Vol.2010-SLP-82 No.7
2010/7/23

雑音下マルチモーダル音声認識評価基盤 CENSREC-1-AV の構築

田村哲嗣[†1] 宮島千代美[†2] 北岡教英[†2] 武田一哉[†2]
山田武志[†3] 滝口哲也[†4] 柘植覚[†5] 山本一公[†6]
西浦敬信[†7] 中山雅人[†8] 傳田遊亀[†9] 藤本雅清[†10]
松田繁樹[†11] 小川哲司[†12] 黒岩眞吾[†13] 中村哲[†11]

本稿では，音声と画像を用いたマルチモーダル音声認識の共通評価基盤 CENSREC-1-AV について紹介する．CENSREC-1-AV では，音声・画像データベースおよびベースラインシステムを提供する．音声は学習用クリーンデータのほか，乗用車走行雑音を付与したものを収録した．画像はカラー映像と近赤外線映像を収録し，ガンマ補正を用いて乗用車走行シミュレーション画像をテストデータとした．ベースラインシステムでは，MFCC と，固有顔ないしはオプティカルフローを特徴量として，マルチストリーム HMM により認識を行った．

CENSREC-1-AV
An evaluation framework for multimodal speech recognition

Satoshi Tamura,[†1] Chiyomi Miyajima,[†2] Norihide Kitaoka,[†2]
Kazuya Takeda,[†2] Takeshi Yamada,[†3] Tetsuya Takiguchi,[†4]
Satoru Tsuge,[†5] Kazumasa Yamamoto,[†6] Takanobu Nishiura,[†7]
Masato Nakayama,[†8] Yuki Denda,[†9] Masakiyo Fujimoto,[†10]
Shigeki Matsuda,[†11] Tetsuji Ogawa,[†12] Shingo Kuroiwa[†13]
and Satoshi Nakamura [†11]

This paper introduces an evaluation framework for multimodal speech recognition: CENSREC-1-AV. The corpus CENSREC-1-AV provides an audio-visual speech database and a baseline system of multimodal speech recognition. Speech signals were recorded in clean condition for training and in-car noises were overlapped for testing. Color and infrared pictures were captured as training data, and image corruption was conducted for testing using the gamma correction technique. In the baseline system, acoustic MFCC as well as eigenface or optical-flow information are adopted as audio and visual features respectively, then multi-stream HMMs are used as a recognition model.

1. はじめに

ノートパソコンや携帯電話などモバイル機器やカーナビゲーションシステムなどにおいて，音声認識はハンズフリーかつスマートなインタフェースとして注目されており，実用化に向けた研究が行われている．しかし，現在の音声認識は，実環境など雑音下において，認識精度が著しく低下してしまうという問題がある．この問題を克服する手法のひとつとして，音声信号に加え，音響雑音の影響を受けない他の信号や情報を用いる「マルチモーダル音声認識（multimodal speech recognition）」が挙げられる．その中でも，発声時の口唇動画像を用いる「バイモーダル音声認識（bimodal speech recognition, audio-visual speech recognition）」が，近年研究されるようになってきている．音声と画像を用いるバイモーダル音声認識は，従来の音声認識の技術に，いわゆる読唇の技術を加えることで，雑音下でも頑健な音声認識を可能とする手法と言い換えることができる．

マルチモーダル音声認識の研究は，コンピュータやカメラが安価になり，また雑音下音声認識の研究が盛んになってきた 1990 年代半ばから広く行われるようになってきた．例として，Potamianos らは，LDA（Linear Discriminant Analysis）と MLLT（Maximum Likelihood Linear Transform）を音声と画像の特徴量抽出に用い，マルチモーダル音声認識の研究を行っている[1)]．熊谷らは，音声と画像のモデルに隠れマルコフモデル（Hidden Markov Model，HMM）を用い，状態ごとに統合することで，音声-画像モデルを作成しマルチモーダル音声認識を行う手法を提案している[2)]．

現在の音声認識の研究では，大規模な音声コーパスやテキストコーパスを構築し，これを

†1 岐阜大学 Gifu University
†2 名古屋大学 Nagoya University
†3 筑波大学 University of Tsukuba
†4 神戸大学 Kobe University
†5 大同大学 Daido University
†6 豊橋技術科学大学 Toyohashi University of Technology
†7 立命館大学 Ritsumeikan University
†8 近畿大学 Kinki University
†9 村田機械 Murata Machinery
†10 NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories
†11 情報通信研究機構 National Institute of Information and Communications Technology
†12 早稲田大学 Waseda University
†13 千葉大学 Chiba University

Vol.2010-SLP-82 No.7
2010/7/23

用いて音響モデルや言語モデルを学習・構築する手法が広く行われている．音声認識の評価もまた，クリーン音声に雑音を重畳したデータや，雑音下で収録された音声データなど，大量のテストデータを用いて行われている．そこで，音声言語情報処理研究会（IPSJ-SLP）の下に設立された「雑音下音声認識評価ワーキンググループ」は，音響雑音に頑健な音声認識の実現を目指し，これまでに，さまざまな雑音環境で収録された学習データとテストデータ，これらを用いたベースラインシステムから成る日本語コーパス「CENSREC（Corpora and Environments for Noisy Speech RECognition）」を構築してきた．

本稿では，CENSREC シリーズの新たなコーパスである「CENSREC-1-AV」について述べる．CENSREC-1-AV は，音声・画像データベースと，ベースライン認識システムから構成される．データベースとしては，オフィス環境で収録したクリーン音声および口唇付近のカラー画像と近赤外線画像を収録した．またベースラインシステムとして，音響特徴量として現在の音声認識で広く用いられている MFCC（Mel-Frequency Cepstrum Coefficients）を，画像特徴量として固有顔（eigenface）とオプティカルフローによるものを抽出し，マルチストリーム HMM（Hidden Markov Model）により認識を行う手法を収録した．

本稿の構成について述べる．第 2 章では，CENSREC-1-AV に収録したデータベースについて説明する．ベースラインシステムについては第 3 章で紹介する．本コーパスを用いて行った予備的実験とその結果を第 4 章で述べる．最後に，第 5 章で本稿のまとめを行う．

2. データベース

2.1 タ ス ク

CENSREC-1-AV のデータベースは，連続数字読み上げタスクにおける音声・画像データから構成される．発話は AURORA-2J（CENSREC-1）[3] に準じて，原則として一発話あたり 1〜7 桁となっている．例えば「1234」であれば「いち に さん よん」と発声される．「0」については「ぜろ」と「まる」の二通りの読み方を収録した．

2.2 収 録 環 境

データベースの収録環境を図 1 に示す．音声・画像ともに雑音の少ないオフィス環境で収録した．発話者は，ブルースクリーンを背景に椅子に座り，襟元にピンマイク（ECM-77B）をつけ，発声している．マイクロホンで収録した音声は，2 台の DV カメラの音声入力端子を通じて録音した．2 台のうち片方のカメラ（VX-1000）はカラー映像を撮影し，残る片方のカメラ（DCR-TRV9）は近赤外線透過レンズフィルタにより近赤外線映像を撮影した．音声のサンプリング周波数 48kHz，量子化ビット数は 16bit である．画像はいずれも，DV

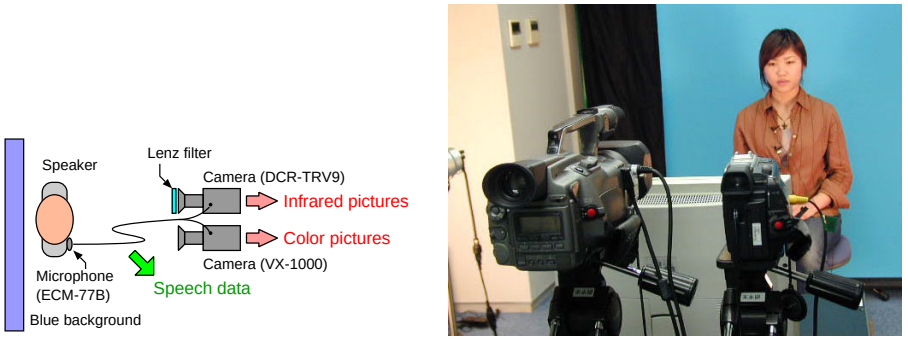


図 1 CENSREC-1-AV の収録風景

ムービー形式で 29.97fps のインターレース映像，横 720 ピクセル × 縦 480 ピクセルである．

2.3 学習セット

CENSREC-1-AV では，音声認識のモデルの学習，主成分分析による画像特徴量の計算における固有ベクトルの導出に，学習セットの音声・画像データを利用する．学習セットは男性 22 名，女性 20 名の各 77 発話，計 3,234 発話から構成されており，音声・画像ともに雑音のないクリーンデータである．

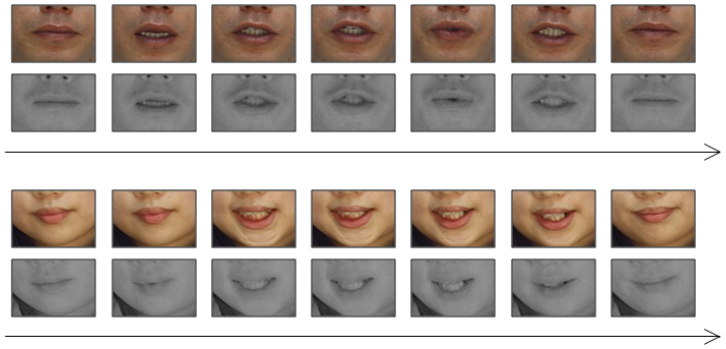
音声データは，DV カメラのムービーから音声信号のみを抽出することで取得した．このときにダウンサンプリングにより標本化周波数 16kHz とした．なお発話の前後には，それぞれ約 0.8 秒の無発声区間を設けてある．得られた音声ファイルは，WAV 形式でデータベースに収録した．

カラー映像，近赤外線映像ともに，ムービーを時系列画像データに分解した．このとき，インターレース解除と同時に，アスペクト比をあわせるため横方向を半分に縮小した．さらに，発話ごとに座標を固定して切り出し窓を設け，口唇付近の画像を切り出した．窓の座標は，OKAO-VISION[4] や HMM による切り出し[5] の結果を基に，手動で修正することで，窓から口唇がはみ出さないようにした．切り出し窓の大きさは全発話で共通とした．これにより，横 81 ピクセル × 縦 55 ピクセルの口唇画像系列を取得し，データベースに収録した．なおカラー画像は 24bitRGB 画像，近赤外線画像はモノクロ変換し 8bit グレースケール画像とし，フォーマットは Windows BMP 形式とした．これらの画像の例を図 2 に示す．

2.4 テストセット

テストセットは，学習セットには含まれていない男性 25 名，女性 26 名による計 1,963 発

Vol.2010-SLP-82 No.7
2010/7/23



上段: カラー画像，下段: 近赤外線画像

図 2 CENSREC-1-AV に収録した画像の例（男性話者，女性話者）

上段: カラー画像，下段: 近赤外線画像

図 3 ガンマ補正後の画像の例

話で構成される．テストセットのデータには，学習セットと同様の方法で収録したクリーン環境の音声・画像に加え，次に述べる雑音や外乱を付与したものを用意した．

音声については，乗用車走行雑音を重畳した．市街地道路および高速道路を走行時の乗用車雑音を，音声収録に用いたものと同じ機材で録音し，雑音データを取得した．これを S/N 比 20dB，15dB，10dB，5dB，0dB，-5dB の 6 種類の S/N 比で，もとの音声にそれぞれ重畳した．これにより，テストセットでは，各発話あたり 13 種類（2 種類の雑音 ×6 種類の S/N 比 + クリーン）の音声を利用することが可能である．

画像については，走行中の乗用車内を明度値のガンマ補正によりシミュレートし，雑音データとした．まず実際の乗用車走行時に，時刻 t における明度値 I_t（256 階調）を記録し，次式によりガンマ値 γ_t を計算しておく．

表 1 CENSREC-1-AV の学習セットとテストセット

	Training set	Test set
# spkr.	22 males and 20 females	25 males and 26 females
# utter.	3,234 utterances (77 utter/spkr)	1,963 utterances (38–39 utter/spkr)
Speech data	clean	clean, in-car noise (city roads, expressway)
	monaural, 16kHz, 16bit, WAV files	
Image data (lip-around pictures)	clean	clean, simulated (Gamma-controlled)
	81×55, 29.97fps, BMP files 24bit RGB (color), 8bit grayscale (infrared)	

$$\gamma_t = \frac{logI_t - log255}{log\bar{I} - log255} \tag{1}$$

ここで $\bar{I}$ は明度の時間平均であり，次式により求められる．

$$\bar{I} = \frac{1}{N}\sum_{t=1}^{N} I_t \tag{2}$$

このガンマ値 γ_t を用い，以下の式により，時刻 t の画像における座標 (x, y) の明度 $I(x, y, t)$ を変更することで，画像に対するノイズ付与後の明度 $I'(x, y, t)$ を求めた．

$$I'(x, y, t) = 255\left(\frac{I(x, y, z)}{255}\right)^{\gamma_{t+\tau}} \tag{3}$$

ここで，τ は発話ごとに決める補正開始位置である．ガンマ補正後の画像の例を図 3 に示す．図 3 では，陸橋の下を通過した時の γ_t を使用している．

以上の学習セットおよびテストセットの仕様をまとめたものを表 1 に示す．

3. ベースラインシステム

3.1 音響特徴量

音響特徴量は MFCC（Mel-Frequency Cepstrum Coefficient）12 次元と対数パワー，それらの Δ，ΔΔ 成分の計 39 次元とした．フレーム長は 25ms，フレームレートは 10ms（100Hz）である．

図 4 固有顔特徴量における固有顔（カラー画像，10 次元）

3.2 画像特徴量

ベースラインでは，画像特徴量抽出に，固有顔（固有唇）とオプティカルフローによる 2 種類の手法を用意した．モデルの学習および認識では，これらのいずれかを画像特徴量として用いるようになっている．画像特徴量のフレームレートは，ムービーのフレームレートと同一の 29.97Hz である．

3.2.1 固有顔特徴量

固有顔では，はじめに，学習セットから 4,620 枚の画像を抽出し，それぞれ 40×27 に縮小しグレースケール化したのち，画像の左上から右下にかけてラスタスキャンを行い，変換行列計算用の 1,080 次元のベクトルを作成した．この 1,080 次元のベクトル 4,620 個を用いて主成分分析を行い，固有値の大きい方から 10 次元に対応する固有ベクトルを算出した．この固有ベクトルを用いて，各画像の主成分得点を計算し，10 次元の特徴量を抽出した．これに Δ，ΔΔ 成分を加えた 30 次元を画像特徴量とした．なお固有ベクトルは，カラー画像，近赤外線画像それぞれ別に用意した．得られた固有ベクトルを画像で表現したもの（カラー画像，10 次元）を，図 4 に示す．このときの累積寄与率は 87.7%である．

3.2.2 オプティカルフロー特徴量

オプティカルフローを用いる場合は，グレースケール化した隣接する 2 枚の画像を用いて，Horn-Schunck 法[6] により，画素ごとにオプティカルフローを計算した．次にフローベクトルの水平方向と垂直方向の分散値 2 次元を計算した[7]．そして，これらの Δ，ΔΔ 成分を求め，6 次元の画像特徴量を算出した．

3.3 音響・画像特徴量

音響特徴量と画像特徴量を連結することで，マルチモーダル音声認識のための音響・画像特徴量を生成した．このとき，3 次元スプライン補間を用いて時間方向に画像特徴量を補間し，フレームレートを 100Hz としたうえで，音響特徴量との連結を行った．

3.4 モデル構築

学習セットの音響・画像特徴量を用いて，マルチモーダル音声認識のモデルを連結学習した．認識用のモデルとして，音声と画像の重みづけが可能な，マルチストリーム HMM を使用した．マルチストリーム HMM では，音響・画像特徴量 $\mathbf{O}_{AV}$ に対する対数尤度 $b_{AV}(\mathbf{O}_{AV})$ は，以下の式で表わされる．

$$b_{AV}(\mathbf{O}_{AV}) = \lambda_A b_A(\mathbf{O}_A) + \lambda_V b_V(\mathbf{O}_V) \quad (4)$$

ここで $b_A(\mathbf{O}_A)$ は音響特徴量 $\mathbf{O}_A$ に対する音響対数尤度，$b_V(\mathbf{O}_V)$ は画像特徴量 $\mathbf{O}_V$ に対する画像対数尤度である．λ_A，λ_V は音響ストリーム重み，画像ストリーム重みであり，以下の制約で変化させる．

$$\lambda_A + \lambda_V = 1 \quad (5)$$

従来の CENSREC コーパスに準じて，モデルは left-to-right 型で，11 種類の数字 HMM および無音 HMM（sil）とショートポーズ HMM（sp）を用意した．HMM の状態数はそれぞれ 16，3，1 とした．

3.5 音声認識・評価

本コーパスを用いて作成したマルチモーダル音声認識システムは，テストデータを用いて評価できる．マルチストリーム HMM におけるストリーム重みは，式 (5) の制約のもと，0.0 から 1.0 まで 0.1 刻みで変化させるものとする．このとき，全ての HMM で同じストリーム重みを使用する．なお挿入ペナルティは設定せずデフォルト値をそのまま使用する．

認識性能の評価では，音声のみ（通常の音声認識），画像のみ（読唇），マルチモーダルの 3 条件ごとに，ノイズあり・なし（音声の場合は S/N 比ごと）の性能比較を行うことができる．これにより，ベースラインシステムと，本コーパスの利用者が作成したマルチモーダル音声認識システムを，共通の基準で評価することが可能になる．

4. 参 考 実 験

4.1 画像特徴量・モデルの性能評価

はじめに，画像特徴量や画像側モデル（画像ストリーム）に関する考察として，画像特徴量の補間の有無，および画像モデルの混合数による認識性能の影響について調査した．固有顔特徴量 30 次元について，補間前（29.97Hz）と補間後（100Hz）の異なるフレームレートのものを用意した．そして補間前特徴量と補間後特徴量それぞれにおいて，連結学習により画像のみモデルを構築した．HMM のトポロジーはベースラインと同じ（left-to-right，状態数 16）である．混合数は，1，2，3，4，6，8，10，12，16，20 の 10 通りを調査した．

また評価には，テストセットのクリーン画像データを用いた．

実験結果を図 5 に示す．カラー，近赤外線ともに，補間前と補間後それぞれにおいて性能最大となるところで比較すると，補間後の方が良い性能が得られた．また補間後の近赤外線の場合を除き，混合数 10 で性能最大となっていることが判明した．なお補間後の近赤外線は，混合数が増えると急激に性能が低下していくように見えるが，これは挿入ペナルティを適切に設定することで対処可能である．

4.2 マルチモーダル音声認識の性能評価

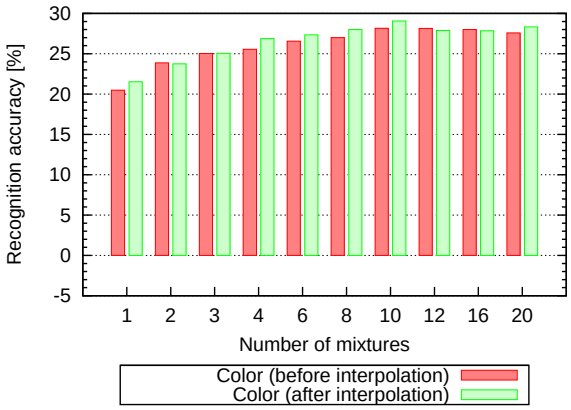
次に，本コーパスのデータベースを用いたマルチモーダル音声認識の実験を行った．画像特徴量に固有顔特徴量 30 次元を用い，モデル構築法として以下の 2 手法を検討した．(1) は特徴量に音響・画像特徴量を用いる以外は従来の音声認識と同様の方法で学習を行い，認識時にマルチストリーム HMM に変換する手法である．(2) では，音響ストリームと HMM 内の遷移確率は音響のみモデルで学習したものを，画像ストリームは音響・画像特徴量で学習したものを利用している．

(1) (a) シングルストリーム HMM（通常の HMM）を用意し，学習セットの音響・画像特徴量 69 次元を用いて，連結学習によりモデルパラメータを学習
(b) 学習後，39 次元の音響ストリームと 30 次元の画像ストリームをもつマルチストリーム HMM に変換

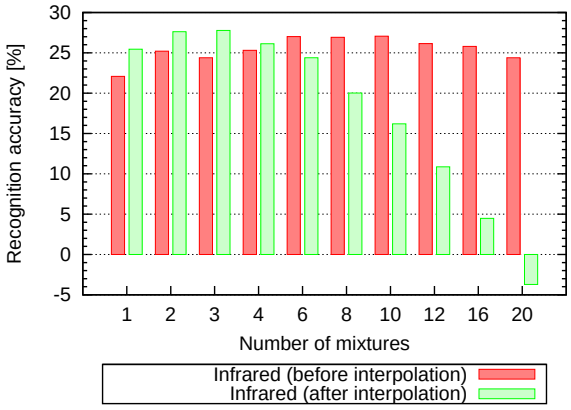
(2) (a) 音響特徴量 39 次元を用いて音響情報のみのモデルを連結学習（通常の音声認識と同様）
(b) 音響・画像用のシングルストリーム HMM を用意（このとき遷移確率として (a) で得られたモデルのものを使用）
(c) シングルストリーム HMM の 遷移確率以外の パラメータを音響・画像特徴量 69 次元で連結学習
(d) 得られたシングルストリーム HMM をマルチストリーム HMM に変換
(e) (d) の音響ストリームのパラメータを (a) の音響モデルのもので置換

認識実験は，音響雑音は市街地走行雑音，高速道路走行雑音ともに 3 種類の SNR（-5dB, 5dB，15dB）で重畳し，画像はクリーンの条件で行った．音響重みと画像重みは，本実験では 0.2 刻みで変化させた．

カラー画像を用いた時の認識結果を表 2 に示す．比較のため，音響特徴のみで学習・認識したもの（Audio only），画像特徴のみで学習・認識したもの（Visual only）の結果をあわせて記載してある．なお今回用いた乗用車走行雑音は低周波領域にパワーが集中しており，



(A) カラー画像



(B) 近赤外線画像

図 5 画像特徴量の補間の有無と混合数による性能の変化

SNR の値と比べて，音声のみでも比較的高い認識性能となっていることに注意されたい．

(1) のモデルは，音響のみの結果よりも性能が低下している．この要因として，音声と画像は同期ずれがあり[8]，これをモデル化するには現在のトポロジーでは状態数が不足している可能性や，音声と比べて性能の低い画像情報を併用したため，モデル学習が有効に行なえ

表 2 マルチモーダル音声認識結果（カラー画像）

手法 (1) の認識結果

		cityroad			expressway		
		-5dB	5dB	15dB	-5dB	5dB	15dB
Audio only		80.92	98.22	99.41	58.05	97.05	99.43
multi-modal	$\lambda_A = 1$	36.07	41.70	65.09	33.45	48.76	74.14
	best	54.86	72.60	92.34	50.57	75.66	94.93
	(λ_A)	(0.4)	(0.4)	(0.4)	(0.4)	(0.6)	(0.6)
	$\lambda_A = 0$	38.09					
Visual only		27.58					

手法 (2) の認識結果

		cityroad			expressway		
		-5dB	5dB	15dB	-5dB	5dB	15dB
Audio only		80.92	98.22	99.41	58.05	97.05	99.43
multi-modal	$\lambda_A = 1$	80.92	98.22	99.41	58.11	97.07	99.43
	best	88.02	98.96	99.57	79.37	98.53	99.52
	(λ_A)	(0.8)	(0.8)	(0.8)	(0.6)	(0.8)	(0.8)
	$\lambda_A = 0$	36.05					
Visual only		27.58					

ていない可能性が考えられる．なお画像のみモデルと，マルチストリーム HMM の $\lambda_A = 0$（画像ストリームのみ使用）の結果と比べると，マルチストリームの方が性能が良い．これは逆に，識別能力の高い音声情報とあわせて学習することで，画像側からみると良いモデルができたためと考えられる．

一方 (2) のモデルは，$\lambda_A = 1$ は音響のみモデルと同等のため，ほぼ同一の性能が得られている．また (1) の考察で述べたように，$\lambda_A = 0$ では，音声情報を用いることで高い性能の画像パラメータが得られたため，画像のみモデルよりも性能が向上している．さらに両者を組み合わせてストリーム重みを最適化することにより，市街地走行雑音-5dB では 37%，高速道路走行雑音-5dB では 51%の誤り率削減に成功した．CENSREC-1-AV のベースラインにおけるモデル構築手法も，これと同類の学習方式を採用する予定である．

5. おわりに

本稿では，音声と口唇画像を用いたバイモーダル音声認識や，読唇の研究に利用できる，雑音環境下マルチモーダル音声認識評価基盤 CENSREC-1-AV について報告した．本コーパスが，新たなマルチモーダル音声認識や読唇技術における特徴量や統合技術の確立，ならびに音声認識性能の改善において活用されることを望みたい．また，新たにマルチモーダル音声認識や読唇手法の研究を考えている研究者の一助となれば幸いである．

謝　辞

本論文にあたっては，国立情報学研究所音声資源コンソーシアムおよび名古屋大学末永研究室よりご協力をいただきました．ここに御礼申し上げます．

参 考 文 献

1) G.Potamianos et al., “Discriminative training of HMM stream exponents for audio-visual speech recognition,” Proc. ICASSP’98, vol.6, pp.3733-3736, Seattle, U.S.A. (1998).

2) 熊谷ほか，「HMM 合成を用いたバイモーダル音声認識」日本音響学会 2000 年秋季研究発表会, 2-Q-11, pp.111-112 (2000).

3) S.Nakamura et al., “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” IEICE Trans. on Information and Systems, Vol.E88-D, No.3, pp.535–544 (2005).

4) `http://www.omron.com/r&d/coretech/vision/okao.html` .

5) 田村ほか，「マルチモーダル音声認識における音響・画像特徴量の融合法に関する検討」日本音響学会 2003 年秋季講演論文集, 3-6-11, pp.123-124 (2003).

6) B.K.P.Horn et al., “Determining optical flow,” Artificial Intelligence, vol.17, pp.185-203 (1981).

7) K.Iwano et al., “Bimodal speech recognition using lip movement measured by optical-flow analysis,” Proc. HSC2001, pp.187-190, Kyoto, Japan (2001).

8) 中村ほか，「バイモーダル音声認識における音素境界を越えた同期性のモデル」日本音響学会 2001 年秋季講演論文集, 1-1-13, pp.25-26 (2001).