モバイル環境雑音に強い音声認識技術

モバイル環境では，雑音の影響を大きく受け，音声認識の性能の劣化が大きい．そこでモバイル環境雑音に強い音声認識の向上を目的として「マルチモーダル音声認識の高度化に関する研究」と「雑音処理技術の研究」を行った．なお，本研究は東京工業大学大学院 情報理工学研究科 古井研究室（古井 貞熙教授）との共同研究により実施した．

張 志鵬（ちょう しほう）　大矢 智之（おおや ともゆき）　杉村 利明（すぎむら としあき）

1. まえがき

モバイル環境では，音声による入力・操作は非常に簡易かつ有効である．しかし，音声認識性能はモバイル環境の雑音の影響を大きく受け，性能の劣化が生じるために性能向上が強く求められている．そこで，これらの課題を解決するために重要な「マルチモーダル音声認識の高度化に関する研究」と「雑音処理技術の研究」の2方向から研究を進めた．

(1) マルチモーダル音声認識の高度化に関する研究

①横顔の動画像を用いたマルチモーダル音声認識

モバイル環境において雑音に頑健な音声認識手法として，横顔口唇動画像情報を用いたマルチモーダル音声認識手法を提案した．これは，横顔の画像情報を利用することにより，ユーザに自然な姿勢での音声入力を提供することができるという手法である．**図1**に，本研究のマルチモーダル音声認識の構成図を示す．本手法では，音響と画像情報はマルチストリーム隠れマルコフモデル（HMM：Hidden Markov Model）を用いて融合されることで，認識性能の向上が達成された．

②マルチモーダル音声認識におけるストリームの重みの最適化手法の検討

上記のマルチストリームHMMを用いた音響と画像情報のマルチモーダル音声認識において，尤度平均化基準による最適化手法の提案を行い，適用化データが少量のとき従来の手法である尤度比最大法を用いた場合に比べ，誤り率を約40％削減できることを確認した．

(2) 音声認識の雑音処理技術の研究

音声認識実験の多くは，入力音声は発声区間が未知の連続入力ストリームである．発声区間が未知の音声，特に雑音条件下の連続入力音声を自動的に認識する技術が必要である．そこで，雑音特性と信号対雑音比（SNR：Signal to Noise Ratio）が時間的に変化するような状況において，発声区間を自動的かつ頑健に検出し，さらに木構造雑音重畳音声モデルによる手法を提案し，認識性能が上がることを確認した．

2. 横顔の動画像を用いたマルチモーダル音声認識

音響雑音の影響を受けない音声認識手法の1つとして，発声時の口唇の動画像から得られる情報を音声情報とともに利用するマルチモーダル音声認識システムが注目され，近年研究が進められている[1]．従来，これらの研究では，正面から撮影された口唇画像が用いられている．しかし，モバイル環境でこのような方式を利用しようとすると，さまざまな問題が生じる．カメラ付き移動端末で音声と画像を入力することを考えると，ユーザは話しながら移動端末を顔の正面に持ってカメラ撮影を行うことになり，負担が

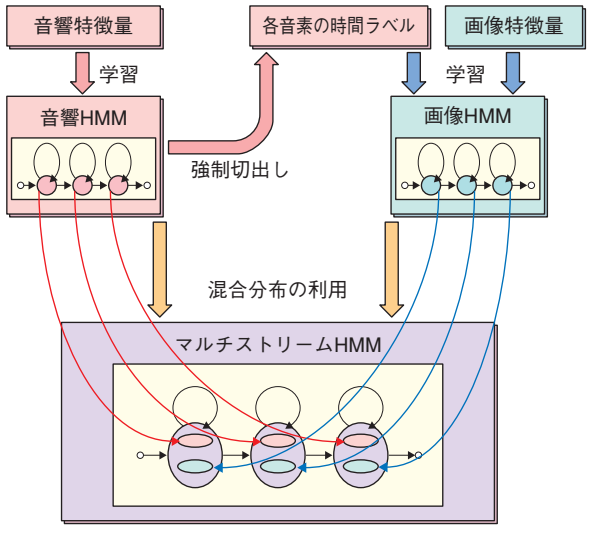

図1　マルチモーダル音声認識の構成図

大きく，自然な音声入力を行うことができない．また，撮影のために移動端末と口との距離を離さなければならず，音声のSNRが劣化するといった問題も生じる．

そこで本研究では，正面の顔画像ではなく，横顔の動画像情報を用いたマルチモーダル音声認識手法を提案する．この方法では，移動端末などのモバイル機器のマイクロフォン部分に微小カメラを搭載し，口唇動画像情報を取得することを想定している．横方向からの口唇の動き情報を音声認識に利用することによって，モバイル環境において自然な体勢で音声入力が可能で，雑音に頑健な音声認識を行うことができる．さらに，本研究では新たに横向きの口唇の形状モデルを用いた口唇ライン抽出による特徴量抽出法も提案した．以下では，口唇ライン抽出について述べ，マルチモーダル音声認識手法，評価実験条件とその結果について述べる．

2.1 口唇ライン抽出

ここでいう口唇ラインとは，「横顔における口唇の最左点を基準点として，この点から上唇，下唇それぞれの口唇領域を最も含むように引かれた2つの直線」である（**図2**）．なお，本研究では，横方向からの口唇画像として話者の右方向から撮影した画像を用いており，以下の3つのステップにより抽出される[2]．

(1) 口唇領域抽出

初期フレーム画像の左右端の決定，上下端の決定，領域抽出のための画像生成，色相画像の口唇以外の点の除去と左右端の決定，上下端の決定，という手順で行う．

(2) 基準点抽出

(1)で得られた口唇領域の画像から，口唇ラインの基準点を決定する．まず，原画像から口唇の内部領域を抽出する．口唇内部は画像中において最も暗くなることから，輝度の値で2値化を行うことで抽出を行う．口唇内の白色部分が得られた領域である．

(3) 上下唇ラインの決定

以上のように求められた口唇領域と基準点から，色相画像に対して探索開始線から直線を右回りに回転させながら，直線上の画素数を求めていき，画素数が最大であった直線を下唇ラインとし，上下唇のラインを決定する．

2.2 横顔動画像を用いた音声認識実験

(1) 特徴量

音声データをPCに取り込む際に，標本化周波数を16kHzにダウンサンプリングした．そして，フレーム長25ms，フレームレート100Hzで発声区間を切り出してメル ケプストラム係数（MFCC：Mel-Frequency Cepstrum Coefficient）と正規化対数パワーを抽出した．映像はデジタルビデオカメラにより収録した幅720×高さ480pixelの24bitフルカラーの画像である．まず180×120pixelに解像度縮小を行う．この画像に対し，口唇ラインの抽出を行う．この口唇ラインの成す角度と，その微分成分の計2次元をフレーム画像ごとに計算し，2次元の画像特徴量を得る．

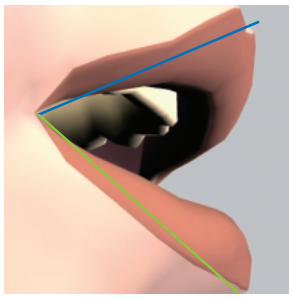

図2　口唇ラインの例

本手法では，音響特徴量と画像特徴量をパラメータの段階で融合して音響・画像特徴量を生成し，認識に利用する．ただし，音響特徴量のフレームレートが100Hzであるのに対して画像特徴量は30Hzであるため，そのままでは特徴量を融合できない．そこで，画像特徴量を3次元スプライン関数によって時間方向に補間を行い，音響特徴量と同じ100Hzにフレームレートを修正したうえでフレームごとに融合を行った．

(2) マルチストリームHMMの構築

本研究では，まず音響HMMと画像HMMをそれぞれ別々に学習し，それらを融合することで構築される．

(3) 実験結果

学習には音響・画像ともにクリーン環境で収録した男性話者11名による連続数字読上げデータを使用した．テストには高速道路走行中の車内で収録した，各話者の2～6桁の数字の発声データを使用した．テストデータ中の音響雑音としてはエンジン音，風切り音やウィンカー音などが観測され，SNRはおよそ10～15dBであった．

上下唇のライン特徴量を用いたときの結果を見ると，音響のみの結果より絶対値で8.0％向上され，提案した画像特徴量の有効性が確認できた．

3. マルチモーダル音声認識におけるストリームの重み最適化手法

マルチストリームHMMのストリームの重みに関しては，よりよい認識性能を得るために，音響・画像それぞれの雑音状況に応じて適切に重み係数を設定する必要がある．そこで，尤度比に基づき適応的に重み係数を推定する，尤度平均化基準によるストリームの重み最適化手法を提案する．

3.1 ストリームの重み最適化手法

(1) 音響・画像特徴量を用いたマルチストリームHMM

本研究では音声認識のためのモデルとして，音響ストリームと画像ストリームより成るマルチストリームHMMを用いている．マルチストリームHMMでは，単語wに対する音響・画像特徴量O_tの対数尤度$b_w(O_t)$は，式(1)のように定義される．

$$b_w(O_t)=b_{Aw}(O_{At})^{\lambda_{Aw}}\times b_{Vw}(O_{Vt})^{\lambda_{Vw}} \quad (1)$$

ただし，tは時刻，$b_{Aw}(O_{At})$は音響特徴量O_{At}，$b_{Vw}(O_{Vt})$は画像特徴量O_{Vt}に対する単語wの対数尤度，λ_{Aw}とλ_{Vw}は単語wのHMMにおける音響／画像ストリームの重みで，以下の制約がある．

$$\lambda_{Aw}+\lambda_{Vw}=1,\ 0\le\lambda_{Aw},\ \lambda_{Vw}\le 1 \quad (2)$$

(2) 尤度平均化基準による最適化手法

雑音環境下では，モデル学習時の環境と認識時の環境の違いにより，各単語のモデルが出力する尤度に偏りやばらつきが発生するため，特定の単語に関連する認識誤りが生じやすくなる．そこで，各モデルが出力する尤度の平均を一定以上の時間でみて等しくなるようにストリームの重み係数を調整し，出力尤度の偏りやばらつきを抑制することで，認識性能を改善する手法を考える．具体的には，式(3)により単語$d\in W$に対する音響ストリームの重みを推定する[3]．

$$\lambda_{Ad}=\frac{\frac{1}{NT}\sum_{t=1}^{T}\sum_{w\in W}b_{Aw}(O_{At})}{\frac{1}{T}\sum_{t=1}^{T}b_{Ad}(O_{At})} \quad (3)$$

この尤度平均化基準による手法には繰返し演算が不要で，計算量・演算時間が削減できるという利点がある．さらに，雑音状況が大きく変わらなければ，比較的少量のデータでも信頼性の高い対数尤度平均値を得ることができると考えられる．

3.2 認識実験

(1) 実験条件

ストリームの重み最適化は教師なし条件で行った．評価用データベースを話者ごとにデータを36個のデータセットに分割した．それぞれのデータセットでは，始めのn個の発声と対応する認識仮説を用いてストリームの重みを最適化し，得られたストリームの重みを用いて，セット内の発話を音声認識した．

(2) 実験結果

尤度平均化法の場合では，少量の最適化データでも認識率が改善し，さらにデータ量が増えるほど認識性能も大きく向上し，従来法を上回る優れた性能を示すことが確認された．従来の尤度比最大法を用いた場合に比べ，誤り率を約40％削減できることを確認した．

4. 音声認識の雑音処理技術

発声区間が未知の音声，特に雑音条件下の連続入力音声を自動的に認識する技術が必要である．雑音特性とSNRが時間的に変化するような状況において，発声区間を自動的かつ頑健に検出し，さらに音声認識性能を向上させるために，音素モデルを準リアルタイムで雑音に適応化する方法について検討する．

4.1 木構造雑音重畳音声モデルの作成

多様な種類の雑音とSNR条件下の雑音重畳音声モデルを学習し，雑音重畳音声のクラスタリングを行って，1つの木構造を作成する．この木構造をルートからリーフ方向にたどり，最適なモデルを選択することにより，入力音声に最適な雑音区分空間を選択できる[4]．**図3**に木構造モデルの概念を示す．木構造で雑音特性を表すことにより，木構造の上層では雑音特性の大局的な特徴，下層では特定の雑音特徴を表現するモデルが得られる．

4.2 連続音声認識

連続入力音声には，明示的に文の区切り情報が与えられないため，まず，固定長の音声（ブロック）を抽出し，ブロックごとに処理する．各入力ブロックに対して木構造空間を上から下にたどり，最適なモデルを選択することによって入力音声に最適なHMMを選択し，認識を行う．認識結果に基づいて以下のように発声区間の検出を行う．認識結果の中に区切りがある場合は，最初の区切りの出現場所を発声区間の始めとする．認識結果の中に区切りがない場合は，この音声ブロックの終点を発声区間の終りとする．さらに選ばれたモデルに対しMLLR（Maximum Likelihood Linear Regression）*[5]によるモデル適応を行って，最終結果を出力する．

4.3 認識実験

(1) 実験条件

音声認識実験のタスクとしては，音声入力による飲食

＊ MLLR：尤度最大化に基づくモデルのパラメータの線形変換手法．

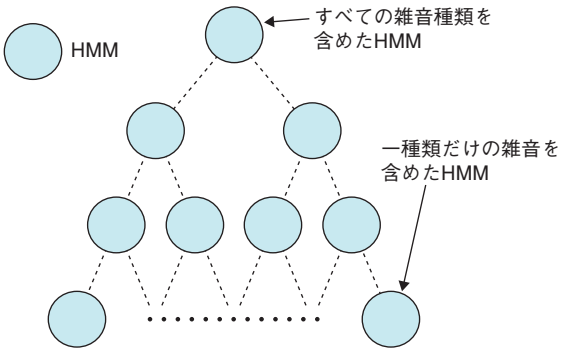

図3　木構造モデルの概念図

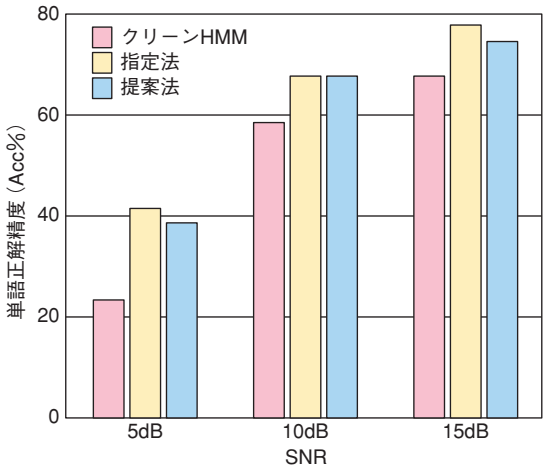

図4　3種類の条件下の認識の精度（Acc％）

店舗検索の対話システムを用いた．音響モデルは2,000状態16混合の状態共有型HMMである．評価用データとしては，10名の話者が発声した計50の対話音声に対し，3種類のSNR（SNR＝5, 10, 15dB）で，学習に用いなかった駅の雑音を重畳させたデータを用意した．

(2) 実験結果

音声実験は，「クリーンな音声で学習したHMMを用いた認識（クリーンHMM法）」「文の正しい区切り情報を指定し，認識を行う（指定法）」「提案法による発声区間検出と認識（提案法）」という条件下で行った．**図4**に，この3種類の条件下における認識の性能（Acc％）を示す．提案手法は，いずれの雑音の場合も，クリーンHMMに比べて大幅に良い性能を示している．また，目標値の区切り指定の場合に近い性能が得られることが分かる．

5. あとがき

本研究では，マルチモーダル音声認識の高度化に関する研究として「横顔の動画像を用いたマルチモーダル音声認識」および「尤度平均化基準による重みの最適化手法」を提案し，データが少量のときに特に有効に機能することが確認された．今後の課題としては，大語彙連続音声認識などへのマルチモーダル音声認識の適用，リアルタイム・マルチモーダル音声認識システムの構築などが挙げられる．

また，音声認識の雑音処理技術の研究として「頑健な区間検出とモデル適応に基づく雑音下音声認識手法」を提案し，短い遅延で高い認識性能が確認された．今後の課題としては，実時間化に向けた改善などが挙げられる．

文　献

[1] C. Bregler and Y. Konig: “‘Eigenlips’ for robust speech recognition,” Proc. ICASSP 94, Vol. 2, pp. 669-672, 1994.

[2] T. Yoshinaga, S. Tamura, K. Iwano and S. Furui: “Audio-visual speech recognition using lip movement extracted from side-face images,” Proc. AVSP 2003, pp. 117-120, 2003.

[3] S. Tamura, K. Iwano and S. Furui: “A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization,” Proc. ICASSP 2005, pp. 469-472, 2005.

[4] Z. P. Zhang and S. Furui: “Noisy speech recognition based on robust end-point detection and model adaptation,” Proc. ICASSP, pp. 981-984, 2005.

[5] C. J. Leggetter and P. C. Woodland: “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, Vol. 9, No. 2, pp. 171-185, 1995.

用　語　一　覧

HMM：Hidden Markov Model（隠れマルコフモデル）
MFCC：Mel-Frequency Cepstrum Coefficient（メルケプストラム係数）
MLLR：Maximum Likelihood Linear Regression
SNR：Signal to Noise Ratio（信号対雑音比）