社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64（1）
2006／12／21

雑音下音声認識評価ワーキンググループ活動報告：認識に影響する要因の個別評価環境

北岡 教英[1] 山田 武志[2] 滝口 哲也[3] 柘植 覚[4] 山本 一公[5]
宮島千代美[1] 西浦 敬信[6] 中山 雅人[6] 傳田 遊亀[6] 藤本 雅清[7]
田村 哲嗣[8] 黒岩 眞吾[4] 武田 一哉[1] 中村 哲[9]

[1] 名古屋大学 [2] 筑波大学 [3] 神戸大学 [4] 徳島大学 [5] 信州大学
[6] 立命館大学 [7] NTT CS 基礎研 [8] 岐阜大学 [9] NiCT/ATR
E-mail: kitaoka@slp.ics.tut.ac.jp

あらまし 雑音下の音声認識の性能向上は音声認識実用化のために急務である．これまでに数多くの研究が行われてきており，これらの手法を客観的に比較評価できる標準評価基盤の構築を目的として，2001 年 10 月，情報処理学会音声言語情報処理研究会の下に 雑音下音声認識評価ワーキンググループを組織した．本稿ではこれまでの標準評価基盤 CENSREC シリーズを振り返り，今年度新たに配付した CENSREC-1-C の概要と位置づけを述べる．さらに，今後どのような方針で新たな評価基盤を設計・構築・配付するのかについての考えを述べる．
キーワード 音声認識，雑音，評価基盤，CENSREC

Progress Report of SLP Noisy Speech Recognition Evaluation WG: Individual evaluation framework for each factor affecting recognition performance

Norihide KITAOKA[1], Takeshi YAMADA[2], Tetsuya TAKIGUCHI[3], Satoru TSUGE[4],
Kazumasa YAMAMOTO[5], Chiyomi MIYAJIMA[1], Takanobu NISHIURA[6],
Masato NAKAYAMA[6], Yuki DENDA[6], Masakiyo FUJIMOTO[7], Satoshi TAMURA[8],
Shingo KUROIWA[4], Kazuya TAKEDA[1], and Satoshi NAKAMURA[9]

[1] Nagoya Univ. [2] Univ. of Tsukuba [3] Kobe Univ. [4] Univ. of Tokushima [5] Shinshu Univ.
[6] Ritsumeikan Univ. [7] NTT CS research Lab. [8] Gifu Univ. [9] NiCT/ATR
E-mail: kitaoka@slp.ics.tut.ac.jp

Abstract Performance improvement of noisy speech recognition is urgent for practical use of speech recognition. Many researchers have been tryin to overcome this problem. We organized a working group under Special Interest Group of Spoken Language Processing in Information Processing Society of Japan, to develop evaluation frameworks of noisy speech recognition to compare many methods for processing of noisy speech. In this paper, we first review the series of CENSREC (Corppus and Environment of Noisy Speech RECognition) and then introduce the CENSREC-1-C, the newest CENSREC. Finally we describe the road-maps of future CENSRECs.
Key words Speech recognition, noise, evaluation frameworks, CENSREC.

1. はじめに

音声認識の性能が飛躍的に向上し実用化が進められている中，実環境のさまざまな雑音下で頑健に動作するシステムはいまだ実現されていない．雑音下の音声認識の研究は多くなされてきており，個々の実験条件の下では十分な性能を得ることができ

ているものも多い．そこで，これらの多くの手法を客観的に比較評価できれば，例えば開発者らが意中の目的にふさわしい手法を見付けることが可能となると考えられる．

これまでに，こうした問題に対して欧州の AURORA 研究プロジェクトでは，雑音抑圧手法の標準評価コーパスと HTK を利用した標準認識スクリプト，および標準評価用 Excel シートをセットとした評価基盤を配布した．これらは，TI digit に雑音を付与した連続数字音声認識タスクの AURORA-2，自動車内で実際に発声した連続数字認識タスクの AURORA-3，Wall Street Journal タスクをベースとした雑音下大語彙連続音声認識タスクの AURORA-4 からなる．[1], [2]．この中でも AURORA-2 はコンパクトで認識評価が容易なことから国際会議論文でもよく利用されている．

我々も上記の活動に追従するように 2001 年 10 月に雑音下音声認識評価ワーキンググループを組織し，日本語における同様の評価基盤の構築を開始した．さらにそれを拡充し，雑音下音声認識手法をさまざまな側面から評価できる評価基盤群を構築すべく活動してきた [3], [4]．その一環として，2006 年 9 月に雑音下音声区間検出評価基盤である CENSREC-1-C の配布を開始した．

本稿では，これまでのワーキンググループの活動を振り返るとともに，基盤群構築のための考え方を述べ，今後の展開について述べる．

2. 情報処理学会 音声言語情報処理研究会 雑音下音声認識評価ワーキンググループ

本ワーキンググループ (WG) では，雑音下音声認識の標準評価基盤として，どのようなものを提供すればよいかを課題として議論を重ねてきた．そして，AURORA プロジェクトの配布形態にならい，共通学習データ/テストデータ，標準学習/認識スクリプトおよび評価用 Excel シートを含むいくつかの評価基盤を配布している．これらは CENSREC(Corpus and Environment for Noisy Speech RECognition) と呼び，CENSREC-1,2,3 の作成・配布を行ってきた．

CENSREC-1 は AURORA-2 の日本語版であり，連続数字発声に種々の雑音を付加したデータによる実験環境である．しかし一般に，雑音付加によるシミュレーション環境は実際の雑音環境をある程度模擬しているものの，実環境下ではそれ以外の要因も作用していると考えられている．CENSREC-2 では，この違いを検討できるように，自動車内の実環境下で同様の内容を実際に発声して収録したデータを用いている．さらに，タスクによる影響も考えられるため，その検討のために，CENSREC-2 を孤立単語認識タスクに変更したものが CENSREC-3 である．すなわち，雑音下音声認識の困難さを「雑音の違い」「シミュレーション/実環境の違い」および「タスクの違い」の軸上で比較できる環境群となっている．以下に，それぞれの説明を行う．

2.1 CENSREC-1/AURORA-2J

欧州 AURORA プロジェクトの作成した雑音環境下連続英語数字音声認識タスクの共通評価フレームワークである AURORA-2 の日本語版である．2003 年 7 月の配布開始から，100 部を越える部数を配布し，各種研究発表で利用されている．

表 1 CENSREC-1 の雑音環境

	加法性雑音	フィルタ特性
Set A	Subway, Babble, Car, Exhibition	G.712
Set B	Restaurant, Street, Airport, Station	G.712
Set C	Subway, Street	MIRS

学習/テストデータとともに学習認識スクリプトとベースライン性能を提供し手軽に実験を開始できる．タスクが連続数字認識とシンプルで，加算性雑音の影響とその対策に絞ってあり，逆にこれからの発展がさまざま考えられるという意味で，本 WG の評価基盤の原点であるといえる．

2.1.1 データ構成

1 から 7 桁の連続数字であり，話者数，男女構成などは AURORA-2 と同一で，発声リストは同じものを日本語に翻訳して使用した．10 種類の数字のうち，英語の" ゼロ" と" オー" に対応して 0 のみが "Z(ぜろ)"，"O(まる)" の 2 種類がある．

学習データは，clean training（クリーン音声によるモデル学習），multicondition training（雑音重畳音声による学習）用にそれぞれ 110 名，8,440 発話（男女 55 名，4,220 発話ずつ）用意されている．Multicondition training の場合にのみ 4 種類の雑音（Subway, Babble, Car, Exhibition）を 5 種類の SNR レベル（clean, 20dB, 15dB, 10dB, 5dB）で重畳した音声（各雑音・SNR で 422 発話ずつの学習データとなる）を用いて学習を行なう．チャネルフィルタとして G.712 の中で規定されているフィルタを用いている．

テストデータは男女 104 名による 4,004 発話を 4 セットに等分した 1001 発話に表 1 のいずれかの雑音を 7 種類の SNR レベル（clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB）で重畳したものである．

2.1.2 評価方法

HTK を用いた HMM の学習および認識スクリプトが添付されている．HMM は 10 数字 (11 種類) の単語モデルに無音 (sil, sp) を加えた 13 モデルで，数字 HMM は 18 状態 (2 状態はダミー)，出力確率は対角正規分布の 20 混合（無音モデルは 36 混合）である．特徴パラメータは，MFCC 12 次元+Δ+ΔΔ + 対数パワー（1 次元）+Δ パワー+ΔΔ パワーの計 39 次元である．

この認識スクリプトを動作させることによりベースライン性能が得られる．これは既に入力された Excel シートが用意されており，独自の手法で雑音対策した結果を入力することによりベースライン性能からの相対的な性能改善が計算される．

詳細は文献 [5] を参照されたい．

2.2 CENSREC-2

CENSREC-1 と同様の連続数字データを自動車内で発話したデータに基づいており，CENSREC-1 の雑音加算シミュレーションに対し実収録データへの効果を確認できる．本評価環境は，2005 年 12 月より配布を開始している．

2.2.1 収録データ

名古屋大学の実験車両を用いて収録を行った．収録時には 5

本 (接話 1 本，ダッシュボード上 2 本，天井 3 本) のマイクロフォン (SONY ECM77B) を用いており，CENSREC-2 では，接話マイクロホンと天井の 1 本 (遠隔マイクロホンと呼ぶ) で収録された音声を用いた [10]．収録条件は，3 種類の走行速度 (アイドリング，低速 (市街地) 走行，高速走行) と 4 種類の車内環境，エアコン On，オーディオ On，窓あけ) を組み合わせた 11 種類の環境である．総発話者数は 104 名 (学習データ：73 名，評価データ：31 名) で，収録音声の総数は 17,651 発話である．

2.2.2 評価環境の設計

前節のような種々の条件で収録された音声データを組み合わせ，そのときの学習データと評価データの収録条件 (マイク種別，収録環境) の一致度合により，以下の 4 種類の音声認識評価条件 (Condition 1〜4) が設定されている．

Condition 1: マイク種別，収録環境がともに一致
Condition 2: マイク種別が一致，収録環境が相違
Condition 3: マイク種別が相違，収録環境が一致
Condition 4: マイク種別，収録環境がともに相違

CENSREC-2 に関する詳細は，文献 [6] を参照されたい．

2.3 CENSREC-3

CENSREC-3 は，CENSREC-1 に続いて，実音響環境での発話を対象として設計されたものであり，実走行車内での孤立単語音声認識の評価環境を提供する．このデータおよび評価環境は 2005 年 2 月より配布を行っている

2.3.1 収録データ

収録環境は CENSREC-2 と同じで，接話マイクロホンと遠隔マイクロホンの 2 種類を用いて 3 種類の走行速度 (アイドリング，低速 (市街地) 走行，高速走行) と，6 種類の車内環境 (通常走行，ハザード On，エアコン (Low)，エアコン (High)，オーディオ On，窓開) を組み合わせた 16 種類の環境 [11] である．内容は，カーナビゲーション等で使用されることを想定した 50 個のコマンドワードである．これらの 18 名 (男性 8 名，女性 10 名) による 14,216 発話が収録されている

2.3.2 評価環境の設計

学習データと評価データの環境の組み合わせにより，欧州 AURORA プロジェクトの AURORA-3 の Well-matched condition(WM)，Moderate-mismatched condition(MM)，High-mismatched condition(HM) に準じた 6 種類の評価環境 (Condition 1〜6) を設定しており，以下のような対応となっている．

Condition 1, 2, 3 マイク種別，収録環境がともに一致 (WM)
Condition 4 マイク種別が一致，収録環境が相違 (MM)
Condition 5, 6 マイク種別，収録環境 (の一部) が相違 (HM)

これらの学習環境条件の元で収録された音素バランス文を用いて学習した Word Internal Triphone HMM で認識を行うスクリプトと，ベースラインからの性能向上が評価できる Excel シートを用意している．また評価時のバックエンド部分の変更レベルにより他手法との比較が可能となるように評価カテゴリを設定している．

詳細は，文献 [7] を参照されたい．

3. 雑音下音声区間検出評価基盤 CENSREC-1-C [8]

これまでは認識対象データに含まれる種々の要因の音声認識性能に及ぼす影響を比較検討できるという観点で基盤群を構築してきた．しかし実環境での利用の状況をより広範囲に鑑みて，それ以外の評価すべき要因が存在することがわかってきた．

そのひとつに音声区間検出手法 (Voice Activity Detection; VAD) がある．近年，音声認識の性能向上を目的として雑音下の VAD の研究が盛んである．正確な VAD は音声区間のみを音声認識することを可能にし，非音声区間からの湧き出し誤りや音声区間からの脱落誤りを減少させて結果的に音声認識性能を向上させることができる．また，VAD は音声強調や音声符号化などの音声処理においてもその精度向上に大きな役割を果たす．ここでは雑音下の VAD 評価環境として構築した **CENSREC-1-C** (CENSREC-1-Concatenated) について述べる．雑音下連続数字発声データと評価ツールからなり，VAD を行った結果を 2 種類の評価基準に基づいて評価できるものである．

3.1 データ構成

連続数字を間隔をあけて発声したもので，個々の発声は CENSREC-1 [5] に準じている．雑音加算によるシミュレーションデータと実際の雑音環境下で発声された実環境データでの評価の両方が行えるよう，以下のデータ群を，目視により作成された音声区間の始端・終端データとともに用意した．

3.1.1 雑音加算によるシミュレーションデータ

CENSREC-1 [5] の音声データを複数接続することにより，連続的な発話のデータを作成している．CENSREC-1 の 1 つの雑音環境で同一話者により発話されている 9 ないし 10 の発話を 1 秒の無音を挟んで接続した．CENSREC-1 の 1 つの雑音環境での話者数は 104 名 (男女各 52 名) であり，話者毎に接続データを作成するため，1 環境におけるデータ数は 104 となる．本評価環境における雑音環境は CENSREC-1 の Set A, B と同様である (表 1 参照)．ただし，Set C は提供しない．

3.2 実環境データ

実環境データの収録は，2 つの雑音環境 (学生食堂，高速道路付近) および 2 つの SNR 環境 (低 SNR，高 SNR) にて行った．マイクロホンは近接位置と遠隔位置を同期収録した (近接マイクのデータは評価対象ではない)．被験者は，男女 5 名，合わせて 10 名 (被験者の年齢内訳は，20 歳前後男女各 3 名，30 歳前後男女各 1 名，40 歳以上男女各 1 名) とした．1 名につき各雑音環境および各 SNR 環境につき 1〜12 桁の連続数字を 8〜10 回，約 2 秒間の間隔で発声した音声を 1 つのファイルとして，計 4 ファイル (総発話数：38〜39 発話) である (ただし 1 名分は意図通りに発話できていないデータとして評価から除外)．

シミュレーションデータおよび実環境データの内容を表 2 にまとめる．

表 2　CENSREC-1-C 収録データの全容

	シミュレーションデータ	実環境データ
各ファイル中の発声	CENSREC-1 の 9〜10 発話を無音を挟んで接続	連続数字を約 2 秒間隔で 8〜10 発話
雑音	CENSREC-1 と同じ雑音を −5〜20dB(5dB 刻み) で加算，またはクリーン	学生食堂 (低 SNR 69.7dBA，高 SNR 53.4dBA) および高速道路付近 (低 SNR 69.2dBA，高 SNR 58.4dBA)
データ量	各条件 104 ファイル，全 5824 ファイル	各条件 4 ファイル，全 160 ファイル（評価対象は 144 ファイル）

3.3 評価尺度

3.3.1 フレーム単位の評価尺度

音声区間検出の性能を測るためのフレーム単位の評価尺度として，FRR (False Rejection Rate) と FAR (False Acceptance Rate) を用いる．

$$\mathrm{FRR} = \frac{N_{FR}}{N_s} \times 100\ [\%], \qquad \mathrm{FAR} = \frac{N_{FA}}{N_{ns}} \times 100\ [\%]$$

ここで，N_s は音声フレーム数，N_{FR} は音声を非音声と検出したフレーム数，N_{ns} は非音声フレーム数，N_{FA} は非音声を音声と検出したフレーム数である．複数のデータを対象とする場合には，データ毎に FRR と FAR を求め，その平均値を用いて評価する．手法によってフレーム長は異なるので，サンプル単位の正解データをフレーム長に換算して評価する．

一般に FRR と FAR はトレードオフの関係にあり，どちらの性能を重視するかは対象とするアプリケーションによって決まる．そこで，閾値によって FRR と FAR を調整することで，ROC 曲線 (x 軸：FAR，y 軸：100 − FRR) を示すことを推奨している．以下の 2 つの ROC 曲線を描くための Excel シートを用意している．

(1) SNR 別 ROC 曲線：雑音レベルによる性能変化の評価
(2) 雑音別 ROC 曲線：雑音種類による性能変化の評価

3.3.2 発話単位の評価尺度

音声認識のための音声区間検出は，一般に発話単位 (単語や文など) で行う．発話単位の音声区間検出性能を評価する尺度には，発話区間検出正解率 Corr と発話区間検出正解精度 Acc を用いる．

$$\mathrm{Corr} = \frac{N_c}{N} \times 100\ [\%], \qquad \mathrm{Acc} = \frac{N_c - N_f}{N} \times 100\ [\%]$$

ここで，N は総発話区間数，N_c は正解発話区間検出数，N_f は誤発話区間検出数である．Corr は，発話区間をどれだけ多く検出できるかを評価する尺度であるのに対し，Acc は，発話区間をどれだけ過不足なく検出できるかを評価する尺度である．複数のデータを対象とする場合には，データ毎に Corr と Acc を求め，その平均値を用いて評価する．

音声認識では，発話区間を短かめに検出すると認識誤りを起こしやすいが，前後にやや長めに検出しても認識結果への影響は比較的少ない．そこで，検出区間の中に，ある 1 つの発話区間全体が含まれ，その前後の発話区間と重なりがなければ，正解検出とし，その他の検出区間は全て誤検出とする．

3.4 ベースライン評価

音声パワーに基づくベースライン VAD を設定し，その結果を公開している．雑音区間と音声区間のフレームごとに音声/非音声の 2 クラスに分類する分類問題と考えて，クラス分類に最適な閾値を判別基準

$$\eta(s) = \frac{\sigma_B^2(s)}{\sigma_T^2}, \qquad \sigma_T^2 = \sigma_W^2(s) + \sigma_B^2(s)$$

を最大とするしきい値 (s) を求めることにより決定する方法である [9]．ここで，$\sigma_W^2(s)$ は平均クラス内分散，$\sigma_B^2(s)$ は平均クラス間分散を示す．この手法で求めた最適なしきい値を初期しきい値 (THR_{ini}) とし，以下の式で $k = -40, -39, \ldots, 39, 40$ の 81 通りに変動させて評価した．

$$\mathrm{THR} = \mathrm{THR}_{ini} + k \cdot \alpha, \qquad \alpha = \frac{(\mathrm{POW}_h - \mathrm{POW}_l)}{K}$$

ここで，POW_l は初期しきい値未満のフレームの平均パワー，POW_h は初期しきい値以上のフレームの平均パワーを示す．

フレームのパワーがしきい値以上となった場合，当該フレームを音声開始フレーム候補とし，次の音声終了フレーム検出を行う．音声開始フレームの候補が決定された後，フレームのパワーがしきい値未満となりかつそのフレーム以降のある一定区間 (ベースライン実験では 500ms) 以上のフレームのパワーがしきい値未満の場合，音声終了フレームとする．ただし，一定区間以上連続してフレームのパワーがしきい値未満でない場合は棄却し，再度音声終了フレームを検出する．さらに，音声終了フレームが確定しても，音声区間が一定区間未満の場合にはその音声区間は棄却する．

フレームベース評価における SNR 別，雑音別の ROC 曲線を図 1 に示す．また，発話単位の性能評価結果を表 3 に示す．これは，最も発話区間検出正解率 (Corr) が高かったしきい値の結果である．ここでの評価では，上記手法で得られた発話区間を前後に 300ms 延長している．

シミュレーションデータに対しても同様の評価が行える．利用者は自身の手法で定められたフォーマットで VAD 結果を出力すれば容易に評価結果が得られ，ベースラインと比較できる．

詳細は文献 [8] を参照されたい．

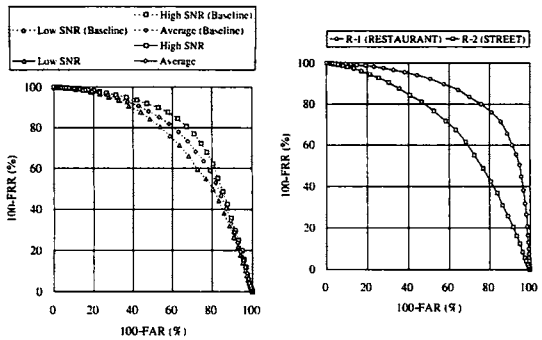
表 3　ベースラインの実環境データに対する発話単位の評価結果 (上: 正解率，下: 精度)

Correct Rate [%]

Real Data		Remote Microphone		
		Restaurant	Street	Average
Correct Rate [%]	High SNR	74.20	39.42	56.81
	Low SNR	56.52	41.45	48.99
	Average	65.36	40.44	52.90

Accuracy [%]

Real Data		Remote Microphone		
		Restaurant	Street	Average
Accuracy [%]	High SNR	21.45	-15.65	2.90
	Low SNR	-43.48	-33.91	-38.70
	Average	-11.02	-24.78	-17.90



(a) SNR 別　　(b) 雑音別

図 1　ベースラインの実環境データに対する VAD の ROC 曲線

4. 今後の展開

さきにも述べた通り，これまで認識対象データに含まれる種々の要因の音声認識性能に及ぼす影響を比較検討できる基盤群を構築してきた．ここでは，それを含めより広い観点から評価すべき要因を挙げ，今後どのような評価基盤を構築するかを考える．

4.1 環　　境

まず，認識対象データのバリエーションについて考える．

4.1.1 加法性雑音

これまで，実在する雑音の種類を集めて重畳したり，自動車の走行条件を変えるなどしてきた．これは実在の雑音のカバレージを向上させるという目的がある．しかし，手当たり次第に集めることで十分なカバレージを得ることは容易ではないし，得たとしても膨大な種類の雑音での評価が必要となってしまう．

そこで，音声認識の性能への影響という視点から，雑音の種類を客観的に評価・分類し，雑音の種類を効率的にサンプリングする方法を探求していきたいと考えている．例えば，WG メンバーにより音声の歪尺度 PESQ [12] と擬似音声 [13] を用いた簡便な認識率推定手法を提案されている [14]．また，雑音はよく SNR や非定常性で分類されるが，同じ SNR でもスペクトルが異なれば音声への影響は異なり，非定常性にもさまざまある．平均スペクトルから非定常性，突発性などを複合的に用いて雑音の姿を記述できる量に関しても議論していく．そして，COSMOS マップ [15] のように俯瞰的・可視的にみることで雑音収集の方向が見えてくるものと考えている．

4.1.2 残　　響

これまで実環境の大きな要素として注目され研究もされながら，音声認識の評価基盤として確立されていないのが残響である．残響もインパルス応答が測定されていれば畳み込みにより残響環境下音声がシミュレーション可能である．しかし，加算性の場合と同様にシミュレーションと実環境下の差異が問題となる．そこで，現在 CENSREC-1 と同様な畳み込みによるシミュレーション残響音声と，実際に残響の存在する場所において収録した音声とを含んだ残響音声認識評価基盤を設計している．

しかし残響も，加算性雑音と同様にそのカバレージをいかに確保するかの問題が生じる．これに対して，メンバーにより，初期残響と音声認識率の関係を明らかにするために第 1 次反射音の遅れ時間と音声認識率の関係についての調査 [16] が行われており，残響が一概に音声認識率の低下を引き起こすばかりではないことが確認されている．このように残響を記述する量を探求する研究結果を基盤構築に生かしていく．

4.2 雑音下音声認識のためのアプローチ

これまで雑音下の音声認識へのアプローチの一般的な分類として，(1) 雑音に強い特徴量，(2) 雑音の除去・抑圧，(3) 音響モデル・適応があった．これらはいずれも与えられた音声があり，それを HMM とマッチングする際のミスマッチを軽減するための方法であり，これまでに提供してきた CENSREC シリーズで評価が可能であった．

しかし最近，より異なるアプローチも多数提案されてきている．そのひとつに，音声区間検出 (VAD) がある．音声と HMM とのミスマッチ軽減という直接的な方法ではなく，不要なマッチングを除外し，必要十分な部分のみを認識することで，大幅な湧き出し誤りと脱落誤りの削減効果がある．この方法は認識そのものによる対処法と独立に開発し，組み合わせることができる．このアプローチの独立した評価基盤として CENSREC-1-C を配布している．

また，フロントエンドの処理としてマルチマイクロフォンを用いる手法が数多く提案され，遅延和アレイによる SN の向上や適応的に特定の到来方向の雑音を抑圧する方法，さらに空間的に音声を除く音成分を取り出して減算することによる空間的スペクトルサブトラクション [17] などが効果的であると言われている．このように，(比較的少数の) 複数マイクロフォンを用いた手法はいくつも提案されてきており，各手法の能力を客観的に比較できる基盤は必要であろう．ただし，マイクロフォンの配置にはそれぞれの手法や研究者の意図によってさまざまな工夫が施されており，共通に利用可能な収録条件が設定できるかどうか難しい面がある．今後，慎重に議論していきたい．

さらに，認識段階における新たなアプローチも見られる．音響的な雑音に強い映像情報を音声とハイブリッドに用いて頑健に音声認識を実現するバイモーダル手法も数多く研究されている．そして，大語彙連続音声認識アルゴリズムにおける耐雑音性の向上も考えられる．まだ研究例は少ないが，雑音レベルに応じて音響尤度と言語尤度の重みを動的に変化させて，高 SNR 時には音響モデルに，低 SNR 時には言語モデルに頼るなどデコーディングの工夫による研究例もある (例えば [18])．英語の AURORA-4 は大語彙連続音声認識タスクにおける評価基盤であり，大規模さのために動作に時間がかかるなど問題があり現在のところあまり利用されているとは言えない．しかしこのようなアプローチにも十分な効果があることが確認されてくれば必要とされよう．研究動向を注視しながら議論を重ねていく．(注1)

このように雑音に対するアプローチも多様化している．そ

(注1)：さらに上位のアプローチとしてアプリケーションレベルもありえよう．すなわち対話的に誤認識を解消したり他の入力デバイスとの併用により認識結果の複数候補から正解を選択するなどがある．しかしこれらは本 WG の対象外としたい．

れに見合った評価基盤の提供が必要である．次節では，その一例として顔映像と音声のバイモーダル情報による認識のための評価基盤として配付準備を進めている CENSREC-1-AV, CENSREC-2-AV を紹介する．

4.3 バイモーダル音声認識評価用基盤 CENSREC-1-AV, CENSREC-2-AV

4.3.1 概　　要

CENSREC-1-AV は，CENSREC-1 と同様に，室内で収録したデータに事後的に雑音を重畳したデータである．音声には，CENSREC-1 と同じ雑音を，映像には車内でダッシュボードを撮影した映像から得た γ 値の変化を加えている．発話内容も CENSREC-1 と同じ連続数字単語であり，約 90 名のデータの公開を予定している．

一方，CENSREC-2-AV は，CENSREC-2 と同様に，実車内でドライバの音声と映像を収録したデータである．名古屋大学のデータ収録車を用いて，アイドリング・走行中に収録した約 50 名のドライバの音声と顔映像を収録している．近年，安全運転支援のための居眠りや脇見の検知，防犯用の顔撮影のためのカメラが実用化され始めており，今後車内カメラの普及が予想され，バイモーダル音声認識にも利用できると期待できる．

音声は 2 つのマイク，顔映像はカラーカメラと可視光カットフィルタ装着した近赤外カメラで収録しており，メディアコンバータでステレオ音声とそれぞれの映像を統合している．[19]

4.3.2 配 布 方 法

バイモーダル音声データベースの映像に関しては，それぞれ 2 つの形式で公開する予定である．一つは，顔全体を含む収録した映像のデータそのものであり，もう一つは，顔映像から口付近のみを抽出した口唇映像のデータである．口唇映像は，口唇位置検出アルゴリズムや追跡アルゴリズムを用いて口唇部分を抽出し，各フレームに分割した画像を，音声・評価用スクリプト等と併せて DVD に収めて配布する．顔映像のデータは容量が大きいため，HDD での配布を予定している．音声と顔映像は，DV フォーマットの AVI ファイルとして保存しているが，AVI ファイルから映像をステレオ音声と映像に分割するためのソフトウェア等も併せて配布する予定である．

音声は MFCC，映像は主成分スコアを特徴量とした認識結果を，比較用の認識結果として掲載する．音声と映像を統合するためのストリーム重みは，学習時は等重みで固定し，認識時にのみ重みの割合を変化させて，各 SNR において認識精度が最も高くなる重みを選択できるようなスクリプトを配布する．

評価用スクリプトや認識結果を集計するための Excel シートの形式については，CENSREC-1, CENSREC-2 に準拠する．これらの他に，映像付加による認識性能の改善率を計算するための Excel シートも追加する．

5. ま と め

本稿では，雑音下音声認識評価ワーキンググループのこれまでの標準評価基盤 CENSREC シリーズを振り返り，今年度新たに構築した CENSREC-1-C の概要と，これまでおよびこれからの CENSREC シリーズ内での位置づけを述べた．今後は，CENSREC-1,2,3 と同様に認識対象データの雑音環境の幅を広げていくとともに，CENSREC-1-C のように雑音下音声認識に対する音声認識そのものだけではない，違ったアプローチに対しても個別評価・比較検討できる環境を提案していく．

文　　献

[1] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next." Proc. Eurospeech 2001, 2001.

[2] Aurora document no. AU/345/01, "Large vocabulary evaluation of front-ends- baseline recognition system description," Mississipppi State University. Aug. 2001.

[3] 中村 哲, 武田一哉, 黒岩眞吾, 山田武志, 北岡教英, 山本一公, 西浦敬信, 藤本雅清, 水町光徳, "SLP 雑音下音声認識評価ワーキンググループ活動報告," 情報処理学会研究報告, 2002-SLP-42-11, pp.65-70, July 2002.

[4] 中村 哲, 武田一哉, 黒岩眞吾, 山田武志, 北岡教英, 山本一公, 西浦敬信, 藤本雅清, 水町光徳, "SLP 雑音下音声認識評価のための WG: 評価データ収集について," 情報処理学会研究報告, 2002-SLP-45-9, pp.51-56, Feb. 2003.

[5] S. Nakamura *et al.*, "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition." *IEICE Transactions on Information and Systems*, Vol. E88-D. No. 3, pp. 535-544. 2005.

[6] 藤本 雅清, 武田 一哉, 中村 哲, "CENSREC-2: 実走行車内における連続数字音声データベースと評価環境の構築," 情報処理学会研究報告, SLP-60-3, pp. 13-18, 2006.

[7] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-3: An Evaluation Framework for Japanese Speech Recognition in Real Driving-Car Environments," *IEICE Transactions on Information and Systems*, (accepted).

[8] 北岡教英ら "CENSREC-1-C：雑音下音声区間検出手法評価基盤の構築," 情報処理学会研究報告, 2006-SLP-63-1, pp. 1-6, 2006.

[9] 大津展之, "判別および最小 2 乗基準に基づく自動しきい値選定法," 信学論, Vol.J63-D, No. 4, pp. 349-356, 1980.

[10] K.Takeda et al., "Construction and Evaluation a Large In-Car Speech Corpus," IEICE Transactions on Information and Systems, Vol.E88-D, No.3, Mar. 2005.

[11] 武田 一哉 他 "走行状況別車内音声データベースとその予備評価," 音講論集, 3-P-10, pp. 185-186, Mar. 2002.

[12] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." Feb. 2001.

[13] ITU-T Recommendation P.50, "Artificial voices," Sept. 1999.

[14] T. Yamada, M. Kumakura, N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 2006-2013, 2006.

[15] 庄境 誠, "複数音声コーパスの俯瞰的分析," 音声言語シンポジウム, 情報処理学会研究報告, 2005-SLP-59-23, 2005.

[16] 西浦敬信, 傳田遊亀 "音声認識における初期反射音の影響についての検討". 日本音響学会 2006 年度春期全国大会講演論文集, 3-1-17, pp. 141-142, Mar. 2006.

[17] Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Noise reduction using spatial subtraction array based on independent component analysis", 電子情報通信学会技術研究報告, EA2006-22, vol.106, no.125, pp.13-18, June 2006.

[18] N. B. Yoma, F. R. McInnes, M. A. Jack, S. D. Stump, and L. L. Ling,"On including temporal constraints in viterbi alignment for speech recognition in noise," IEEE Transactions on Speech and Audio Processing, Vol. 9, pp. 179-182, 2001.

[19] 根本 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, "バイモーダル車内音声認識評価用データベースの構築", 情報処理学会研究報告, 2005-SLP-55-7, Feb. 2005.