論文 / 著書情報
Article / Book Information

論題(和文)	対話音声を対象とした不特定話者マルチモーダル音声認識の検討
Title(English)	
著者(和文)	高山 俊輔, 松尾 俊秀, 岩野 公司, 古井 貞熙
Authors(English)	Toshihide Matsuo, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2007年春季講演論文集, Vol. , No. 2-9-13, pp. 61-62
Citation(English)	, Vol. , No. 2-9-13, pp. 61-62
発行日 / Pub. date	2007, 3

対話音声を対象とした不特定話者マルチモーダル音声認識の検討*

◎高山俊輔, 松尾俊秀, 岩野公司, 古井貞熙 (東工大)

1 はじめに

音声認識の耐雑音性を向上させるため，口唇の動画像情報を併せて利用するマルチモーダル音声認識の研究が盛んに行われている．国外ではこれまでに，ニュース音声などの大語彙連続音声認識に関する研究が行われているが [1]，日本語を対象とした研究は，タスクが連続数字などで語彙数が小さいものがほとんどである [2]．近年になり文献 [3] で大語彙を対象とした研究が行われるようになったが，話者が 1 名と限定的で，不特定話者を対象とした本格的な研究はまだ行われていない．そこで本論文では，日本語の対話音声をタスクとした，不特定話者マルチモーダル音声認識についての検討を行う．学習と評価に用いる音声・画像データを収集し，それを用いて大語彙タスクに適した音響・画像情報の融合方法について検討する．

2 データベースの構築

録音室において，男性話者 25 名からなるマルチモーダルデータベースの収録を行った．詳細を Table 1 に示す．話者は壁に背を向けて椅子に座り，そこから 1m20cm 離れたディスプレイに表示された文章を発声する．画像はディスプレイ上に設置したカメラで顔全体を撮影し，音声はディスプレイ手前に設置した無指向性マイクで収音した．ここで，評価データのタスクである模擬対話文は，先行研究 [4] で構築した音声対話システムの入力となるような，「柏で定食ありますか？」などの文章となっている．

3 マルチモーダル音声認識システム

Fig.1 に，本研究で構築したマルチモーダル音声認識システムの流れを示す．音声・画像はそれぞれ標本化周波数 16kHz，60Hz でサンプリングを行い，それぞれ 100Hz の音響特徴量と 60Hz の画像特徴量に変換する．その後フレームレートを合わせて 2 つの特徴量を融合し，マルチストリーム HMM により認識を行う．音響特徴量には，CMN-MFCC12 次元とその Δ，$\Delta\Delta$ 成分，および対数パワーの Δ，$\Delta\Delta$ 成分の計 38 次元を用いた．画像特徴量にはオプティカルフローに基づく特徴量 [2] と主成分分析 (PCA) に基づく特徴量の 2 種類を用いた．

3.1 画像特徴量の抽出

画像特徴量の抽出は OpenCV[5] を用いて行った．始めに口唇位置の検出を行い，切り出した口唇画像に対して画像特徴量を抽出する．切り出す口唇画像のサイズは，オプティカルフロー特徴量では 180×120，PCA 特徴量では 110×90 とした．

オプティカルフローは，「画像中の明度パターンの見かけ上の速度分布」と定義される．フローベクトルは，連続する 2 フレームの口唇画像を用いて Lucas-Kanade 法 [6] により計算される．得られたフローベクトルの水平・垂直成分の分散値を計算し，この 2 次元を画像特徴量として用いる [2]．

Table 1 データベースの詳細

学習データ	タスク	ATR 音素バランス文
	話者	男性話者 15 名
	発声数	計 1509 文
	総時間長	約 2 時間半
評価データ	タスク	模擬対話文
	話者	男性話者 10 名
	発声数	計 400 文
	総時間長	約 30 分

PCA に基づく画像特徴量抽出のための固有空間の構築には，学習セットである 15 話者の発声画像のうち，ランダムに選ばれた約 2700 枚から切り出した口唇画像を用いる．入力口唇画像をこの空間に射影し，得られた主成分得点を画像特徴量として用いる．次元数は 5 とした．

3.2 マルチストリーム HMM

マルチストリーム HMM では，音響・画像特徴量 O_t の観測確率は，対数尤度 $b(\mathrm{O}_t)$ を用いて以下のように表される．

$$b(\mathrm{O}_t) = \lambda_A b_A(\mathrm{O}_{At}) + \lambda_V b_V(\mathrm{O}_{Vt}) \qquad (1)$$

ただし，t は時刻，$b_A(\mathrm{O}_{At})$，$b_V(\mathrm{O}_{Vt})$ はそれぞれ音響特徴量 O_{At}，画像特徴量 O_{Vt} に対する対数尤度，λ_A，λ_V は音響，画像ストリーム重みで，認識時には以下の条件で変化させる．

$$\lambda_A + \lambda_V = 1 \ , \ 0 \leq \lambda_A, \lambda_V \leq 1 \qquad (2)$$

4 音響・画像情報の融合

4.1 従来の融合方法

マルチストリーム HMM により音響・画像情報の融合を行う従来手法の 1 つとして，フレームごとに融合された音響-画像特徴量を 1 つのベクトルとして扱い，通常の HMM と同じように学習を行った後，ストリームの分割を行うことでマルチストリーム HMM に変換する方法がある．この手法は大語彙タスクを対象とした研究 [3] などで用いられている．しかし，画像情報には唇の動きといった調音様式の一部のみしか反映されないため，音響情報に比べて音素識別性能が低く，2 つの情報を単一のベクトルとして扱うこの手法では，画像特徴量の次元数が増えるにつれて学習時の音素境界の推定精度が劣化し，音響パラメータや状態遷移行列の学習に悪影響を及ぼす可能性がある．

この問題を解決する融合方法として，小語彙タスクを対象として行った先行研究 [7] では，音響 HMM と画像 HMM を別々に学習し，それを融合する手法について検討している．ただし，画像情報単体では音素境界の推定精度が低いため，画像 HMM 学習の際には，事前に学習した音響 HMM により切り出した音素境界の時間情報を与えて学習し，モデル単位で

* Speaker independent multimodal speech recognition for spoken dialog systems. by Shunsuke Takayama, Toshihide Matsuo, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology).

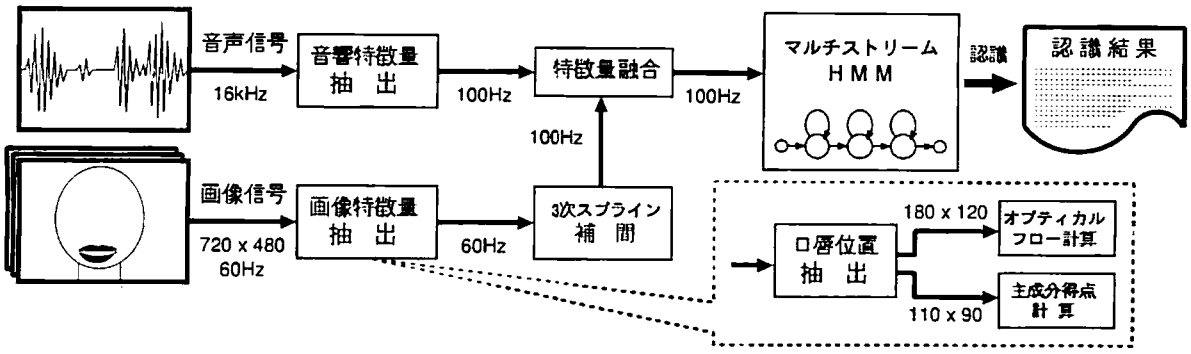

Fig. 1 マルチモーダル音声認識システム

Table 2 各融合方法の認識結果 (単語正解精度)

SNR	音響のみ	従来手法	提案手法
10dB	7.1 %	11.6 %	13.9 %
15dB	24.7 %	31.9 %	35.2 %
20dB	47.2 %	53.8 %	53.9 %
clean	74.3 %	75.4 %	75.6 %

※オプティカルフローに基づく画像特徴量を使用

融合している．しかし，本研究のように大語彙をタスクとした場合には，学習データ不足を補うためトライフォン間で状態の共有化を行うため，共有化構造の異なる音響，画像 HMM を音素境界の同期をとりながら学習し，両者をモデル単位で融合することは困難である．

4.2 本研究における融合方法

本研究では，[7] の手法の問題を解決するため，事前に学習した音響 HMM の共有化構造で画像 HMM の学習を行い，状態を単位として融合を行う手法を提案する．具体的には，音響情報のみから作成された音響 HMM を用いて，Viterbi アルゴリズムにより学習データの時間情報つきラベルを作成する．このラベルには音素の各状態ごとに時間情報が付与される．これにより画像特徴量を用いて各状態ごとに GMM を学習する．得られた 2 つのモデルを状態毎に融合し，音響-画像マルチストリーム HMM を構築する．

5 実験

5.1 実験条件

認識デコーダには，マルチストリーム HMM が扱えるように改良を施した Julius を用いた．言語モデルは 2-gram，逆向き 3-gram であり，学習にはシステムとの対話を想定して作成した 1,206 文の模擬対話文を用いた．語彙数は 6.839 単語である．評価データに数種の雑音 (白色，エレベータ，ステーション) を様々な SNR 条件で重畳して実験を行った．すべての雑音条件で提案法による認識率の改善が見られたが，本稿では，SNR=10, 15, 20dB の白色雑音を加えたときの結果を示す．

5.1.1 融合方法に関する実験・考察

音響・画像情報を単一のベクトルとして学習し，後にストリーム分割を行う手法 (従来手法) と提案する手法の比較実験を行った．この実験では画像特徴量にはオプティカルフローに基づく特徴量を用いた．混合数は，事前の実験ですべての HMM について 8 が良かったためこれを用いた．Table 2 に，最適なストリーム重みでの各融合方法の単語正解精度を示す．これから，提案法では音響のみに比べ，SNR=10dB で最大 6.8%，15dB で最大 10.5%の改善が見られた．すべての条件で従来手法と比べて同等以上の結果となり，提案法により各パラメータの学習が適切に行われることが示された．

Table 3 各画像特徴量の認識結果 (単語正解精度)

SNR	OPF	PCA	OPF+PCA
10dB	13.9%	11.1%	15.6%
15dB	35.2%	31.1%	37.9%
20dB	53.9%	51.7%	56.5%
clean	75.6%	75.9%	76.7%

5.1.2 特徴量に関する実験・考察

提案する融合法において，画像特徴量の違いに関する検討を行う．オプティカルフローに基づく特徴量のみ (OPF)，PCA に基づく特徴量のみ (PCA)，および OPF，PCA の両方を用いて 3 ストリーム HMM により認識した場合 (OPF+PCA) について比較を行った．各ストリームの混合数は，事前の実験により，音響，OPF は 8 を，PCA は 1 を用いた．Table 3 に，最適なストリーム重みでの各画像特徴量での単語正解精度を示す．これから，PCA についても音響のみに比べて正解精度が向上しているが，OPF の方がよい結果となった．また，2 つの画像特徴量を組み合わせることで正解精度が向上し，音響のみに比べ，SNR=10dB で最大 8.5%，15dB で最大 13.2%の改善が見られた．

6 まとめ

本論文では，日本語対話音声を対象とした不特定話者マルチモーダル音声認識についての検討を行った．データベースの収集を行い，構築した認識システムを様々な雑音条件で評価した．トライフォンでの状態共有を考慮した新しい音響・画像情報の融合方法を提案し，認識実験を行った結果，すべての条件で従来手法と同等以上の結果となり，手法の有効性が確認された．今後の課題として，最適なストリームを自動推定する手法の検討などが挙げられる．

参考文献

[1] S. Basu et al., Proc. MMSP'99, pp.475-481, (1999).
[2] 田村 他, 音講論, 1-1-14, (2001-10).
[3] 石川 他. FIT2002, pp.203-204, (2002-9).
[4] 田熊 他. 人工知能学会 研究報告, pp.21-26 (2002-6).
[5] http://opencvlibrary.sourceforge.net
[6] B. D. Lucas et al., Proc. DARPA Image Understanding Workshop, pp.121-130 (1981).
[7] 田村 他, 音講論, 3-6-11 (2003-9).