# HMMを用いた読唇方法の検討

5H－7

永井 論　中村 哲　鹿野清宏*

奈良先端科学技術大学院大学 情報科学研究科

## 1　はじめに

音声は、人間同士の通信手段としてとても便利なものである。しかし、実環境ではさまざまな雑音が存在するため、音声だけでは認識が困難なことも多い。その際、人間は相手の口の形を参考にすることが知られている。そこで本稿では、人間が行なっている読唇（Lipreading）に着目し、HMM（隠れマルコフモデル）を用いて、唇画像の情報から発話内容を認識することを試みる。さらに、音声と画像の情報を統合した認識実験も行ない、唇情報による認識率の改善について述べる。

## 2　HMMを用いた読唇方法

### 2.1　方法

本稿では、HMMを画像認識に適用する。HMMを用いる従来の研究では、いくつかの問題点がある[1,2]。それは、口の特徴抽出の問題やデータベースの量によって学習が十分行なわれない問題、またGaussian-Mixtureを用いた時のモデルの学習の問題である。そこで本稿では、次のような方法を試みる。微妙な位置のずれを吸収するため、口の周辺画像に2次元FFTを行なう。重要語5240単語の特定話者のデータベースを収録する。Tied-Mixtureを使った学習を行ない、モデルの精度を改良する。以上の方法を適用することにより、問題点を改善する。

### 2.2　実験

データベースに、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ファイルを用いる。処理の手順は、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、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は次のように定義する。音素単位で55個の文脈独立モデルとし、状態数は2状態、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。学習する単語は、4740単語である。

### 2.3　結果

実験結果を表1,2に示す。表1はパラメータの数を変えた場合の500単語の認識実験の結果、表2は70次元における画像の認識結果である。()内は4混合のGaussian-Mixtureの認識率を表す。

表1：画像の500単語の認識率（単位%）

| | CLOSED | OPEN |
|---|---|---|
| 96次元 | 69.6(60.4) | 47.0(44.4) |
| 70次元 | 71.4(69.4) | 60.4(58.2) |
| 48次元 | 68.0(65.0) | 49.0(46.8) |

表2：70次元における画像の認識率（単位%）

| | CLOSED | OPEN |
|---|---|---|
| 100単語 | 87.0(83.0) | 85.0(84.0) |
| 500単語 | 71.4(69.4) | 60.4(58.2) |

表1の結果から、高周波成分を削除した70次元の認識率が最も良いことが分かった。またTied-Mixtureの認識率は、どの次元でもGaussian-Mixtureに比べて高く、結果からTied-Mixtureの有効性が明らかになった。

## 3　画像と音声の統合

### 3.1　統合方法

音声の情報に画像の情報を統合すれば、より良い認識が可能だろう。そこで実際に、以下の2通りの方法で実験を行ない、比較検討する。処理方法を図1に示す。

*Visual Speech Recognition based on Hidden Markov Models, Ron NAGAI, Satoshi NAKAMURA and Kiyohiro SHIKANO, Graduate School of Information Science,Nara Institute of Science and Technology;

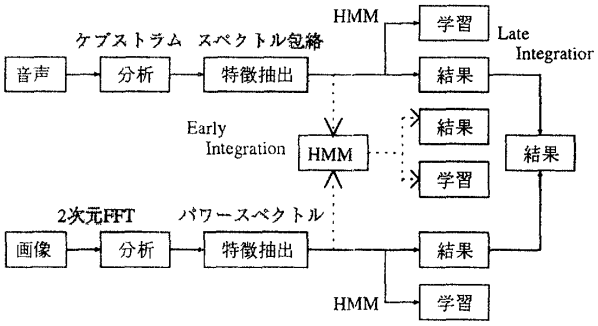


図 1: 処理方法

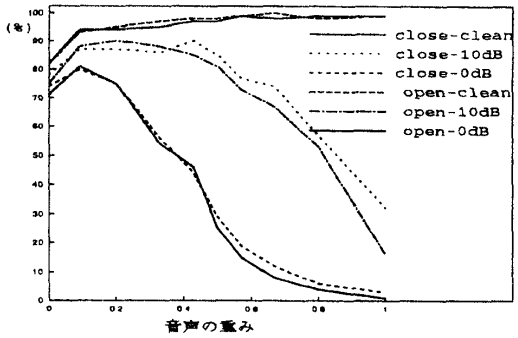


図 2: 初期統合 (EARLY INTEGRATION)

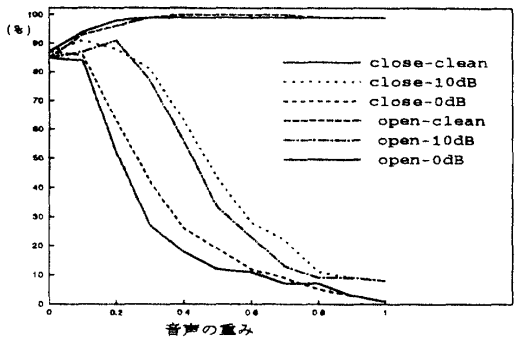


図 3: 結果統合 (LATE INTEGRATION)

1. 初期統合 (Early Integration)
   音声および画像の特徴を統合し、混合ベクトルを作成する。その際、音声と画像のフレーム周期が違うので、同じ画像を補間することが必要となる。このベクトルを音声 3 ストリーム、口の形 2 ストリームに分け、音声 (画像) の重みを変化させ実験を行なう。

2. 結果統合 (Late Integration)
   結果を統合する方法で、音声と画像それぞれの結果を出力する際に、すべての単語に対する尤度を求める。その音声の結果 (対数尤度) と画像の結果 (対数尤度) を重みづけ計算し、最終的な結果とする。($S_{total} = \lambda_a S_{audio} + (1-\lambda_a) S_{visual}$ : $\lambda_a$は音声の重み)

### 3.2 実験内容

音声は 22.050$kHz$を 12$kHz$にサンプリングする。前処理にメルケプストラムを用い、16 次で分析した後、33 次元 ($16MFCC + 16dMFCC + dPower$) の特徴パラメータを求める。HMM モデルは、文脈独立で 55 音素・3 状態、分布は Tied-Mixture で行なう。画像は 2.2 と同様。学習および認識単語は、それぞれ、4740 単語、100 単語で行なう。

### 3.3 実験結果および検討

2 つの統合方法による認識結果を図 2,3 に示す。初期統合と結果統合の結果を比較すると、どちらも統合することにより認識率が改善された。しかし目立った認識率の差はなく、SNR が 0dB の時に、結果統合の方が若干良い結果となった。これは初期統合の際に、画像のフレーム周期が粗い (30 フレーム/s) ため、画像を補間して認識しており、このことが認識率の劣化の一つの原因と考えられる。

## 4 おわりに

本稿では、画像の認識率と、種々の SNR の下で音声と画像を統合させた時の認識率を実験によって求めた。またその統合方法についても検討し、音声と画像を統合する有効性を示した。本稿では、特定話者のみの実験しか行なわなかったが、今後は不特定話者の唇画像認識や統合方法の改善を行なう予定である。

## 参考文献

[1] Mamoun Alissali, Paul Deleglise and Alexandrina Rogozan. Asynchronous Integration of Visual Information in an Automatic Speech Recognition System. *Proc.ICSLP*,P34-37,1996.

[2] Qin Su and Peter L.Silsbee. Robust Audiovisual Integration useing Semicontinuous Hidden Markov Models. *Proc.ICSLP*,P42-45,1996.

[3] C.Bregler,H.Hild,S.Manke,and A.Waibel. Improving connected letter recognition by lipreading. *Proc.IEEE Int. Conf. on Acoustic,Speech and Signal Processing, Minneapolis*, 1:557-560,1993.

[4] P.Duchnowski,M.Hunke,D.Busching,U.Meier,and A.Waibel. Toward movement-invariant automatic lipreading and speech recognition. *Proc.ICASSP*, P109-112,1995.