音声言語情報処理 15−17
（1997. 2. 8）

HMMを用いた音声と唇画像の統合による音声認識と唇画像生成

中村 哲、 山本 英里、 永井 論、 鹿野 清宏

奈良先端科学技術大学院大学 情報科学研究科

〒630-01 奈良県生駒市高山町 8916-5

E-Mail:nakamura@is.aist-nara.ac.jp

概要 ヒューマンマシンインタフェースのための要素技術の進展に伴い、これらの技術をどのように有効に統合するかということが問題となってきている。本稿では、音声情報と唇情報の統合的利用とその変換法について述べる。まず、音声認識に発話者の唇画像の情報を HMM を介して統合し、利用することで音声認識性能の改善を試みる。実験により、Tied-Mixture HMM の導入による画像認識精度の改善、統合方法による差、統合による認識精度の改善を示す。次に、HMM を用いて音声から唇画像を生成する試みについて述べる。HMM による方法では、入力された一発話の音声から最適な画像系列を生成する。さらに、後続音素に依存したモデルによる改善を行ない、唇パラメータの誤差評価により変換手法の有効性を示す。

Speech Recognition and Lip Movement Synthesis by HMM-based Audio-Visual Integration

Satoshi NAKAMURA, Eli YAMAMOTO, Ron NAGAI, Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama-cho, Ikoma-shi, Nara, 630-01 JAPAN

E-Mail:nakamura@is.aist-nara.ac.jp

Abstract While individual technologies for human-machine interface have achieved remarkable developments, how to integrate different modalities effectively becomes an urgent problem. This paper tries to apply the integration of speech and image to speech recognition and lip movement image synthesis. Integration based on HMM is used to improve lip image recognition. Experiment results show that Tied-Mixture HMM improves lip image recognition accuracy and that the speech and lip image integration improves speech recognition accuracy under various kinds of SNR environments. Lip movement image synthesis through HMM is also proposed. The proposed method is able to synthesize an optimal lip movement image sequence for the whole utterance. The method based on the right context dependent HMM is also proposed. Lip image synthesis experiment shows the effectiveness of the proposed methods.

1 はじめに

ヒューマンマシンインタフェースの高度化を実現するための要素技術の進展に伴い、これら要素技術をどのように統合するかの問題が重要な課題になり始めている。例えば、昨今の音声認識技術によれば、確率モデルの利用により静かな環境では、用途によっては充分な認識性能を得ることができる。しかし、雑音などにより環境が劣化した場合、著しい認識率の劣化を生じてしまう。この点、人間は種々の情報を統合的に利用し発話内容を理解している。音声が聞き取り難い場合、発話者の顔、特に唇の動きに注目して発話内容を理解しようとする。逆に、唇の動きと音声が不一致の場合、唇の動きに影響されて発話内容を誤って理解してしまう。このように、人間による発話内容の理解には、唇の画像と音声の情報の統合的利用が極めて重要であり、自動認識装置においてもこの統合的利用により環境に頑健な理解が可能となることは、疑いのないところであろう。このような、複数の感覚情報の統合に関しては、日本では平成3年から平成7年までセンサーフュージョンプロジェクトのものとで、種々感覚情報を種々のレベルで統合する研究が行なわれた。また、欧州でも、ヒューマンマシンインタフェースに焦点を当てて、ESPRITの中の一つのプロジェクトMIAMI(Multimodal Integration Advanced Multimedia Interface)などで種々の研究が行なわれている。これらのプロジェクトにより、人間の情報処理における感覚統合の方法の解明、感覚統合の工学的応用の研究が進められている。著者らは、これらの研究成果を鑑みながら音声認識の性能向上を目指したマルチモーダル情報の統合およびその逆である他モードの情報の生成について検討を行なっている。本稿では、特にHMMに基づいた音声認識における唇画像の情報の利用、さらに音声情報から唇画像への変換法について述べる。

2 音声と唇画像による音声認識

2.1 音声と唇画像の情報の統合

音声認識は、HMMの利用により認識性能が改善されたが、それでも前章に述べたように雑音や残響の影響で認識性能は大きく劣化してしまうことが知られている。人間は、Face-to-Faceで対話を行なう時、雑音や残響などで聞き取り難い環境の時、知らず知らずのうちに発話者の唇を注視して懸命に発話内容を探ろうとする。この極端な場合が、音声情報が全く聞こえない場合である。このことから、このようなマルチモーダルな情報の統合的利用を認識アルゴリズムに応用して、音声認識性能を改善できないかという期待が生まれてくる。実際、極めて雑音の少ない環境や、電話回線で相手の顔の情報が全く利用できない場合を除いて、通常のヒューマンマシンインタフェースにおいては、種々のモードの情報の併用が可能である。これまで、このような観点から1984年のPetajan[5, 6]の論文を始めとして、ニューラルネットを利用するもの[7, 8, 9, 14, 15, 16]、HMMを利用するもの[10, 11, 12, 13, 17, 18, 19, 20, 21]など数々の研究が行なわれている。特に、最近HMMに基づいて音声と唇画像を統合して音声認識の性能改善を試みる研究が盛んになってきている。この理由としては、HMMの標準ツールの流通や、データベースの共用があげられる。しかしながら、HMMの学習のための音声と画像が同期したデータベースや、情報の統合方法、モデリング方法は充分とはいい難い(例えば[4])。本章では、日本語の音声画像同期データベースの収録、さらにサブワード単位のHMMを用い、唇画像の情報を音声に加えて利用することによる日本語単語音声認識の改善の試みについて述べ、その結果について報告する。

2.2 方法

前節で述べたように、画像と音声の統合による音声認識にHMMが用いられているが、口の形状の特徴抽出の問題、データベースのサイズの問題、データベースが小さいことによるHMMの学習精度の問題がある。そこで本稿ではこの問題に対処するため、次のような方法で解決を試みる。

- 微妙な位置のずれを吸収するため、口の周辺画像を2次元FFTし、2次元パワースペクトルを特徴量とする。
- 大量の音声と同期した画像のデータの収録
- Tied-Mixture HMMによる学習・認識

実際に認識を行なう際、どのように画像と音声を統合法するのかが非常に重要な問題となる。ここでは、代表的な次の2つの方法を比較評価する。

1. 初期統合 (Early Integration)
   初期の段階で統合する方法。具体的には、音声および画像のパラメータを独立のパラメー

タストリームとしそれぞれの HMM の出力確率の積を各状態で求めてその状態の出力確率とする。この際、各ストリームの出力確率の冪乗の重みを与える。

2. 結果統合（Late Integration）
上記の方法と反対に、結果を統合する方法。つまり、音声と画像それぞれで学習および認識を行ない、音声と画像と別々にすべての単語に対する尤度を計算しておく。最後に、同一の単語に対する音声の対数尤度と画像の対数尤度を重み付けして加算しその単語の対数尤度とする。

2.3 実験

データベースとして、ATR の日本語データベース (SetA) の語彙を使用し、特定話者 1 名の 5240 単語を収録する。その際、頭を固定し、口の周りの画像のみを撮影する。同時に、同期をとりながら、音声も収録する。この際、画像のフレーム周期は 33.3msec で、音声のフレーム周期は、8msec である。ファイルフォーマットには AVI ファイルを用い、同期情報を保ったまま格納する。処理の手順は、まず、各フレームごとの JPEG 画像（160x120）を、256 階調の濃淡画像に変換する。その画像に対して、256x256 で 2 次元 FFT を行なう。ここで、空間周波数領域におけるパワースペクトルを計算し、対数スケールのスムージングを行なう。さらに、フレーム間の差分をとることで、動的な特徴を求める。HMM は音素単位で、パワースペクトルに 256 分布、その時間差分に 256 分布の Tied Mixture を用いる。音声については、22.050kHz の音声を 12kHz にダウンサンプリングして用いる。パラメータとして、メルケプストラムを用い、16 次で分析した後、3 ストリーム 33 次元 (16MFCC+16 Δ MFCC+Δ Power) の特徴パラメータとする。HMM は画像と同様、monophone HMM で、MFCC に 256 分布、デルタ MFCC に 256 分布、Δパワーに 128 分布の Tied-Mixture としている。また、音素は統合するために、音声・画像ともに 55 音素、状態数は音声が 3 状態・画像が 2 状態、用いる分布は Tied Mixture、学習単語数、認識単語数は、それぞれ、4740 単語、100 単語とした。初期統合法による認識の際には、音声と画像のフレーム周期が異なるため、音声 4 フレームの間に同じ画像を 4 フレーム挿入する。また、特徴ベクトルは、音声 3 ストリーム、口の形 2 ストリームに分け、音声と画像の重みをさまざまな値に変化させ、HMM で学習および認識を行なう。

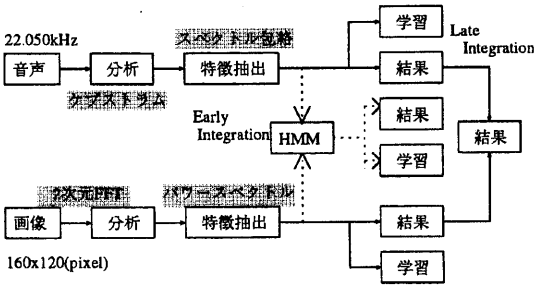

図 1: 処理方法

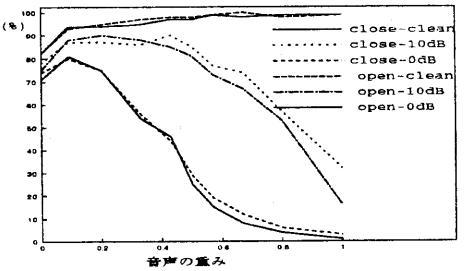

図 2: 初期統合（EARLY INTEGRATION）

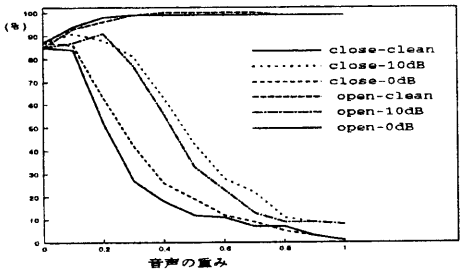

図 3: 結果統合（LATE INTEGRATION）

2.4 結果

まず、HMM の構造による性能差としては、Tied-Mixture の導入により、4 混合の混合ガウス分布 HMM と比較して、画像認識精度が学習データで 83%から 87%、テストデータで 84%から 85%への改善が得られた。次に、統合に関する実験結果を図 2および図 3に示す。この結果から、統合することにより、特に SNR が劣化した場合に認識率が改善されることがわかる。また、初期統合と結果統合の結果を比較すると、統合方法による有意な性能差は見られず、SNR=0dB の時、結果統合の方が若干良い結果となっ

ている。これは、初期統合の際、カメラのフレーム周期が粗く、統合の際に同じ画像を補間したため、画像のみの認識率が数%劣化したことが一つの原因と思われる。しかし、人間の情報処理においては同期情報を利用していると考えられるので、重み係数をSNRに依存させる[16]など重み付けや同期の方法などにさらに検討の余地があるものと考えられる。

3 音声から唇画像への変換

音声と画像の情報が強い相関をもつことを利用して、音声から唇画像を合成することを試みる。音声から、自然な動きの唇画像を合成することができれば、電話音声を唇画像に変換する、ラジオやテレビの放送を唇画像に変換するなどのさまざまな聴覚障害者の補助や、コンピュータエージェントの高度化、あるいは、発話画像伝送における唇画像の補間に役立てることができる。一般に、唇画像を合成する為には、音声から画像へ変換する方法と、テキストから画像を生成する方法があるが、著者らは、上述の用途を目指して、音声から画像へ変換するシステムについて検討している。この音声から唇画像への変換についても、いくつかの先行研究がなされてきている。例えば、[22, 23, 24]は、ベクトル量子化やニューラルネットワークを用いて、1フレーム毎に、音声パラメータから画像パラメータへと変換するシステムを提案している。ニューラルネットワークを用いることで改善されてはいるものの、フレーム毎に逐次変換する為、発話全体で見ると歪みが大きくなる問題がある。本章では、この問題に対し、サブワード単位のHMMを用いて、一発話全体で最適な変換を行うシステムについて述べ、実験により有効性を示す。

3.1 HMMを用いた合成アルゴリズム

まず学習方法について説明する。音声と画像を同期させたデータを用意する。音声のHMMを用い、1発話毎に全フレームにわたって、遷移確率と出力確率分布から尤度を計算し、最も尤度の高い遷移パスを選択する(アライメントをとる)。これにより、フレーム毎に対応する音素と対応する状態が決まる。学習で使用される全てのフレームのうち、同じ音素、同じ状態をとるフレームを選び、その画像パラメータの平均値をとる。これにより、図4の様に各音素の各状態毎に、画像パラメータの平均値が求まる[1]。次に合成では、テスト用の音声データを分析し、音声のHMMにより1発話毎にアライメントをとる。この時、音素毎状態毎に、学習した画像パラメータを対応させる。この画像パラメータを元に、1フレーム毎に3次元の唇画像を合成し、1発話毎に接続したものを画像系列として出力する。

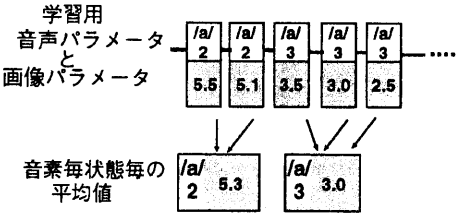

図4: 画像パラメータの学習

3.2 実験条件

合成のための画像データは、前章の認識用のデータベースより精密である必要があるため、3次元位置測定装置を用いて音声と画像のデータを収録する。音声と画像は、125Hzで同期がとられている。分析パラメータは、音声はメルケプストラム係数16次、差分係数16次、差分エネルギーが1次である。画像パラメータは、唇上に張り付けられたマーカーの位置座標から、口の縦の開き x、口の横の開き y、口端の奥行き z, の3変数に変換して使用する[25]。学習とテストに、1名の特定話者(女性)が発話した316単語を使用する。学習には音韻バランスのとれた216単語を用い、テストにはそれ以外の100単語を使用する。一方、アライメントをとるのに使用した音声HMMは次の条件で作成している。モデル数は、54音素に発話前pauseと発話後pauseを加えた56個。状態数は3状態。出力確率分布はTied-Mixture Gaussian型である。音声の周波数は12kHzのサンプルに、窓長32msecのハミング窓をかけ、分析周波数を125Hzとする。このモデルでの音素認識率は、認識率を(正答音素数−湧出し音素数)/全音素数で定義する場合、学習内単語で91.8%、テスト単語で69.4%の結果となる。

[1]このアルゴリズムは、画像を音声とHMMの状態の条件付確率として定義しEMアルゴリズムで画像の分布の平均値を推定する場合の近似となっている。

3.3 2乗誤差による評価

画像データを学習の結果は、合成画像パラメータ x_s, y_s, z_s と収録画像パラメータ x_o, y_o, z_o 間の2乗誤差 E で評価する。

$E = \sqrt{(x_s - x_o)^2 + (y_s - y_o)^2 + (z_s - z_o)^2}$ (cm)

表1に学習単語とテスト単語での1フレームあたりの平均2乗誤差を示す。正答時と音素認識の場合の

表 1: 誤差

	216 単語	100 単語
正答時 E	1.448	1.503
音素認識 E	1.473	1.530

誤差には大きな差がみられない。具体的な合成画像パラメータを図5-(a) に示す。横軸はフレーム番号、縦軸が E である。合成画像は実線、収録画像は破線 (全体が滑らか) で描かれている。縦線は、音声区間の区切りを示す。

3.4 後続音素依存の合成方法

HMM による合成方法で、誤差が大きく寄与する音素を調べたところ、/h/や促音/Q/, また発話前 pause が顕著であることが分かった。/h/の例を図5-(b) に示す。/h/と/Q/の特徴は、後続音素に依存した口形をとることである。そこで、HMM を用いる過程で、後続音素に依存した合成方法に拡張する。まず、画像パラメータの学習において、アライメントをとる時に、後続音素に注目する。そして、音素毎状態毎の画像パラメータを平均化する際に、後続音素別に平均値をとる。ただしこれでは、合成画像パラメータのパターン数が多くなりすぎ、学習できないパターンが出てくる。そこで、後続音素の第一状態の画像パラメータを、クラスタリングする。この結果出来るクラスを、ここでは viseme と呼ぶ。画像パラメータの平均値は、音素毎状態毎、更に後続音素の viseme 毎に用意することになる。合成時にも、アライメントをとる時には、後続音素の viseme に注目する。入力フレーム毎に、音素、状態、後続音素の viseme を見て、平均値画像パラメータを出力する。

後続音素依存の結果を表2に示す。全体として後続音素依存の方法により、誤差値が改善されている。また画像パラメータの変化を図5-(c) に示す。図では、/h/の部分の大きな誤差が低減されている。

表 2: 誤差 (後続音素依存)

	216 単語	100 単語
正答時 E	1.113	1.203
音素認識 E	1.155	1.277

4 まとめ

本稿では、HMM に基づく音声と唇画像の情報の統合による音声認識の高度化と音声から唇画像の生成を行なう方法について提案し、結果について述べた。前者については、SNR が劣化する場合に唇画像の利用により認識率の改善を行なえることを示した。しかしながら、初期統合と結果統合の差がないことなどから、統合方法にさらに検討の余地があることが明らかとなった。また、後者については、後続音素依存型の HMM 法を用いて一発話で最適化を行なう変換法が良好な変換結果を与えることが示された。今後、知覚実験を通した評価および評価尺度の検討などを行なっていく予定である。

5 謝辞

唇画像生成用のデータ、唇画像生成ソフトウェアの使用を許可して頂いた ATR 人間情報通信研究所の東倉社長、Bateson 博士、ICP の Benoit 博士に感謝致します。

参考文献

[1] 石川正俊、山崎弘郎、"センサフュージョンプロジェクト"、日本ロボット学会誌, Vol.12 No.5, pp.650-655, 1994

[2] 高橋弘太、"視覚と聴覚を統合するシステム"、電子情報通信学会誌 Vol.79 No.2,1996.2

[3] http://i60s30.ira.uka.de/miami/

[4] D.G.Stork, M.E.Hennecke, "Speechreading by Humans and Machines", NATO ASI Series, Springer, 1995

[5] E.Petajan, "Automatic Lipreading to Enhance Speech Recognition", Proc.CVPR'85

[6] E.Petajan, "Automatic Lipreading to Enhance speech Recognition", Proc.IEEE GLOBCOM,1984

[7] B.Yuhas, M.Goldstein. Jr,T.Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks", IEEE Communications Magazine, pp65-71, 1989

[8] J.Wu, S.Tamura, H.Mitsumoto, H.Kawai, K.Kurosu, K.Okazaki, "Neural Network Vowel-Recognition Jointly Using Voice Features and Mouth Shape Image", IEICE trans, D-II, Vol.J73-D-II,No.8,1990

[9] C.Bregler, H.Hild, S.Manke, A.Waibel, "Improving Connected Letter Recognition by Lipreading", Proc.IEEE ICSLP93

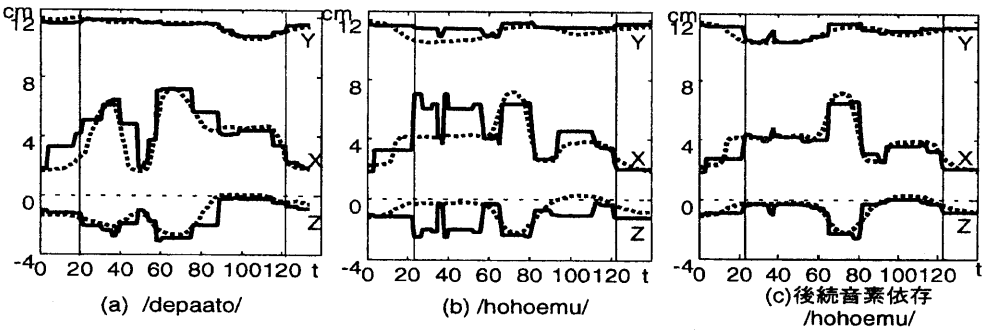

図 5: 唇パラメータの合成

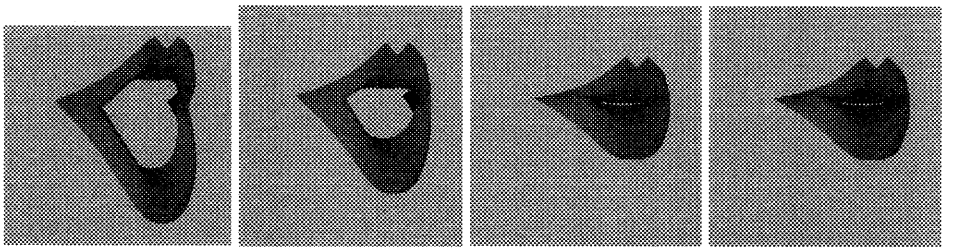

図 6: 生成画像 (/hohoemu/ 左から 45 フレーム目/h/ HMM 法、HMM-後続音素依存法、85 フレーム目/u/ HMM 法、HMM-後続音素依存法)

[10] A.Shintani, A.Ogihara, Y.Yamaguchi, Y.Hayashi, K.Fukunaga, "Speech Recognition Using HMM based on Fusion of Visual and Auditory Information", IEICE Trans,Fundamentals.Vol.E77-A,No.11 1994

[11] A.Adjoudani, C.Benoit, "Audio-Visual Speech Recognition Compared Across Two Architectures", Proc.EUROSPEECH95

[12] P.Silsbee, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition", IEEE Trans. on Speech and Audio, Vol.4. No.5,1996

[13] P.Cosi, E.Caldognett, F.Ferrero, M.Dugatto, K.Vagges, "Speaker Independent Bimodal Phonetic Recognition Experiments", Proc.ICSLP96

[14] P.Duchnowski, U.Meier, A.Waibel, "See Mee, Hear Me: Integrating Automatic Speech Recognition and Lip-Reading", Proc.ICSLP94

[15] P.Duchnowski, M.Hunke, D.Busching, U.Meier, A.Waibel, "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition", Proc.IEEE ICASSP95

[16] U.Meier, W.Hurst, P.Duchnowski, "Adaptive Bimodal Sensor Fusion for Automatic Speechreading", Proc.IEEE ICASSP 96

[17] R.Andre-Obrecht, B.Jacob, C.Senac, "Words on Lips:How to Merge Acoustic and Articulatory Informations to Automatic Speech Recognition", Proc.EUSIPCO '96

[18] M.Tomlinson, M.Russell, N.Brooke, "Integrating Audio and Visual Information to Provide Highly Robust Speech Recognition", Proc.IEEE ICASSP96

[19] P.Jourlin, "Handling Disynchronization Phenomena with HMM in Connected Speech", Proc.EUSIPCo '96

[20] M.Alissali, P.Deleglise, A.Rogozan, "Asynchronous Integration of Visual Information in an Automatic Speech RecognitionSystem", Proc.ICSLP96

[21] I.Matthews, J.Bangham, S.Cox, "Audiovisual Speech Recognition Using Multiscale Nonlinear Image Decomposition", Proc.ICSLP96

[22] S.Morishima, H.Harashima, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface", IEEE Journal on Selected Areas in Communications,Vol.9,No.4,1991

[23] F.Lavagetto, "Converting Speech into Lip Movements:A Multimedia Telephone for Hard of Hearing People", IEEE Trans. on Rehabilitation Engineering, Vol.3.No.1,1995

[24] S.Curinga, F.Lavagetto, F.Vignoli, "Lip Movements Synthesis Using Time Delay Neural Networks", Proc.EUSIPCO96

[25] T.Guiard-Marigny, A.Adjoudani, C.Benoit, "A 3D model of the Lips and of the Jaw for Visual Speech Synthesis", Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis, 1994.