ノンバーバルな情報を再現して伝えることが困難となる。これらの課題を克服し音声翻訳と共に画像の翻訳、つまり話者の表情を保ちつつ口形状の翻訳が可能となれば、より親しみのあるコミュニケーションを実現できるであろう。

口形状を変形した顔画像を生成する研究[2]は、過去にもアプローチはいくつかあった。しかし、画像は音声に比べ情報量が多い為、大規模データベースを用意するのは困難であり、話者が限定される等の汎用性が乏しいという制約があった。

そこで本研究では、話者の表情を保つ為に、口やその周囲の情報以外は原言語発話時の顔動画像をそのまま用い、口領域については任意の話者に適合できる3次元モデルを用意し、双方を合成することを試みた。3次元モデルを用いれば、音声合成に用いた音素の表記と継続長情報を基に口形を生成することができ、顔の位置や向きにも対応する合成画像を得ることが可能である。また3次元モデルには画像データベースは必要がなく、用意するのは口形状を表現する為のワイヤフレーム格子点のベクトル移動量のみである。これにより、比較的小規模のデータベースによって音声のみならず顔画像をも翻訳できる、ビデオ翻訳システムが可能となる。

本稿ではまず、このビデオ翻訳システムの全体像について触れ、顔画像合成における3次元口形モデルの生成について記述する。次に音声合成部より

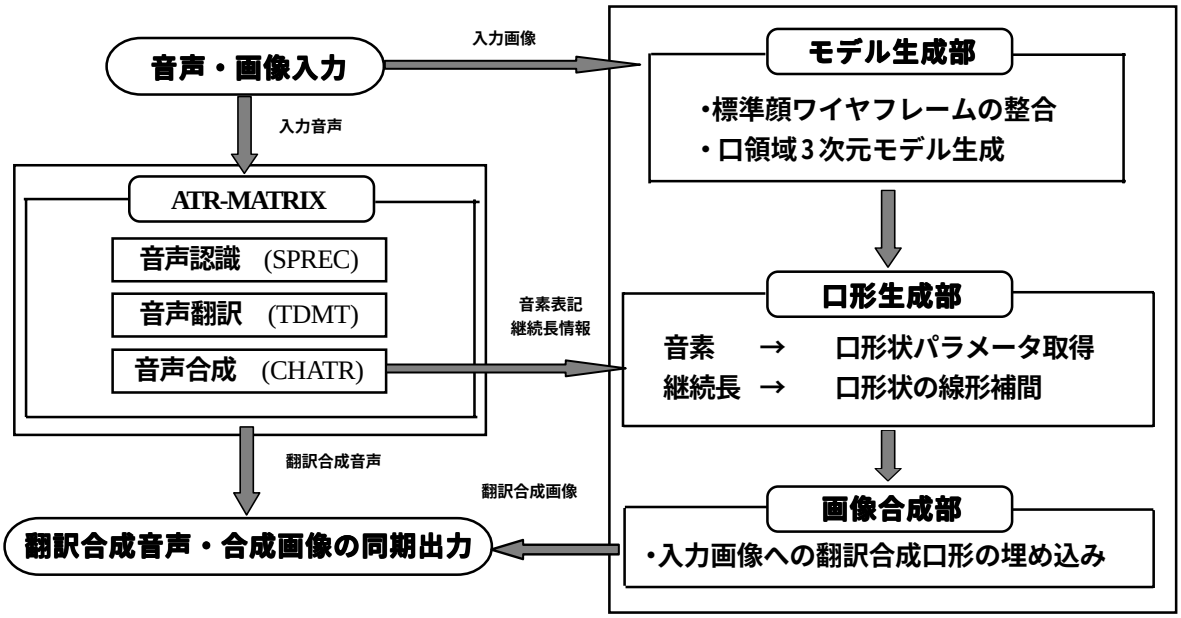


図1　システム全体像

得られる音素表記と継続長情報の2つのパラメータから、発話に対応する口形をモデル上に生成する手法について説明する。その後モデルと入力画像の合成手法について触れ、そして最後に、このシステムにおける研究課題について考察する。

2．システム全体像

図1に、本研究におけるシステムの全体像を示す。システムは大別すると音声翻訳部と画像翻訳部に分かれている。音声翻訳部はATR音声翻訳通信研究所で開発されたATR-MATRIX[1]により行われる。ATR-MATRIXは、音声認識を行うSPREC、テキストに変換された言語情報を翻訳するTDMT、翻訳されたテキストから合成音声を生成するCHATR[3]より構成されている。このうちの翻訳合成音声を生成するCHATRより出力される音素表記と音素継続長の情報は、顔画像翻訳に利用される。

画像翻訳部における第一段階は、入力画像から標準顔ワイヤフレームを整合することにより、話者別の口領域の3次元モデルを生成することである。話者により顔面の骨格が異なる為に、個人ごとにモデルを生成しておかなければならないが、この工程は話者1人につき1度踏まえればよい。

画像翻訳部第二段階は、発話に対応する口形生成部である。音声合成に用いた音素表記から、各音素に対応する口形状パラメータをデータベースより取得し、口領域モデルを変形させる。また音素継続長情報は口形状の線形補間に利用する。このときに使用する口形状パラメータは音素ごとに定めた口形状のワイヤフレームのベクトル移動量としている為、話者に依存することはない。

画像翻訳部の最終段階は、入力画像に3次元口形モデルを埋め込む画像合成部である。この工程でモデルと入力画像の色、スケールを一致させる。入力した顔画像が発話時に運動していても、モデルは3次元情報を所持している為、自然な画像合成を行うことが可能である。

システムの最終工程では、翻訳された合成音声と合成画像を30[frame/sec]で同期させて出力する。

3．口領域の3次元モデルの生成[5][6]

3-1.3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、口等の要素を構成する形状や位置

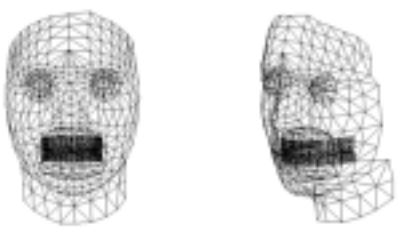


図2　3次元頭部モデル

(a) 入力画像　(b) 整合　(c) 口領域モデル

図3　生成された口領域3次元モデル

は、個人によって微妙に異なる。CG(Computer Graphics) 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でかつ演算量の少ない3次元モデルを構成することが必要となる。

本研究では、成蹊大学情報通信研究室において研究・開発されている、3次元頭部モデル[図2]を用いて、口領域の3次元モデルを生成することを試みた。本研究では顔領域全体の整合を行った。

このモデルは約1500ポリゴンの三角形パッチより構成されていて、格子点数は約800からなる。

3-2. 口領域モデルの作成過程

3-1節で導入した3次元頭部モデル上に、人物画像の口領域を正確にテクスチャマッピングする為には、ワイヤフレームモデルと対象人物の入力画像の整合を行わなければならない。整合は、任意方向から撮影した複数画像と、専用のGUIツールを用いて行い、この工程を経てモデルに3次元情報を付加することが可能である。現状では、整合に多くの画像を用いるほど精密なモデルを生成することができるが、人手による作業量も増すことになる。

本研究においては、話者1人につき、正面・側面に加えさらに2つの斜め方向より、計4方向から撮影した画像を用いて、口領域モデルの整合を行った。図3(c)が、入力画像(a)の人物の生成された3次元口領域モデルを表している。

4. 発話口形の生成[4][6]

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。

本研究と同様、発話顔動画像の生成をモデルベースからアプローチしている、文献[4]の報告では、被験者の口領域にマーカを置き、3次元的移動量を計測することで運動学的データを採取し、アニメーションに再現する方法を試みている。しかし今回は、話者への依存性を削減するという点から、次節に示す手法を用いた。

4-1. 標準口形状データの設定[6]

口領域の動きを定量的に表現する為に、文献[6]では口領域の制御点として図4のように7点を定めている。各々の制御点はワイヤ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フレームモデルをその参照画像に近づけるよう変形させたとき得られる、各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為、一度基本口形を用意すれば、すべての話者に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。

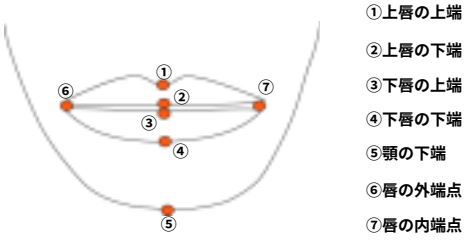


図4　制御点の位置

4-2.VISEME による音韻分類

4-2-1.VISEME の定義

VISEME とは、音素である “ phoneme ” から作られた造語である。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音単位の意である。例として、英語における “ me ” と “ knee ” という単語の発音は、騒音下などでは音のみよって区別するのは非常に困難である為、同一音とする場合があるが、同一音であっても視覚的要素によってどちらの単語であるか区別するのは容易である。もし話者の口が閉じていればそれは “ me ” であればあり、そうでなければ “ knee ”である。またそれとは逆に、“ bat ” と “ pat ”のような単語にみられる、視覚では区別できず聴覚によって区別が可能な /b/, /p/ 等の音韻を「視覚素」、すなわち VISEMEと呼ぶ。

本研究では、音声合成部CHATRより出力される音素表記をVISEMEに基づき、英語については22種類に、日本語については5母音をそれとは別に分類し、さらに無音区間を加えた計 28 種類の基本口形をデータベースとして用いた。

表1　VISEMEの分類

VISEME No.	CHATRより返還される音素表記
1	/ae/　　英語
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/
23	/a/, /A/　　日本語
24	/i/, /I/
25	/u/
26	/e/, /E/
27	/o/, /O/
28	/#/　　無音

図5　音素表記/ah/に対応する基本口形

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としてさらに分解し、運動情報は所持しない形状の情報のみで分類した。複合 VISEME については後述する。

表1に本研究で分類したVISEMEとCHATRの音素表記の対応を、図5に3次元モデルで表現される音素表記 /ah/ の基本口形を、例として示す。

CHATR では英語合成音声に、英国英語（British English）と米国英語（American English）を用意している。音声合成には、それぞれ別々の音素辞書を用いるが、本研究では British English の音素辞書を VISEME に対応付けた。また日本語の音素辞書には、外来語（カタカナ語）に多く用いられる「ヴェ」、「デュ」等の音素が存在するが、これらについては対応付けるには至らなかった。その他、共通にみられる「笑い声」や「いびき」等についてもCHATR側には表記として存在したが、本研究においては対応付けはしていない。

4-2-2. 英語における複合 VISEME

英語には、音素表記1つに対して時間的に複数の基本口形から構成されるVISEMEが存在する。発音記号 [au] や [ei]、[ou] 等の音素がそれに当たる。本研究ではこのようなVISEMEに対して基本口形単位で分解し新たに分類し、複合VISEMEと呼ぶことにする。表2はCHATR より出力される、これらの複合VISEMEで表される音素表記を示す。

音素表記は個別に音素継続長の情報を持っている。しかしこのように音素表記が複合VISEMEと対

表2　英語における複合VISEME

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

応する場合はその継続長情報についても基本口形の個数によって分解する必要がある。

本研究では、音素表記が2つの基本口形から構成される場合において、前半に現れる基本口形に30%の音素継続時間を、後半に残りの継続時間を経験的に割り当てた。これはあくまで経験的に定めた値であるが、音素別に特徴を定量化できた場合、データベースを容易に変更することが可能であり、この点は本手法の特徴であるといえる。

4-2-3. 日本語の子音分類

日本語の子音は英語に現れるものより少ない為、本研究では英語のデータベースより引用した。しかし一般的に日本語の子音口形は英語に比べ運動の変化が少ないことが知られている。そこで今回、データベースより引用する日本語の子音基本口形は、英語の基本口形の移動量の 60% におけるものと定めた。

また日本語では、文末の母音が無声化することが多い。CHATR では、母音「う」の音素表記に有声音 /u/ と無声音 /U/ がある。そこで本システムでは無声化した場合の唇の移動量を考慮にいれて、子音のときと同様に、 /u/ の基本口形 60% を /U/ の基本口形とした。

さらに日本語子音の特殊な例として、「は行」がある。発音記号 [h] に表される子音は、主に口内で生成される音である為、唇の動作等の視覚要素に反映されることは少ない。これを考慮に入れ、システムにおける音素表記 /h/ に対しては、後に続く母音基本口形を割り当てた。

4-3. 口形状の補間

システムの口形状データベースには28種類の基本口形があることは前節までに触れた。しかしある基本口形から次の基本口形に移行するまでのデータは存在しない。

本節では、音声合成部より出力されるもう1つのパラメータである音素継続長情報より、基本口形間の線形補間を行う手法を述べる。

4-3-1. 口唇運動の軌跡

人間が言葉を話すとき、唇は絶えず運動をする。しかし同じ発話内容であっても、1音ずつ音節を区切りながら発音するときと、文章として発話するときとでは口唇運動の軌跡が異なる。これは人間が滑らかに発話する場合、口唇の運動軌跡は効率良く推移していく傾向がある為である。

本システムに使用する口形状データベースは、基本口形に変形させたワイヤフレームの格子点のベクトル移動量で定義されている。そこで、ある基本口形の移動量と次に現れる基本口形のベクトル移動量を加減算することにより連続的な口唇運動を実現する手法について次節で説明する。

4-3-2. 口形状の線形補間

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モデルワイヤ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ものと定義した。従って図6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

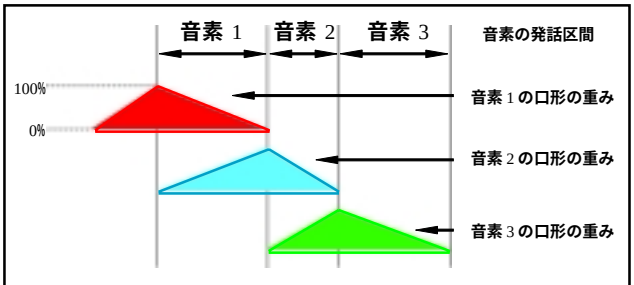


図6　口形状の線形補完に用いた重み付け

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%になるように線形補間を行った。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、格子点のベクトル移動量を0%から100%に線形補間する。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フレームを変形する為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はいないが、口唇運動を近似的に再現できるものと考える。

4-3-3. 合成音声との同期

システムの最終工程では合成音声と合成画像を同期させなければならない為、口領域モデルもそれに合わせて生成する必要がある。

音声合成部CHATRで扱う最小時間長は、1[msec]である。そこで本研究では、前節で説明した線形補間の手法を用いて、事前にベクトル移動量の百分率を1[msec]単位で算出した。最終的に30[frame/sec]で動画を生成する事を考慮し、33[msec]毎にそのベクトル移動量を改めてサンプリングする。演算量を少なくする為に、口領域モデルの生成に必要なデータのみを操作するのである。

実際には、100[msec]毎に1[msec]の割合でこのサンプリングされたベクトル移動量の補正を行うことにより、時間的誤差の減少を図っている。

図7　画像合成部のGUIツール

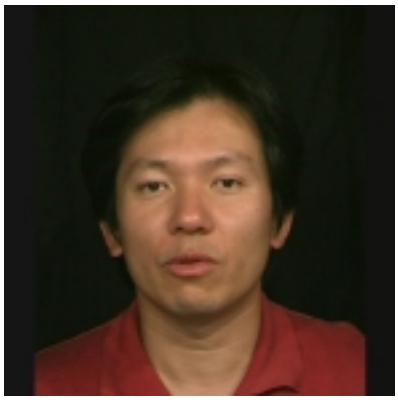
5. 出力合成画像の生成

本システムの画像合成部には入力画像に対して口領域モデルのスケールを一致させる工程と入力画像とモデルの色調補正を行う工程、そして入力動画像にモデルを追跡し自然な合成画像を得る工程があるが、これらの工程は現時点では全て人の手による補正が必要となり、完全な自動化には至っていない。

そこで本研究では、この画像合成の工程に必要な一連の機能を備えたGUIを開発し[図7]、このツール上で合成画像の生成を行った。ツールにはCHATRより出力されるパラメータファイルを読み込む機能が実装されており、口形状は自動で生成することが可能である。

(a) 補正前　　(b) 補正後

図8　色補正を行った例



(a) 入力画像

(b) 翻訳合成画像

図9　入力画像と翻訳合成画像

5-1. 口領域モデルの色補正

翻訳された合成口形を入力画像に埋め込む際、話者別に生成したモデルは、入力画像の撮影条件によっては色補正を必要とする。より自然な画像に見せる為、色補正の後、口領域モデルと入力画像の境界は、色調の透過率を徐々に変化させた[図8]。

色補正の手法に、モデルと入力画像の色ヒストグラムの平均値の調整と、$L^*a^*b^*$(注1)を用いた光源情報の補正を用いたが、今回は完全な自動化には至っていない。今後は補正手法についても検討の必要があるものと考える。

5-2. 入力画像の間引き・繰り返し

入力画像は入力音声の継続時間長分の情報を所持している。しかし、言語を翻訳することにより、音声の時間長は変化する。

本研究では、入力動画像を連続する静止画像の時系列、画像シーケンスとみなし、翻訳された合成音全体の継続時間長から、合成時に使用する入力画像のシーケンス数を操作する手法をとった。合成音声が入力音声に対して短くなる場合には、一定の割合で画像シーケンスを間引き、反対に合成音声が入力音声に比べて長くなる場合においては、時系列に沿って一定の割合で画像を繰り返し用いることで、合成音の継続長と合成画像の継続長を調整し、同期出力させた。

5-3. 翻訳合成口形の埋め込み

前節までの工程を経た口形状モデルは、3次元形状、発話口形、発話時間長、スケール、色の情報を所持している為、入力顔画像に対して自然な合成を得ることが可能である。図9に合成画像の1例を示す。

モデルは対象人物の鼻の下から喉仏までの情報を所持している。入力画像の口形はモデルによって覆い隠される為、これより入力画像の原国語の発話口形に依らず翻訳画像を生成することができる。

また、モデルに覆い隠されない顔の部位に、ノンバーバル情報が現れている場合、その情報を維持することが可能であるのも、本手法の特徴である。

4章で述べたように、発話時間に応じて口形の補間が行われ、その情報に基づき画像シーケンスが生成される。画像シーケンスを30[frame/sec]でCHATRの生成する合成音声と同期させ出力することにより、翻訳合成動画像を得る。

以上の処理を用いて、日本語から英語の翻訳、反対に英語から日本語の双方向の翻訳が可能なシステムを構築した。

6. まとめ

本研究の成果として挙げられるのは、小規模な口形データベースから話者に依存することなくさまざまな発話口形画像を生成できることと、口領域以

(注1) CIE$L^*a^*b^*$ 色空間を使用。

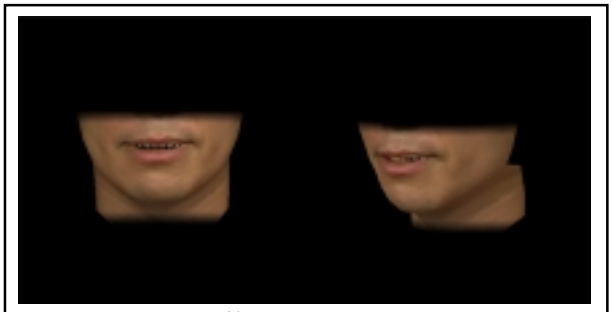


図10　現状における/th/の口形

図11　入力画像に対してモデルが適合しない例

外は入力動画像を用いることで、ノンバーバルな情報を維持できたこと、それによって翻訳合成音声を生成する際に出力されるパラメータから、日英双方向ビデオ翻訳システムを実現できたことである。

さらに、今回の研究において確認できた今後の研究課題について述べる。

まず、口形状モデルについては現在、舌のモデル化が行われていない為、英語の[th]の発音等は不完全である[図10]。舌モデルも唇と同様にパラメータ制御が可能なモデルを作成することが必要と考える。また歯のモデルについては、スケール変換はできるが、未だ人工的な印象を与える。これについては、ライティングの設定或いはテクスチャマッピングを施すことで改善できるものと考える。

今回、入力画像に対するモデルのトラッキングは全て手作業で行った。自動トラッキングの分野の研究は盛んである為、システムに取り入れる余地はあると考える。

そしてまた、今回のシステムは全てオフラインで行ったに過ぎない。このシステムのリアルタイム化の実現にはまず第1に、画像処理の高速化が必要である。その1つの方法として提案するのが、入力動画像のキャプチャと口領域モデルの完全分離処理である。また、オンラインでシステムを稼動させる為には、今回のように入力画像の継続時間を調節するのではなく、音声合成側の継続時間を操作する手法を確立することが必要であると考える。

また発声方法や発話様式、話者の感情も考慮したシステムを構築する必要がある。図11に示すのは、入力画像において話者が笑顔であるとき、合成画像に笑顔を再現できていない状況である。これについては、感情によってワイヤフレームの移動量を新に定義することを考えている。

本手法に用いた基本口形を構成するベクトル移動量を定義した値、線形補間法についてもまた、更なる改善の余地があるものと考える。

7. 参考文献

[1] 菅谷，竹澤，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

[2] Hans Peter Graf, Eric Cosatto, Tony Ezzat
“Face Analysis for the Synthesis of Photo-Realistic Talking Heads”
PROCEEDINGS FOURTH INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION 28-30 MARCH,2000.,GRENOBLE,FRANCE pp189-194

[3] Nick Campbell, Alan W. Black
“Chatr : a multi-lingual speech re-sequencing synthesis system”
電子情報通信学会技術研究報告， sp96-7, pp.45, 1995

[4] T. Kuratate, H. Yehia, E. Vatikiotis-Bateson
“KINEMATICS-BASED SYNTHESIS OF REALISTIC TRACKING FACE”
International Conference on Auditory-Visual Speech Processing - AVSP’98, pp.185-190, 1998

[5] 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研究報告， Vol99, No582, pp7-12, 2000

[6] 伊藤，三澤，武藤，森島
「仮想空間上におけるリアルな三次元口形状の作成」
電子情報通信学会総合大会， A-16-24, pp328, 2000