E-054

画像特徴量誤りによる視聴覚融合音声認識の認識率低下の抑制法

The Recognition Error Reduction Method for Audio-Visual Automatic Speech Recognition against Visual Feature Errors

吉田孝博　　半谷精一郎
Takahiro Yoshida　　Seiichiro Hangai

東京理科大学 工学部 電気工学科
Dept. of Electrical Engineering, Tokyo University of Science

1．まえがき

近年，音声認識の耐雑音手法の一つとして，音声とは別のモダリティを認識に用いる視聴覚融合音声認識が研究されている．この視聴覚融合音声認識では，音声情報は雑音の影響を受けて信頼度が低下するが，画像情報（口唇動作や口唇画像）は音響的雑音の影響を受けないとしている．実際に，視聴覚融合音声認識の研究では，画像情報は常に雑音の影響がなく，誤りが無いという前提のもと，両情報の統合方法の検討[1]や音声認識に適した画像特徴量の検討[1][2]が行われてきた．

しかし，視聴覚融合音声認識を実際に使用する際には，話者の顔の向きや位置，照明条件の変動等で画像特徴量にも欠落や誤りが生じる．また，画像特徴量が得られる状況と得られない状況が時間と共に刻々と変化する場合もありうる．そのため，実環境にロバストな視聴覚融合音声認識の実現のためには，認識部においても画像情報の誤りに対する適応的な耐雑音手法が必要である．

そこで本研究では，画像特徴量の誤りの一形態である一部欠落時の認識率への影響を認識実験により調査すると共に，誤りフレームが特定された場合に認識精度低下を回避する手法を提案し，効果を確認した．

2．画像特徴量誤りによる認識精度低下抑制法

画像特徴量誤りによる認識精度低下を抑制するため，認識時の出力確率計算において，画像特徴量の抽出誤りと判定されたフレームの出力確率を 0 とすることにより，誤りによる悪影響を低減する．この提案手法を組み入れた初期統合型のマルチストリーム HMM を用いた視聴覚融合音声認識系を図 1 に示す．

まず，顔動画像から口唇形状や口唇領域画像等の画像特徴量を抽出する処理にて，抽出不可（欠落・誤り）と判定したフレームの抽出成功フラグ C(t)に 0 を，抽出成功フレームでは C(t)に 1 を出力させる．次に，認識時の画像ストリームの HMM の出力確率計算にて，C(t)=0 であるフレームの出力確率を 0 とする．この処理は，状態 j，フレーム t における視聴覚融合音声認識系の HMM の出力確率計算式に C(t)を組み入れた次式により行う．

$$\log b_j(o_t) = W_{Sp} \cdot \log b_j(o_{t\text{-}Sp}) + C(t) \cdot W_{Lip} \cdot \log b_j(o_{t\text{-}Lip}) \quad (1)$$

ここで，W_{Sp} は音声ストリームの重み，W_{Lip} は画像ストリームの重みである．

なお，抽出成功フラグ C(t)を得る方法は，画像特徴量の種類や口唇形状抽出アルゴリズムにより異なる．本研究では，L*a*b*色空間の a*により赤みを帯びた口唇周辺を切り出し，テンプレートマッチングにより口唇領域を決定し，投影法により口唇領域の縦幅・横幅を測定する口唇形状抽出方法[3]を用いているので，口唇領域のピクセル数や抽出した縦幅・横幅の閾値，前後フレームからの変化量の閾値等にて判定する方法が必要である．

3. 単語認識実験

3.1　実験方法

画像特徴量の一部欠落時の認識率への影響と，提案手法の効果を検討するため，認識実験用に画像特徴量を欠落させたテストセットを用意した．欠落方法は，1 単語内の全フレームに対して規定の比率のフレームを単語中央部から欠落させる．欠落部の補正処理として，欠落直前の値を用いるサンプルホールドと，欠落部の直前と直後の値から線形補間を行う場合との 2 通りを用意した．図 2 に，30%欠落時と 70%欠落時の線形補間した画像特徴量の例を示す．なお，今回の実験で用いた抽出成功フラグ C(t)は，実験的に既知として用意したため，現段階では検出処理は行っていない．

これらの実験的に一部欠落させたテストセットにて行った 50 単語認識の実験条件を表 1 に示す．今回のテストセットの元となる顔画像音声データベースは，当研究室で用意したものであり，語彙はカーナビゲーションでの使用を想定した 50 単語である．また，HMM による認識系は

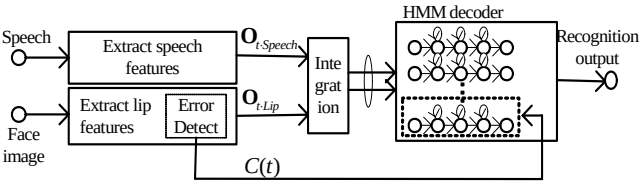

図 1　提案手法を組み入れた視聴覚融合音声認識系

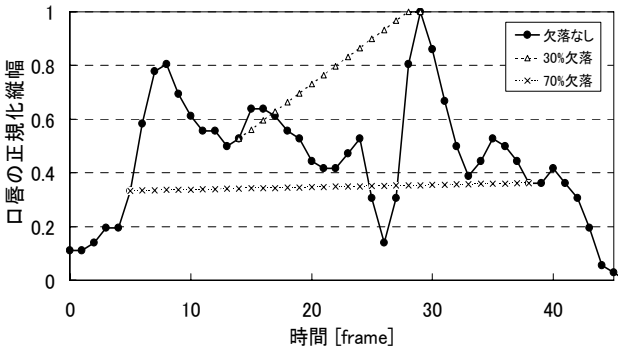

図 2　単語中央部から欠落させ，線形補間した画像特徴量の例（発声単語：「現在地」）

表 1　単語認識実験の実験条件

顔画像音声データベース（50 単語）	男性話者 20 名，各話者各 4 回発声 学習用…2 回発声分(2000 データ) 認識用…残りの 2 回分(2000 データ)
音声特徴量	蝸牛フィルタ出力 20 次:Δf (86 次)の DCT 圧縮とΔt (86 次)の DCT 圧縮 分析窓長：30ms，分析周期：11.0625ms
両特徴量のフレーム周期	11.0625ms(90fps)（画像特徴量:NTSC30fps を線形補間により 90fps 化）
画像特徴量（口唇形状）	計 4 次：口唇の横幅 $x(t)$，縦横比 $r(t)$，横幅の時間差分 $\Delta x(t)$，縦幅の時間差分 $\Delta y(t)$
音響雑音	HMM 学習時：無雑音&白色雑音(SNR=0dB) 認識時：白色雑音(Clean, SNR=10,5,0,-5,-10dB)
HMM	3 状態 Left-to-Right 型，計 33 音素モデル，混合数(ガウス分布):音声 6，画像 2

HTK (HMM Took Kit) Ver.3.2 により構築した．HMM の学習は，各話者 2 回発声分（計 2000 発話）の，音響雑音が無いクリーンな音声と，SNR=0dB の白色雑音を加えた音声と，欠落の無い画像特徴量により行った．一方，認識時は，学習時に用いなかった残りの各話者 2 回発声分の音声（計 2000 発話）に対して，白色雑音を SNR=Clean，10dB，5dB，0dB，-5dB，-10dB となるように加えた音声と，実験的に一部欠落させた画像特徴量により行った．

また，本実験の視聴覚融合認識系の音響 HMM と画像 HMM のストリーム重み（融合比）は，各 SNR で最適値を用いた．

3.2　実験結果

画像特徴量のみを用いて行った 50 単語認識実験の結果を図 3 に，視聴覚融合単語音声認識での結果を図 4 に示す．なお，図 3 の画像特徴量のみの認識実験結果では，1 発話毎の画像特徴量の全フレームに対する欠落フレームの比率を 0～80%の範囲で 10%刻みで変化させて単語認識率を示してある．また，図 4 の視聴覚融合単語音声認識では，横軸を音声の SNR とし，画像特徴量の欠落 0%（欠落なし），30%，70%の結果について示した．

図 3 より，画像特徴量のみを用いた認識において，欠落部の補正を線形補間等で行っても，画像特徴量の誤りが増すにつれて認識率が低下し，欠落 50%で認識率が半減した．一方，提案手法を用いることにより，画像特徴量の誤りの増加に伴う認識率低下が他の補正処理よりも抑えられている．欠落 50%時において線形補間処理と比較すると，12.5%の改善が得られた．

図 4 より，視聴覚融合音声認識においても同様に，提案手法を用いることにより大幅に認識率低下を抑制できている．特に，視聴覚融合時に画像特徴量の重みが大きくなる高雑音環境下で効果が大きく，音声の SNR が 0dB の環境において線形補間処理と比較すると，欠落 30%の場合では 8.8%の認識率改善，欠落 70%では 48.2%の改善が得られた．

認識時に，誤った画像特徴量の値を持つフレームにおいて計算される出力確率値が，本来正解である単語の出力確率よりも他の不正解単語の方が大きい値となり，誤認識を引き起こす要因となると考えると，本提案手法は，その誤

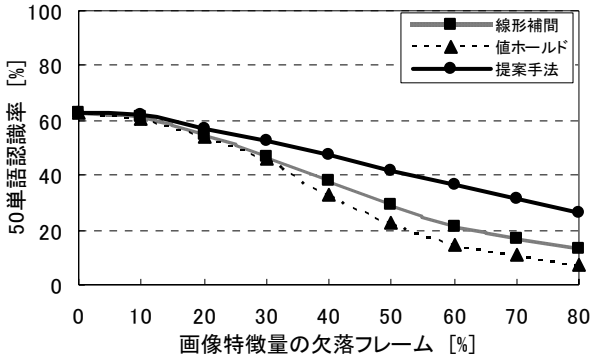

図 3　画像特徴量欠落時の単語認識率
（画像特徴量のみでの認識時）

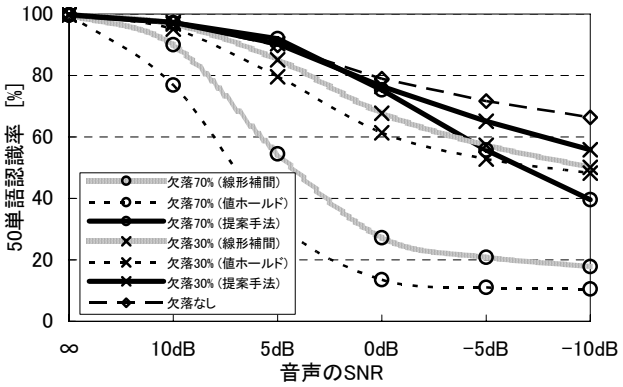

図 4　画像特徴量欠落時の単語認識率
（視聴覚融合認識時）

ったフレームにおいて計算される出力確率を無視するため，誤認識が抑制できるといえる．さらに，出力確率を 0 とするフレームが増すことにより，相対的に画像ストリームの比率が下がる点も，有効に作用していると考える．

4. まとめ

視聴覚融合音声認識において，画像特徴量の誤りフレームの増加に伴い，認識率低下が生じることを確認した．また，認識時の出力確率計算において，誤りフレームにて計算される画像ストリームの出力確率を 0 とする提案手法により，視聴覚融合単語音声認識では，SNR=0dB において 48.2%の認識率改善が得られ，画像特徴量の誤りによる認識率低下を抑制できることが示された．

参考文献

[1] G. Potamianos, et al., “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, Proc. of IEEE, vol.91, no.9, Sep. 2003

[2] Q. Zhi, et al., “HMM Modeling for Audio-Visual Speech Recognition”, Proc. of IEEE ICME’01, pp. 136-139, 2001

[3] T. Yoshida, T. Hamamoto, S. Hangai, "A Study on Multi-modal Word Recognition System for Car Navigation", Proc. of URSI ISSSE'01, pp.452-455, 2001