Internal Use Only （非公開）

TR-S-0009

音声合成に合わせた顔画像翻訳技術の研究

On multi-modal translation system

緒方信 Shin Ogata
村井和昌 Kazumasa Murai

2000.9.29

成蹊大学森島研究室と共同で、音声翻訳システムに、話者の口元の動きを、翻訳した言語に合わせて動かすマルチモーダル翻訳システムを開発した。

©2000 ＡＴＲ音声言語通信研究所
©2000 by ATR Spoken Language Translation Research Laboratories

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。

しかし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。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顔の情報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。

本研究では、従来よりある音声翻訳システムに加え、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次元モデルとして用意し、双方を合成することを試みた。3次元モデルは、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。これにより、音声のみならず顔画像をも翻訳できるシステムの実現性を示すものである。

本稿ではまず、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得られる音素表記と継続長情報から、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。音声翻訳部は従来のATR-MATRIXにより行われる。このうちの翻訳合成音声を生成するCHATRより返還される音素表記と音素継続長の情報は、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移動量としている為、話者に依存することはない。

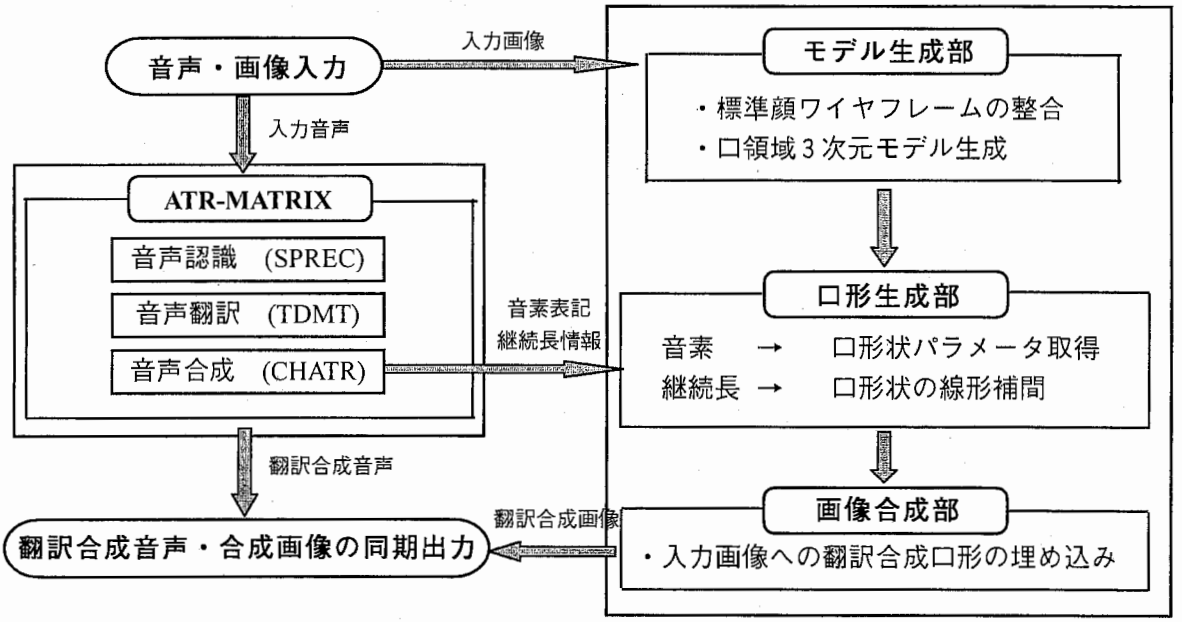


図1　システム全体像

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[][]

3-1. 3次元頭部モデル

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。

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。

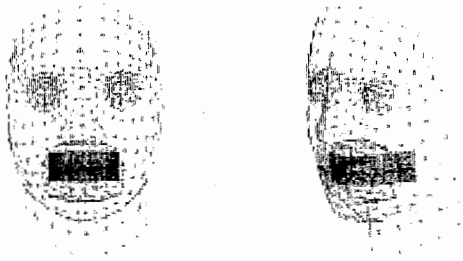
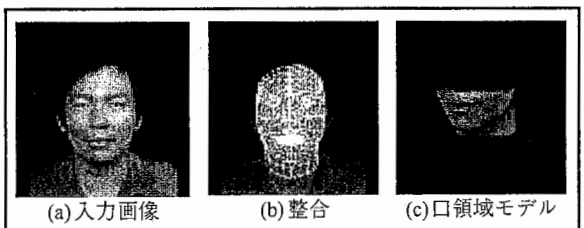


図2　3次元頭部モデル

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を用いることにより、モデルに3次元情報を付加することが可能である。本研究においては、話者1人につき、正面・側面・斜めの3方向より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]。



(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

4. 発話口形の生成[][]

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、正確な定量化が必要である。

4-1. 標準口形状データの設定[]

口領域の動きを定量的に表現する為に[4]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、口領域の大きさより正規化された、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。このデータは正規化した移動量としている為、話者適用を必要としない。

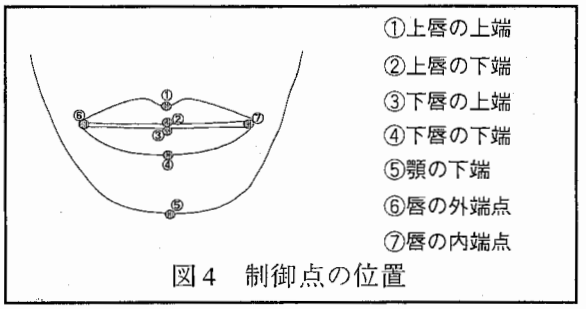


図4　制御点の位置

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1国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より返還される音素表記をVISEMEに基き、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にその分類を示す。

CHATRでは英語合成音声に、British Englishと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。

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

4-2-2. 英語の複合VISEME

英語には、音素表記1つに対して複数のVISEMEから構成される音素が存在する。発音記号[au]や[ei]、[ou]等の音素がそれに当たる。表2はCHATRより返還される、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と対応する場合はその継続長情報をVISEME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つの複合VISEMEから構成される場合において、前半に現れるVISEIME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。

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

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運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、音声合成部より返還されるもう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の手法を述べる。

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音づ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、口唇の運動軌跡は最もエネルギー効率が良いように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移動量を加減算することにより、近似的に最良のエネルギー効率を得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

音素が継続している間は基本口形の要素を持った移動量がモデルワイヤフレーム上に存在する。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。従って音素継続時間の始点での、基本口形状を構成する格子点の移動量を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う。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、現音素の継続時間長を基に、格子点の移動量を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つの移動量を加算したものを、基本口形間におけるワイヤフレームを変形する為の移動量として算出する。この手法により、口唇運動の最良エネルギー効率を近似的に再現することが可能であると考える。

4-3-3. フレームレートの制御

システムでは最終的に音声と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。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]毎にその移動量をサンプリングし、口領域モデルの生成を行った。

実際には、100[msec]毎に1[msec]の補正を行うことで、時間的誤差の減少を図っている。

5. 出力合成画像の生成

本システムの画像合成部は、現時点では完全な自動化に至っておらず、人の手による補正が必要となる。本研究では、トラッキングやスケール変換等、画像合成に必要な一連の機能を備えたGUIを開発し[図5]、このツール上で合成画像の生成を行った。

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、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6]。

色補正はさまざまな方法があり色ヒストグラムの平均値やLabを用いた補正から行ったが、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。

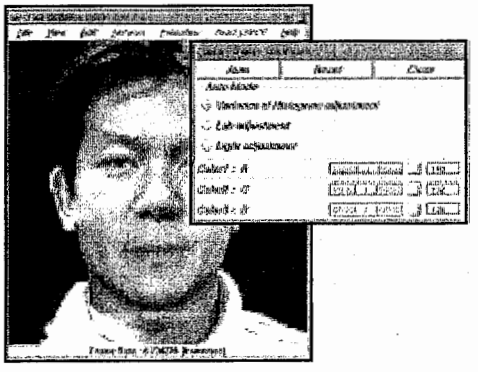


図5　画像合成部のGUIツール

(a)補正前　(b)補正後

図6　色補正を行った例

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静止画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、一定の割合で静止画像の列を間引き、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7に合成画像の1例を示す。

(a)入力画像　(b)合成翻訳画像

図7　入力画像と翻訳合成画像

6. まとめ

本節では、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。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。

さらに、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率することが必要であると考える。

最後に本研究の成果として挙げられるのは、少ない口形データベースで、さまざまな発話口形を生成でき、話者に依存しない画像翻訳システムとしては、妥当であり、発展性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] Nick Campbell, Alan W. Black
"Chatr : a multi-lingual speech re-sequencing synthesis system"
電子情報通信学会信学技報，sp96-7, pp.45, 1995

[3] 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

[4] 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