2005－SLP－55　(7)
2005／2／4

バイモーダル車内音声認識評価用データベースの構築

根木 大輔† 前野 俊希† 北坂 孝幸† 森 健策† 末永 康仁† 宮島 千代美†
伊藤 克亘† 武田 一哉† 板倉 文忠†† 佐野 昌己††† 二宮 芳樹††††

†名古屋大学大学院情報科学研究科　††名城大学理工学部情報工学科
†††名古屋商科大学経営情報学部　††††(株)豊田中央研究所

E-mail: †{dnegi, tmaeno}@suenaga.m.is.nagoya-u.ac.jp,
†{kitasaka, kensaku, suenaga, miyajima, itou, takeda}@is.nagoya-u.ac.jp
††itakuraf@ccmfs.meijo-u.ac.jp †††sano.masami@nucba.ac.jp ††††ninomiya@mosk.tytlabs.co.jp

あらまし　近年，現実の雑音環境下の様々なシーンにおいて音声認識率を向上させるために，音声情報と映像情報を統合したバイモーダル音声認識への関心が高まっている．映像情報は音響雑音の影響を受けない情報源として，音声認識において重要な役割を果たすものと考えられる．しかし，大規模バイモーダルデータベースが少ないことなどから，映像情報は実際の音声認識システムにおいて十分に利用されるには至っていない．そこで我々は，これまでに構築されている雑音環境下音声認識評価用共通データベース AURORA-2J/AURORA-3J の仕様に則り，高品質カラー映像と近赤外映像を付加して収録を行い，新しいデータベース AURORA-2J-AV（室内），AURORA-3J-AV（自動車内）を構築している．本稿ではこれらのデータベースの詳細について述べる．

キーワード　バイモーダル音声認識，マルチメディアデータベース，AURORA

Construction of Bimodal Database for Evaluating In-Car Speech Recognition

Daisuke NEGI†, Toshiki MAENO†, Takayuki KITASAKA†, Kensaku MORI†,
Yasuhito SUENAGA†, Chiyomi MIYAJIMA†, Katsunobu ITOU†, Kazuya TAKEDA†,
Fumitada ITAKURA††, Masami SANO†††, and Yoshiki NINOMIYA††††

†Graduate School of Information Science, Nagoya University
††Department of Information Engineering, Faculty of Science and Technology, Meijo University
†††Faculty of Management Information Science, Nagoya University of Commerce & Business
††††Toyota Central R&D Lab., Inc.

E-mail: †{dnegi, tmaeno}@suenaga.m.is.nagoya-u.ac.jp,
†{kitasaka, kensaku, suenaga, miyajima, itou, takeda}@is.nagoya-u.ac.jp
††itakuraf@ccmfs.meijo-u.ac.jp †††sano.masami@nucba.ac.jp ††††ninomiya@mosk.tytlabs.co.jp

Abstract　There are remarkable interests on bimodal speech recognition, which integrate audio and visual information, to improve speech recognition rates. Visual information plays a very important role in speech recognition since it is not affected by acoustic noises. However, such kind of information has not been fully used in existing actual speech recognition systems because of the lack of large-scale bimodal databases. Therefore we are building new databases called *AURORA-2J-AV*(indoor) and *AURORA-3J-AV*(in-vehicle) that contain aural signals and high quality facial images taken by color and near-infrared cameras. The utterance tasks of these databases are the same as those of our AURORA-2J/AURORA-3J database for evaluating speech recognition method under noisy environments. This paper describes the detailed specification of the databases.

Keywords　Audiovisual automatic speech recognition，Multimedia database，AURORA

1. はじめに

近年，自動車内でのカーナビゲーションシステムのインタフェース利用や擬人化エージェントとの対話インタフェースを初めとして，実環境における音声認識の重要性が増加している．このため雑音が多く含まれる実環境下での音声認識の研究が盛んに行われている[1]．現在の音声認識システムはクリーンな発話に対し90%以上の安定した認識率を示すが，実環境の雑音を多く含む発話では著しく性能が低下する．そのため，街中や自動車内でインタフェースとして音声認識を利用するためには，雑音に対するロバスト性のさらなる向上が必要である．雑音に影響されずに発話内容を認識可能な特徴として，話者を撮影した映像情報が挙げられる．音声情報に加え映像情報を用いて音声認識を行

うことで，雑音下での認識率向上が期待できる．このようなバイモーダル音声認識の研究が近年盛んに行われている[2].

我々の研究グループでも音声と映像を統合して自動車内における音声認識の精度向上を目指す研究を行ってきた．発話区間の検出精度は音声認識率に大きな影響を与えるため，正しく認識を行うために発話区間を正確に推定することが必要となるが，非定常雑音環境下では音声情報のみを用いて発話区間を決定することは非常に難しい．そこで，口唇画像から唇の動きを検出し，雑音環境下でも安定して発話区間を推定する研究が行われている[3]．また我々は，音声情報を併用することで発話区間の推定精度の向上を図る研究を行ってきた[4][5]．これらの研究では，話者の口唇領域等の映像から得られる特徴量を音声情報に加えて利用することで，音声認識の雑音に対する頑健性が向上することが示されている．

映像情報は音響雑音による影響を受けない有望な情報源であるが，実際の音声認識システムにおいて十分に利用されるには至っていない．この一因として，公開されている大規模バイモーダル音声データベースの数およびそれに含まれるサンプル数が少ないことから，実環境の雑音に適したバイモーダル音声認識手法とその効果が示されていないことが挙げられる．実環境の雑音を収録した公開データベースがあれば，同一の音声認識タスクに対して複数の研究グループがそれぞれの手法を適用し，性能改善を競い合うことで，より優れた音声認識手法を確立することが期待される．

音声情報のみを用いた音声認識の分野ではこのようなデータベースの作成が行われており，AURORA プロジェクトの AURORA-2 データベースがある[6]．AURORA-2 では共通の発話に様々な実環境の雑音を重畳して用いることで，様々な音声認識システムの性能や，雑音に対する頑健性を比較することが可能である．数字を日本語に翻訳し，同一の雑音データを重畳した AURORA-2J データベースも広く利用されている[7]. また同様のデータベースに，自動車内での連続数字／コマンドタスクからなる AURORA-3 がある．AURORA-3 では，アイドリングや市街地走行を行いながらオーディオやエアコンの ON/OFF を切り替え，様々な雑音条件下で収録を行っている．AUROR-3 の日本語版に相当する AURORA-3J の構築も現在進められている[7].

AURORA-2/AURORA-2J ， AURORA-3/AURORA-3J は，音声データのみをもつデータベースである．音声情報と映像情報を統合し，より優れた雑音に頑健なバイモーダル音声認識システムを開発するためには，音声情報だけでなく映像情報を同時に収録したデータベースが必要になる. しかし，この分野は比較的新しく，そのような公開されたデータベースは非常に少ないのが現状である[2]．これまでに作成されたデータベースは比較的小規模であり，室内のみのものが多い．自動車内で収録されたものもあるものの[8]，一般に配布されていない. そこで本稿では, AURORA-2J, AURORA-3J データベースの仕様に則り，さらにカラー映像と近赤外映像を加えて収録を行った AURORA-2J-AV，AURORA-3J-AV データベースについて述べる．このデータベースは一般への配布を前提としている点，室内だけでなく自動車内のデータを含む点において従来とは一線を画す．さらに，AURORA データベースには，性能改善の際の評価の基準となる最低限の認識性能を出す標準スクリプトが含まれている．AURORA-2J-AV, AURORA-3J-AV はバイモーダルデータベースであるため，この標準スクリプトも音声と映像の両者を用いたものにする必要がある．そこで，映像情報を用いたベースとなる音声認識手法を実装し，その性能評価を行う．

以下，2.で既存の AURORA 音声データベースについて述べ，3. で今回収録した AURORA-2J-AV，AURORA-3J-AV について説明する．4.でデータベースの応用例を簡単に示す. 5.で映像情報のみを用いた音声認識の予備実験について説明する．

2. AURORA 音声データベース

2.1 概要

AURORA プロジェクトのデータベース（AURORA-2, AURORA-3）は，複数の研究チームに共通の雑音を重畳した発話データを配布し，様々な音声認識システムの認識性能，雑音に対する頑健性を比較するためのものである．これらのデータベースは連続数字やコマンドなどの比較的単語数の少ない発話タスクを収録したものであるため，実環境に近い過酷な雑音環境下でも音声認識が比較的容易である．また，認識から評価までの枠組みがあらかじめ揃っており，評価の基準となるベースライン性能を計算するスクリプトが用意されていることが利点である．また，これらのタスクを日本語化した AURORA-2J データベース，AURORA-3J データベースも作成されている．

2.2 AURORA-2/AURORA-2J

AURORA-2/AURORA-2J データベースには，連続数字コーパス（TI digit に現実環境の雑音を重畳したもの）と，それを認識するための HTK（Hidden Markov Model Toolkit）[9]を用いた標準スクリプト，標準スクリプトを改良せずに音声認識を行った場合のベースライン性能からの性能改善率を算出する Microsoft Excel ワークシートが含まれる．地下鉄やレストラン，人混みなどの様々な実環境のノイズを SNR を変化させながらクリーンな発話音声に重畳することで，実環境下での発話をシミュレートする．HMM の学習条件や SNR の違いによって複数の条件が規定されており，各条件における音声認識率が算出される. この結果を Excel ワークシートに入力することで，その音声認識システムの性能をベースライン性能からの相対性能として自動的に集計，比較することができる．

AURORA-2 の発声リストは，1〜7 桁の連続数字で構成され，“1234567”や“O53Z”などと表記される（“0”は/zero/もしくは/oh/と発声し，それぞれを“Z”および“O”と表記する）．AURORA-2 では学習と評価の条件が明確に規定されており，それぞれに用いる発声リストも分けられている. 学習セットは発声リスト約 40 文，評価セットは約 80 文で構成される．被験者 1 人が学習セットまたは評価セットの 1 つのセットを発声し，データベース全体では学習セット 110 名（男女 55 名），評価セット 104 名（男女 52 名）分のデータで構成される．

AURORA-2J は，AURORA-2 の発話リストを日本語に翻訳し，同一の雑音を重畳したものである．日本語

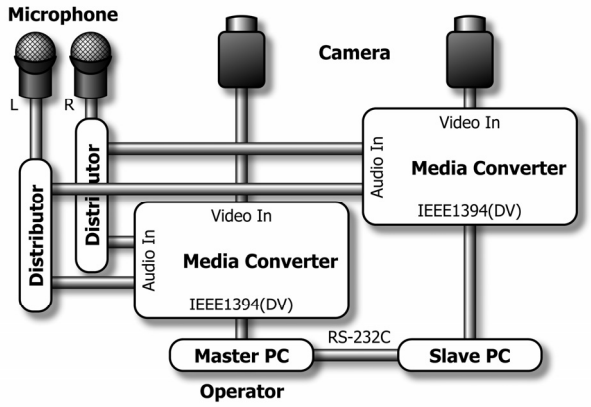

図 1　収録システム構成

表 1　収録データの仕様

File Format	MSDV Type-1 AVI
Transfer Rate	3.6MB/s (28.8Mbps)
Video	NTSC インターレース, 720×480, 29.97fps
Audio	48kHz, 16bit, Stereo

では“0”を「ぜろ」「れい」「まる」と発声するが，「ぜろ」と「まる」と発声される比率が高い．そこで，AURORA-2 の日本語化に際して，“Z”を「ぜろ」，“O”を「まる」としている．これと同様の理由で，“4”を「し」と発音する，あるいは，“7”を「しち」と発声することはしていない．また，“2”や“5”の長母音化（「にー」，「ごー」）については特に指示せず，被験者の自由としている．

2.3 AURORA-3/AURORA-3J

AURORA-3 は，自動車内での連続数字／コマンドタスクで構成される雑音環境下における音声認識評価用共通データベースである．被験者は実際の自動車で運転席に座り，アイドリングや市街地走行を行いながら，オーディオやエアコンの ON/OFF を切り替え，様々な走行条件下で収録を行う．音声の収録には，接話マイクとハンズフリーマイクを用いる．AURORA-2 では雑音を後から人工的に重畳するが，AURORA-3 はオーディオ，ロードノイズ，エンジン音，車のすれ違い音など，様々な雑音下で発話音声が収録されていることが特徴である．AURORA-2 と同様に，ベースライン性能を計算する標準スクリプト，ベースライン性能からの性能改善率を算出する Microsoft Excel ワークシートを含む．また，AURORA-3 の日本語版の AURORA-3J データベースの整備も現在進められている[7].

3. 映像付き AURORA 音声データベース

3.1 収録システム

自動車内で収録された顔映像を利用するためには，夜間は可視光の照明が利用できないため，近赤外照明による近赤外映像が用いられる．今回のデータベースでは通常のカラー映像に近赤外映像を加え，2 系統で収録を行った．収録システムの構成を図 1に示す．2 台の DV カメラを用いて，被験者正面から発話中の顔を撮影する．1 台はカラー映像，もう 1 台は近赤外映像とする．音声は 2 つのモノラルピンマイクを被験者の襟と胸元に取り付けて収録を行う．収録された映像と音声は DV フォーマットでエンコードされ（以下この処理を単に DV エンコードと書く），PC に保存される．今回は可視光画像，近赤外画像の 2 つを撮影するために異なった種類の DV カメラを 2 台使用している．DV エンコードに DV カメラが内蔵する機能を利用すると，それぞれの機器が持つ A/D コンバータや AGC（Auto Gain Control）の特性の違いにより記録される音声データに差異が生じてしまうため，DV エンコード用に 2 系統でそれぞれ同一のメディアコンバータを使用した．メディアコンバータはアナログ映像入力とアナログ音声入力をディジタル変換してインターリーブし，IEEE1394(DV)ストリームに変換して出力する．各系統で DV カメラ，メディアコンバータ，Windows PC が接続される．映像は DV カメラからメディアコンバータに入力される．音声は，2 つのマイクそれぞれに接続された分配器で 2 系統に分配され，2 つのメディアコンバータの L/R チャンネルに接続される．したがって，カラー映像＋ステレオ音声，近赤外映像＋ステレオ音声の 2 系統が収録される．

音声と映像を同期させて収録するにあたり，データの取り扱いやすさや画質と音質などを考慮して，MSDV Type-1 AVI ファイル形式で収録を行った．これは DV テープに再エンコードなしで書き戻すことができるフォーマットであり，映像・音声共に高いビットレートで保存されるため，画質や音質の劣化が比較的少ない．音声は 48kHz･2ch とした．収録データの仕様の詳細を表 1に示す．

メディアコンバータから出力される IEEE1394(DV)ストリームを，PC の IEEE1394 インタフェースに入力し，DV ストリームを AVI ファイルとして保存する．オペレータはマスタ PC で録画開始・停止などの操作を行い，これらの情報はシリアル通信でスレーブ側に送信され，スレーブ PC はマスタ PC と同様の動作をする．マスタ PC は，動作指示をスレーブ PC に送信し，スレーブ PC の動作開始を待ってから動作を開始するため，録画開始・停止などはマスタ側が最大で数百 msec 遅れる．このため，それぞれの PC に保存される AVI ファイルの録画開始時刻やデータの長さなどは若干異なるものになるが，ファイル内の映像ストリームと音声ストリームは同期している．

3.2 AURORA-2J-AV

3.2.1 データベースの内容

AURORA-2J-AV は，雑音環境下連続日本語数字音声認識タスクの共通評価フレームワーク AURORA-2J に，発話中の被験者の顔を撮影したカラー映像と近赤外映像を加えて収録を行ったデータベースである．

発話セットは AURORA-2J のものをそのまま採用した．個々のセットで用いられる数字とその読みを表 2に，発話の例を表 3に示す．被験者数は学習セットが 41 名（男性 19 名，女性 22 名），評価セットが 49 名（男性 23 名，女性 26 名）である（表 4）．被験者の男女比ができるだけ等しくなるように，また 20 代から 50 代の幅広い年齢層を対象とした．データのサンプル映像を図 2に示す．

3.2.2 収録方法

オペレータがマスタ PC で録画開始を指示すると，2

表 2　数字の読み

数字		数字	
1	いち	7	なな
2	に	8	はち
3	さん	9	きゅう
4	よん	Z	ぜろ
5	ご	O	まる
6	ろく		

表 3　AURORA-2J-AV 発話例

記号表記	カナ表記
83966	ハチ/サン/キュウ/ロク/ロク
Z845768	ゼロ/ハチ/ヨン/ゴー/ナナ/ロク/ハチ
Z	ゼロ
33O31	サン/サン/マル/サン/イチ

表 4　被験者の年齢層と男女別人数

	学習セット		評価セット		
	男	女	男	女	合計(人)
20 代	9	9	9	7	34
30 代	6	8	7	9	30
40 代	4	5	4	5	18
50 代	0	0	3	5	8
合計(人)	19	22	23	26	90

（カラー映像）　　（近赤外映像）

図 2　AURORA-2J-AV 映像サンプル

系統で録画が開始される．録画開始 2 秒後に被験者前方の液晶モニタに発話内容が表示され，被験者は表示された連続数字を読み上げる．読み上げが終了したら，オペレータは録画停止を指示し，2 秒後に録画が停止する．

被験者前方の液晶モニタには常に被験者の映像が表示されている．被験者は鏡を見ている状態になり，画面の中央からずれていないか，頭頂部や顎が画面からはみ出ていないかなどを自分自身でチェックすることができる．なお，カラーカメラは SONY VX1000 を，近赤外カメラには SONY DCR-TRV9 に可視光カットフィルタを取り付けて使用した．

3.3 AURORA-3J-AV

3.3.1 データベースの内容

AURORA-3J-AV は，実環境（自動車内）日本語連続数字データベース AURORA-3J の仕様に則り，発話音声に被験者の顔を撮影したカラー映像と近赤外映像を加えて収録したものである．

収録を行った発話内容を表 5，表 6に示す．発話セットは 10 種類あり，それぞれの発話セットは 1 桁の独立数字 4 つ，10 桁数字が 4 つ，16 桁数字が 1 つで構成される．10 桁数字や 16 桁数字は，3 桁ないし 4 桁に区切って発話を行う．このような 9 つの数字列を，走行条件 2 種類（アイドリング，低速走行）と，車内環境条件（ノーマル，オーディオ ON，ハザード ON，エアコン ON）の組み合わせからなる 7 つの収録条件で繰り返し収録する（表 7）．AURORA-3 データベースではアイドリング，低速走行，高速走行の 3 つの走行条件で収録が行われているが，今回はアイドリングと市街地走行のみとした．また，各発話セットで年齢層，男女比のバランスを取るように被験者を配分した（表 8）．

表 5　発話セット内容

発話内容	収録数(文)
1 桁数字	4
10 桁数字(3/4/4 桁に区切って発声)	4
16 桁数字(4/4/4/4 桁に区切って発声)	1

表 6　AURORA-3J-AV 発話例

記号表記	カナ表記
1 桁数字	8, 3, 9, 1
10 桁数字	218-670-5439, 913-649-9567
16 桁数字	9376-0549-3856-5248

表 7　収録条件

アイドリング	ノーマル
	オーディオ ON
	ハザード ON
市街地走行	ノーマル
	オーディオ ON
	エアコン ON
	オーディオ ON＋エアコン ON

表 8　被験者の年齢層と男女別の人数

	男	女	合計(人)
20 代	3	5	8
30 代	2	2	4
40 代	1	3	4
50 代	3	4	7
合計(人)	9	14	23

AURORA-2J-AV と異なり，運転中のタスクが含まれるために発話内容を視覚的に提示することはできない．そこで，被験者の片耳にイヤフォンを装着させ，音声により発話内容を指示する．被験者は，イヤフォンから聞こえてきた数字列をリピートする．収録したデータのサンプル映像を，図 3に示す．

3.3.2 収録方法

収録には，名古屋大学統合音響情報研究拠点（CIAIR）で構築された収録車[10]を用いた．カラーカメラ（SONY DXC-200A（カメラ・レンズ一体型））をダッシュボード右側に，近赤外カメラ（カメラ本体 SONY XC-E150, レンズ SONY VCL-08YM, 可視光カットフィルタ装着）をステアリングコラム上部に取り付けた．また，近赤外 LED ユニット（波長 870nm）2 基をダッシュボード上部に取り付けた．AURORA-3 では，接話マイクとハンズフリーマイクの 2 種類のマイクを用いて収録され

（カラー映像）　　（近赤外映像）

図 3　AURORA-3J-AV 映像サンプル

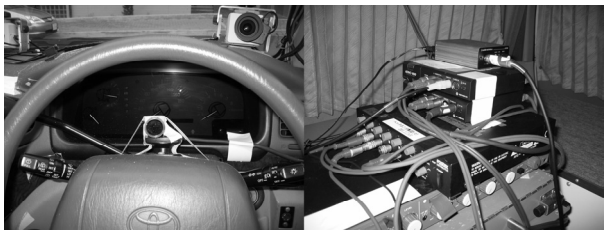

図 4　収録システム

ている．しかし，接話マイクが顔に被ることを避けるため，襟元にマイクを付けることとし，ハンズフリーマイクはサンバイザ付近に取り付けた．その他に，赤外線 LED ユニットドライブ装置，収録用 PC，マイクアンプ，メディアコンバータなどを収録車に搭載した（図 4）．

10 桁数字の場合，発話内容は 3 つに区切って，16 桁数字であれば 4 つに区切って被験者に提示される．オペレータがマスタ PC で録画開始を指示すると，2 系統で録画が開始される．録画開始 2 秒後に，イヤフォンを通して被験者に最初の区切りの発話内容が指示される．発声が終了すると，オペレータは次の区切りの音声を再生する指令を出し，これを 3 回ないし 4 回繰り返す（1 桁数字の場合は 1 回のみ）．最後の区切りの発声が終了すると同時に，オペレータは録画停止を指示し，2 秒後に録画が停止する．

4. データベースの応用例

本データベースの応用例として，バイモーダル音声認識が挙げられる．我々はこれまで，画像情報と音声情報を統合し，雑音の多い実走行車内環境で利用可能な発話区間検出手法を開発してきた[4][5]．発話特徴量として，画像情報からはドライバの口唇画像における低輝度画素領域の面積変化を，音声情報からは対数パワーの変化を利用している．また，夜間への適用を考慮し，色情報は使用していない．

このような手法の評価と改良を行うためには，音声と映像が同期して収録されたデータベースが必要である．これまでは，名古屋大学 CIAIR 実走行車内音声データベース**エラー! 参照元が見つかりません。**[11]で収録されたドライバ映像と音声のデータを用いて評価を行ってきた．しかしながら，このデータは音声と映像が完全には同期していないこと，また，映像の解像度が低く，低画質であることが問題であった．今回収録した AURORA-3J-AV データベースは，音声と映像が同期された状態で収録され，映像も比較的高画質であり，バイモーダル音声認識手法の評価に非常に有用である．今後，映像と音声を統合した音声認識・発話区間検出手法を本データベースに適用し，様々な雑音環境下における音声認識の精度向上を図る予定である．

図 5　固有画像

（上段左から右に向かって第 1,2,･･･,5 固有画像，
中段左から右に向かって第 6,7,･･･,10 固有画像，
下段左から右に向かって第 11,12,13,14 固有画像）

5. 口唇領域画像を用いた発話認識

データベースと同時に配布される標準スクリプトを用いることで，手法の改善を行う際に評価の基準となるベースライン認識性能を簡単に算出できることは AURORA データベースの大きな特徴である．今回収録を行った室内の AURORA-2J-AV および自動車内の AURORA-3J-AV はバイモーダルデータベースであるため，ベースライン認識手法も音声と映像の両方を用いたものに変更する必要がある．そこで，HMM を用いて，AURORA-2J-AV データベースの映像のみを使用した予備的な発話認識実験を行った．なお，HMM の学習および認識には HTK を使用した．

5.1 実験

認識用の画像として AURORA-2J-AV の近赤外映像をグレースケール変換し，インターレース映像のトップフィールドのみを使用した．ボトムフィールドは使用しないため，画像サイズは縦方向が半分の 720×240 ピクセル，フレームレートは 29.97fps となる．口唇領域は 120×45 ピクセルとした．学習セット 47 名，合計 3,608 発話について，手動で与えた鼻領域でテンプレートマッチングを行い，検出された鼻の位置を基準に口唇領域の切り出しを行い，全フレームについて目視でチェックを行った．約 45 万枚の口唇領域画像からランダムで選んだ 20,000 枚について主成分分析を行い，累積寄与率 80%以上となる 14 次元までの固有画像を得た（図 5）．ただし，発話映像には前後に 1.5 秒程度の非発話区間があり，唇がほとんど動かないため，各発話の前後 30 フレームは主成分分析には用いていない．

HMM を用いて学習および認識を行うために，これらの固有画像に対する各フレームの口唇領域画像の主成分得点を認識特徴量とした．今後，音声情報を用いたバイモーダル音声認識を実現するため，特徴量のフレームレートが音声と同期していることが望ましい．AURORA-2J 標準スクリプトでは，音声特徴量について 10msec のフレームシフトと規定している．映像は 29.97fps（30000/1001）であるため，画像特徴量の主成分得点が 10msec 間隔になるように，1001/300 倍のオーバサンプリングを行った．

上記の特徴量を用いて HMM の学習を行った．パラメータ種別は USER_D_A（ユーザ定義データとその 1 次および 2 次差分値）とし，次元数は 14×3 の 42 次元である．発話数字の HMM は 5，10，18，20，22，25 状態と変えながら学習した．なお，挿入ペナルティは，挿入誤りと削除誤りが等しくなる値を事後的に与えた．それ以外の設定はすべて AURORA-2J の標準スクリプトの仕様と同一である．

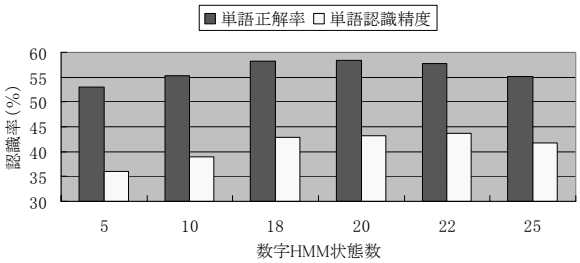

図 6　数字 HMM 状態数と認識率

表 9　HMM 認識結果

数字 HMM 状態数	18	20
正解単語数	1482	1485
削除誤り単語数	388	394
置換誤り単語数	675	666
挿入誤り単語数	391	387
全単語数	2545	2545
単語正解率	58.23%	58.35%
単語認識精度	42.87%	43.14%

認識率の評価には単語正解率と単語認識精度を用いた．単語正解率は正解単語数／全単語数，単語認識精度は（正解単語数－挿入誤り単語数）／全単語数と定義する（ただし，全単語数＝正解単語数＋置換誤り単語数＋削除誤り単語数）. AUROAR-2J-AV の評価セット 20 人（男女 10 人ずつ）について認識を行った結果を図 6と表 9に示す．AURORA-2J 標準スクリプトと同じ数字 HMM 状態数 18 では単語正解率 58.23%, 単語認識精度 42.87%という結果を得た. 状態数 20 では単語正解率 58.35%，単語認識精度 43.14%と最適であった．

5.2 考察

今回の実験では AURORA-2J-AV の室内映像を用いたため，認識特徴量が主成分得点のみでも比較的良好な結果を得られた．しかし，AURORA-3J-AV の車内映像は照明変化や顔向きの変化が激しいため，よりロバストな認識特徴量が必要になる．我々は車内環境における発話区間認識のために口唇領域画像内の低輝度画素数の変化などを特徴量とする試みを行っている[4]．今回は予備実験として口唇画像のみを用いた音声認識の実験を行った．今後，画像にも照明変化などの人工雑音を付加し，雑音を重畳した音声と統合したバイモーダル音声認識のベースラインの評価実験も行う予定である．

6. むすび

本稿では，雑音環境下音声認識評価用共通データベース AURORA-2J，AURORA-3J にカラー映像と近赤外映像を加えて収録を行った，AURORA-2J-AV，AURORA-3J-AV データベースについて述べた．本データベースは音声と映像が同期した状態て収録されており，映像と音声を統合した音声認識手法の開発，評価が可能になる．

今後，さらなる被験者数の増加，データの整理，発話区間や発話内容などの正解情報の付与，標準音声認識スクリプトなどの整備を行い，本データベースを公開する予定である. また, 英語版データベースの収録，音声と映像に加え，マイクロホンアレイを用いた音声データや, 車外の映像, アクセル開度やブレーキ踏力, GPS 位置情報などを含めたマルチモーダルデータベースを作成し公開する予定である．

謝　辞

収録を進めるにあたり，熱心に御討論頂いた名古屋大学大学院情報科学研究科メディア科学専攻末永研究室，武田研究室諸氏に感謝する．なお，本研究の一部は文部科学省科学研究費補助金, 21 世紀 COE プログラム「社会情報基盤のための音声映像の知的統合」および，文部科学省科学研究費補助金 COE 形成基礎研究費（課題番号 11CE2005）の援助を受けて行われた．

文　献

[1] 北岡教英, 赤堀一郎, 中川聖一, “スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識,” 電子情報通信学会論文誌, Vol.J83-D-II, No.2, pp.500-508, 2000.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior, “Recent Advances in the Automatic Recognition of Audiovisual Speech,” Proc. IEEE, vol.91, no.9, pp.1306-1326, 2003.

[3] 村井和昌，野間啓介，熊谷建一，松井知子，中村哲，“口周囲画像による頑強な発話検出,” 情報処理学会研究報告，2000-SLP-34-13，2000

[4] 前野俊希，根木大輔，北坂孝幸，森健策，末永康仁, 二宮芳樹，“車載カメラ映像を用いたドライバの発話区間検出の改善,”電子情報通信学会技術報告，PRMU2003-257，2004.

[5] 坂義秀, 前野俊希，二宮芳樹, 森健策, 末永康仁, “車載カメラ映像を用いたドライバの発話区間検出,” 電子情報通信学会技術報告, PRMU2002-03, 2003.

[6] H. G. Hirsh and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” ISCA ITRW ASR2000, 2000.

[7] 山本一公，中村哲，武田一哉，黒岩眞吾，北岡教英，山田武志，水町光徳，西浦敬信，藤本雅清，“AURORA-2J/AURORA-3J データベースとその評価ベースライン,” 情報処理学会研究報告，2003-SLP-47-19，2003.

[8] G. Iyengar and C. Neti, “Detection of faces under shadows and lighting variations,” Proc. Workshop Multimedia Signal Processing, pp.15-20, 2001.

[9] S. Young, The HTK Book(for HTK Version 3.2.1), http://htk.eng.cam.ac.uk/, 2002

[10] 河口信夫，松原茂樹，山口由紀子，武田一哉，板倉文忠，“CIAIR 実走行車内データベース,” 情報処理学会研究報告，2003-SLP-49-24，2003.

[11] 河口信夫，松原茂樹，岩博之，梶田将司，武田一哉, 板倉文忠，“実走行車内における音声データベースの構築,” 情報処理学会研究報告，2000-SLP-30-12，2000.