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

2004－SLP－54（24）
2004／12／21

〔招待講演〕

実環境下音声認識の評価の標準化とその動向

中村　哲[1]　武田　一哉[2]　黒岩　眞吾[3]　北岡　教英[4]　山田　武志[5]
山本　一公[6]　西浦　敬信[7]　佐宗　晃[8]　水町　光徳[9]　宮島千代美[2]
藤本　雅清[1]　遠藤　俊樹[1]

[1] ATR 音声言語コミュニケーション研究所 [2] 名古屋大学 [3] 徳島大学 [4] 豊橋技術科学大学 [5] 筑波大学 [6] 信州大学 [7] 立命館大学 [8] 産業技術総合研究所 [9] 九州工業大学

あらまし　現在の音声認識は，実使用環境に存在する雑音などの外的要因により性能劣化を免れない．このため，これまで数々の研究が行われてきた．しかしながら，異なるタスク，異なる評価データが用いられてきたため性能の比較が非常に困難であった．このため，米国や欧州で種々のプロジェクトが企画された．本稿では，これらのプロジェクトと日本において著者らが進めている雑音下音声認識の評価フレームワーク構築の活動についての現状と今後の予定，狙いについて述べる．

キーワード　音声認識，評価フレームワーク，雑音，残響

Towards Standardization and Evaluation Framework for Noisy Speech Recognition

Satoshi NAKAMURA[1], Kazuya TAKEDA[2], Shingo KUROIWA[3], Norihide KITAOKA[4], Takeshi YAMADA[5], Kazumasa YAMAMOTO[6], Takanobu NISHIURA[7], Akira SASOU[8], Mitsunoti MIZUMACHI[9], Chiyomi MIYAJIMA[2], Masakiyo FUJIMOTO[1], and Toshiki ENDO[1]

[1] ATR Spoken Language Translation Research Laboratories [2] Nagoya University [3] University of Tokushima [4] Toyohashi University of Technology [5] University of Tsukuba [6] Shinshu University [7] Ritsumeikan University [8] National Institute of Advanced Industrial Science and Technology [9] Kyushu Institute of Technology

Abstract　Performance degradation by environmental interference such as noise and reverberation is inevitable for the current state of the art speech recognition. So far there have been many researches to overcome this problem. However, it has been very difficult to know actual improvements and compare those methods since those methods were developed for individual tasks and on different corpus. Recently, two projects have been organized in USA and Europe. This paper introduces those projects briefly, and also introduces current activities and a future road-map of a common standardized framework for noisy speech recognition organized by the authors.

Key words　Speech recognition, common evaluation framework, noise, reverberation

1. はじめに

現在の音声認識装置は，実際に利用されるような雑音のある環境で利用しようとすると，未だ大きな性能劣化があり，性能改善の余地があると言わざるを得ない．これまで，多くの研究がなされてきたものの，タスクやコーパスの違いにより比較が非常に困難な状況である[1]．現在，音声認識の主流になっているHMMやN-gramは，1980年代に開始された音声認識に関する米国国防総省のDARPAプロジェクトで開発された様々な技術に負うところが大きい．このプロジェクトでは，参加した研究チームに同一の学習データとテストデータを用いて同一のタスクに対する性能改善を競わせる．この手法により，不特定話者大語彙連続音声認識のシステム構成基本技術が確立された．

このような音声認識の音響環境に対する頑健性の問題に対しても，これまでに米国DARPA主催のSPINEプロジェクト[2]と，欧州におけるAURORAプロジェクト[3]の２つの研究プ

ロジェクトが進められた．特に，後者は標準化のためのコーパスとそれを認識するためのソフトウェア，評価のためのスクリプトを提供した点で非常に大きな貢献をしたといえる．

日本でも，2001 年の 10 月に情報処理学会音声言語情報処理研究会の中に，ワーキンググループ（以下，IPSJ-SLP-NOISEWG）を作り，雑音下日本語音声認識の評価のための議論を進めてきた [4], [5]．このワーキンググループの目的は，雑音下音声認識の要素技術のアセスメントのための計画，標準コーパスの構築，共通評価手法の開発，標準パッケージの配布である．

本稿では，欧州の AURORA グループの配布コーパスとタスク，現在の動向について若干ふれ，主として IPSJ-SLP-NOISEWG の活動，ロードマップについて述べる．以下，第 2 章において，欧州の ETSI AURORA WG の活動について述べ，第 3 章において IPSJ-SLP-NOISEWG の活動について述べる．

2. ETSI-AURORA WG

欧州の通信技術標準化団体 ETSI 傘下の AURORA グループが，分散音声認識のための雑音下音声認識の前処理に関するスペシャルセッションを Eurospeech, ICSLP で開催した [6]．このグループは ETSI のもと，標準化に向けて技術標準化を行っていたが，これに並行して，さらに雑音下音声認識の発展のため，標準化のためのコーパス（TI digit+Noise）とそれを認識するための HTK を利用した標準スクリプト，標準スクリプトで得られるベースライン性能からの性能改善率を求める Microsoft Excel Spread Sheet を研究者に配布したものである．これまでに，TI digit に雑音を付与した連続数字音声認識タスクである AURORA-2，自動車内の連続数字／コマンドタスクである AURORA-3 をそのスクリプトと共に配布している [7], [8]．この AURORA のメリットは，タスクが連続数字と比較的小さく，1）大語彙連続音声認識に比べて簡単であること，2）ベースライン性能が配布される HTK スクリプトにより容易に得られること，があげられる．さらに，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である AURORA-4 も配布が開始されている [9]．次に，AURORA での標準コーパスとタスクについて述べる．

[AURORA2] TI-DIGITS に雑音を重畳したデータを評価するタスクである．

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）共に 110 名，8,440 発話（男女 55 名，4,220 発話ずつ）である．clean training の場合はこのデータに雑音を重畳しないで学習を行ない，multicondition training の場合は 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．

テストセットは大別して，

[テストセット A] 雑音は Subway, Babble, Car, Exhibition. チャネルフィルタは G.712. clean training ではチャネル条件がクローズ，multicondition training では雑音条件およびチャネル条件がクローズ．

[テストセット B] 雑音は Restaurant, Street, Airport, Station. チャネルフィルタは G.712. clean training, multicondition trainig 共にチャネル条件のみクローズ．

[テストセット C] 雑音は Subway, Street. チャネルフィルタは MIRS. clean training, multicondition trainig 共にチャネル条件がオープン．

の 3 種類となっている．基本となるテストデータは 104 名，4,004 発話（男女 52 名，2002 発話ずつ）で，テストセット A／B ではこれを 4 分割し各種雑音を 7 種類の SNR レベル（clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB）で重畳，テストセット C では半分の 2,002 発話をさらに 2 分割して各雑音を重畳している（各雑音・フィルタ条件に対して 1,001 発話）．同じ雑音・フィルタ条件ならば，SNR が違っても発話内容は同じである．

[AURORA3] 雑音を加算するのではなく実際に雑音下で収録した音声を用いて評価を行うことを目的としている．SpeechDat-Car プロジェクトで収録した自動車内発話音声を評価に用いる．対象言語をフィンランド語，イタリア語，スペイン語，ドイツ語，デンマーク語とし，実環境で遭遇する可能性のある Well-matched condition, Moderate-mismatched condition, High-mismatched condition の 3 つの状況を設定し評価項目としている．

[AURORA4] Noisy WSJ-large vocabulary evaluation: Wall Street Journal データベースに雑音とフィルタ特性を付加した大語彙データベースを用いて評価を行う [10], [11]．

3. IPSJ SLP 雑音下音声認識評価 WG

雑音下音声認識の技術の向上をめざして，AURORA と同様な活動を検討すべく，情報処理学会音声言語情報処理研究会の内部に，雑音下音声認識ワーキンググループを 2001 年 10 月に発足させた．本ワーキンググループの議論は大きく 2 つに分けられる．1 つめは，騒音下の音声認識を本来どのように評価すべきかという課題，もう一つは欧州で進んでいる AURORA プロジェクトとの関係である．WG としては，一つ目の課題を十分時間をかけ調査などをしながら進めつつ，内容がすでに確定し比較の容易な AURORA と同様のデータを収録していくこととした．基本的な方向としては，研究用に逐次難易度が高くなるように発話内容と雑音を設計したコーパスとベースラインスクリプトの配布しそれに基づいて各方法の優劣を比較する枠組みと，実アプリケーションに近いレベルの評価用に，典型的な数種類の雑音の選定および配布とそれに基づいて SNR を既定して平均認識率を測定する枠組みの 2 種類を設定している．

図 1 に WG にて，これまでに議論した結果のコーパスと評価タスクの開発に関するロードマップを示す．日本語版の AURORA-2 を構築した後，自動車内発話の単語，非定常雑音，残響，文発話，複数話者と発展させる予定である．また，この独自の評価フレームワークを CENSREC(Corpus and ENvironments for Noisy Speech RECognition) と呼ぶことにする．AURORA-2J が CENSREC-1, AURORA-3J が CENSREC-2 に対応する．

3.1 AURORA-2J/CENSREC-1

AURORA-2J は，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 月に配布を開始し，すでにこれまで AURORA-2 に対応する日本語連続数字コーパス，評価スクリプト AURORA-2J を 100set 以上配布した．本章では，AURORA-2J コーパスの

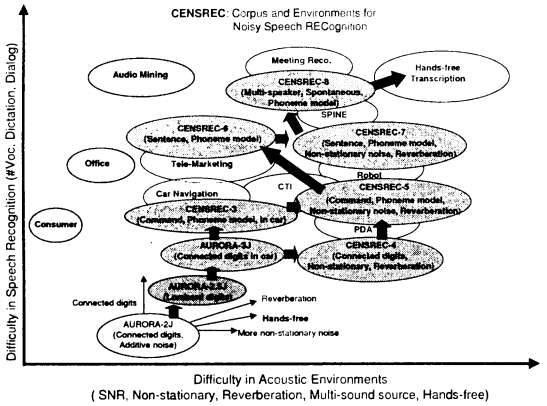

図 1　雑音下音声認識評価ロードマップ

Fig. 1　A road-map of noisy speech recognition evaluation.

収録と，その共通評価環境およびベースライン性能について述べる．

3.1.1　収録する数字列

AURORA-2J の発声リストは AURORA-2 と同一のものを用いた．また，話者数，男女比も同一で話者毎の発声リストも同一となっている．ただし，発声者は日本人，数字の読みは日本語でこの 2 点が AURORA-2 と異なる．

AURORA-2 では，“0” に対し/zero/，/oh/の 2 種類の発声が定義されており，発声リストおよびファイル名は “Z” および “O” と明確に区別されている．日本語の場合，“0” は「ぜろ」「れい」「まる」等と発声されるが，電話番号やクレジットカード番号を電話でオペレータ等に伝える場合，「ぜろ」と「まる」の比率が高い．そこで，AURORA-2J では “Z” を「ぜろ」，“O” を「まる」と発声させた．また同様の理由で，「し」，「しち」等の読みは採用しなかった．一方，“2” や “5” の長母音化に関しては発声者の自由とした．

3.1.2　ダウンサンプリングおよび雑音重畳

AURORA-2J では AURORA-2 のデータ作成方法と同一の手法により音声に雑音を重畳した．そのために必要となる雑音信号 (Subway, Babble, Car, Exhibition, Restaurant, Street, Airport, Station の 8 種類)，各種フィルタ (G.712, MIRS の 2 種類)，およびソースプログラムとスクリプトファイルは，すべて AURORA プロジェクトから提供を受けた．この提供より AURORA-2J では，AURORA-2 と全く同じ条件によるデータの作成を実現した．AURORA-2J では，収録した 16kHz の日本語音声信号を 8kHz にダウンサンプリングした後に，AURORA-2 プロジェクトから提供を受けた雑音信号と各種フィルタを用いて，雑音重畳作業およびフィルタリング作業を行った．

3.1.3　学習／テストデータの構成

学習およびテストデータの構成は，学習は Clean Training と Multicondition Training，テストは Set A, Set B, Set C の AURORA-2 のものをそのまま採用している．

3.1.4　評価用スクリプト

評価用ベースラインスクリプトは，AURORA-2 と同様に HTK [12] を用いて HMM の学習および認識実験を行うよう，AURORA-2 で配布されているスクリプトをベースとして作成されている．HMM トポロジー，特徴量など複数の条件で実験を行ない，様々な議論を重ねた結果，AURORA-2 を踏襲する形でベースラインスクリプトの仕様を以下のように定めた．

- スクリプトは sh(bsh) スクリプトであり，一部（初期モデル生成プログラムなど）は perl スクリプトで書かれている．
- HMM は先に述べた 10 数字（11 モデル）と，長さの異なる 2 種類の無音 (sil, nn sp) の計 13 モデルである．
- 数字 HMM は 18 状態（出力分布を持つ状態は 16），長い無音モデル (sil) は 5 状態（同じく 3 状態），短い無音モデル (sp) は 3 状態（同じく 1 状態）のモデルである．sp の出力分布は sil の真中の状態と共有される．
- 各状態のガウス混合分布は 20 混合（無音モデルは 36 混合）である．
- ベースラインの特徴パラメータは，HTK の HCopy により特徴抽出された MFCC（12 次元）+ΔMFCC（12 次元）+ΔΔMFCC（12 次元）+log power（1 次元）+Δpower（1 次元）+ΔΔpower（1 次元）の計 39 次元とする．分析条件は，$1-0.97z^{-1}$ のプリエンファシス，ハミング窓，25ms の分析フレーム長，10ms のフレームシフトとする（ただし，64Hz 未満は使用しない: LOFREQ=64 と設定している）．

3.1.5　ベースライン性能と認識性能比較

前節のスクリプトによりベースライン性能が得られる．このベースライン性能は，Microsoft Excel Spread Sheet により，各テストセット，各雑音，SNR 毎ごとに平均の認識性能が得られる．ここで，各雑音毎の平均は，SNR 20dB〜0dB の平均値である．さらに，この Spread Sheet に各機関で得られる認識率を入力すれば，認識実験結果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することができる．これによって機関毎の認識性能比較を容易に行うことができる．

3.1.6　CENSREC-3

AURORA-2J に続く雑音下音声認識の標準環境である，CENSREC-3 を作成し，ベースライン性能評価のための標準スクリプトを含めて配布する [19]．また，AURORA-2J が人工的に雑音が付与された連続数字の認識タスクであったのに対し CENSREC-3 では，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．

音声データの収録は，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 種類の走行速度と 6 種類の車内環境を組み合わせた 16 種類の環境下で行っており，これらの音声データを用いた 6 種類の評価環境 (Condition 1〜6) を提供する．CENSREC-3 で設定する 6 種類の評価環境は，AURORA3 の 3 種類の評価環境である，Well-matched condition, Moderate-mismatched condition, High-mismatched condition に準じており，以下のような対応となっている．

Condition 1, 2, 3　学習データと評価データのマイクロホン種別，走行環境が一致する条件下で評価を行う．この評価環境は，AURORA3 の Well-matched condition に相当する．

Condition 4　学習データと評価データのマイクロホン種別は一致するが，走行環境が異なる条件下で評価を行う．この評価環境は，AURORA3 の Moderate-mismatched condition に相当する．

Condition 5, 6　学習データと評価データのマイクロホン種別，走行環境（の一部）が共に異なる条件下で評価を行う．この評価環境は，AURORA3 の High-mismatched condition に相当する．

CENSREC-3 の認識対象は，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り，自動車内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 により認識を行う．また，研究機関毎の認識性能比較を容易にするためのスプレッドシート（表 1）の配布と，評価時のバックエンド部分の変更（HMM の学習方法，トポロジーの変更，特徴量の変更など）に対する評価カテゴリーを設定する．CENSREC-3 の詳細は，文献[19] を参照されたい．

表 1　CENSREC-3 スプレッドシート
Table 1　CENSREC-3 spread sheet.

CENSREC-3 Evaluation Results

CENSREC-3 Baseline Results (%)						
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Average
88.43	99.31	78.36	52.95	36.20	25.50	

CENSREC-3 Word Accuracy (%)						
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Average

CENSREC-3 Relative Improvement						
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Average

3.1.7　CENSREC-2/AURORA-3J

実走行車内での孤立単語音声認識を対象とした CENSREC-3 に対して，現在，実走行車内での連続数字認識を対象とした CENSREC-2 (AURORA-3J) の作成を予定している．

CENSREC-2 においても，CENSREC-3 と同様の収録条件，評価環境を提供することを予定しており，収録音声の発話内容は，AURORA2-J の 4 種類の発話セットのうち，各数字の出現頻度のバランスが最も良い 1 セットを用いる．収録する発話者数は 104 名 (男女各 52 名) で，一人当たりの発話数は 9 もしくは 10 文である．1 環境あたりの総発話数は 1001 文であり，これらを走行環境毎に収録する．また，評価環境は，CENSREC-3 と同等の 6 種類程度を予定している．

3.1.8　CENSREC-1.5/AURORA-2.5J

雑音環境下で発生された音声は，時間領域で加算される音響雑音による歪みに加え，種々の非線形な歪みの影響を受ける．非線形歪みの原因の一つは，我々が雑音環境下でより効率的な音声によるコミュニケーションを実現するために，聴覚により得られる外界からの情報に応じて発声機構を適応的に変化させているためである．この非線形な歪みは，ロンバード効果と呼ばれている．

本ワーキンググループでは，AURORA-2J のテストセットの一部を対象とし，ロンバード効果も考慮した雑音環境で録音されたコーパス CENSREC-1.5 の構築を行う予定である．本評価環境は，AURORA-2J と同一の連続数字タスクを用いるが，収録方法は CENSREC-2,3 と同様に実環境下での雑音重畳であるため（AURORA-2J は計算機上での雑音重畳），CENSREC-1.5/AURORA-2.5J と名付ける．具体的には防音室内にて，AURORA-2J で用いられている雑音あるいは広帯域ランダム雑音のいずれかをスピーカにより再生した雑音環境において収録を行う．雑音の大きさは，SNR ではなく，話者位置における音圧レベル dB(A) により規定する．

3.1.9　CENSREC-1,2-AV/AURORA-2,3J-AV

近年，音声に映像情報を併用した Audio-Visual 音声認識の研究が多く行われている [20]．しかし，複数の研究機関で共用できる Audio-Visual 音声認識の評価用データベースは現在のところ公開されていない．そこで我々は，Audio-Visual 音声認識の性能評価環境を提供するために，名古屋大学末永研究グループと共同で，音声に顔映像を加えた CENSREC-1-AV/AURORA-2J-AV, CENSREC-2-AV/AURORA-3J-AV のデータ収録・整備を進めている [21].

CENSREC-1-AV では，AURORA-2J と同様に，室内の静かな環境で収録した音声に人工雑音を事後的に重畳する．一方，顔映像はブルーバックで撮影しており，顔映像への背景雑音・照明雑音の付加手法についても現在検討中である．CENSREC-2-AV では，CENSREC-2/AURORA-3J と同様に，アイドリング・市街地走行・高速走行における運転手の音声と顔映像を収録しており，実車内環境での Audio-Visual 音声認識の性能評価が可能となる．なお，両者ともに，夜間などの可視光照明が得られない状況も想定し，顔画像は，通常のカラー映像に加えて，近赤外映像も併せて収録している．

3.2　性能評価手法

3.2.1　各種認識率による評価方法

音声認識には話者依存性があるために，認識性能を評価する際には多数の話者のデータを用いるのが一般的である．一方，評価指標である認識率は，話者を区別して算出されないことが多く，このままでは話者依存性の影響を正確に評価できないという問題がある．そこで，筆者らは，話者毎の認識率に基づく評価指標を考案した [13]．具体的には以下の 3 つの指標である．

- 話者毎の認識率の最大，最小，平均，標準偏差
- 話者毎の認識率のヒストグラム
- 認識率が x%以上である話者の割合

以下では，ETSI の DSR 用標準フロントエンドである ES 202 050 [14] の性能評価を AURORA-2J を用いて行い，これらの評価指標の必要性と有効性を示す．

3.2.2　話者毎の認識率の最大，最小，平均，標準偏差

表 2 に話者毎の単語正解精度（Overall）の最大，最小，平均，標準偏差を示す．

ES 202 050 の最大値を見ると，Clean training の場合は 88.04%，Multicondition training の場合に至っては 96.03%を得ていることが分かる．また，ES 202 050 の標準偏差はベースラインよりも小さくなっており，一定の改善が認められる．しかし，最大値と最小値の差は依然として大きいことが分かる．

3.2.3　話者毎の認識率のヒストグラム

Clean training，テストセット A，Subway，SNR5dB の場合の，話者毎の単語正解精度のヒストグラムを図 2 に示す．図中の横軸は単語正解精度の範囲，縦軸は話者数である．

表 2　話者毎の単語正解精度（Overall）の最大，最小，平均，標準偏差
Table 2　Maximum, minimum, average value and standard deviation of the word accuracy per speaker.

		Baseline	ES 202 050
Clean training	Maximum	61.24	88.04
	Minimum	26.33	62.08
	Average	45.51	77.82
	SD	6.06	5.14
Multicondition training	Maximum	93.98	96.03
	Minimum	74.30	79.44
	Average	85.83	90.99
	SD	3.64	3.01

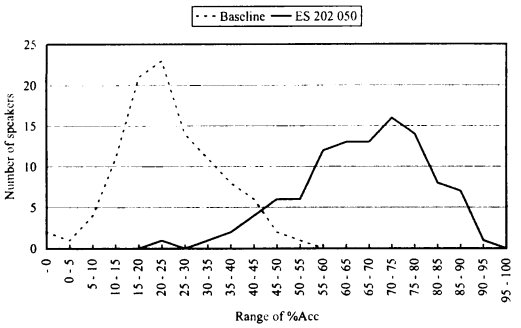

図 2　話者毎の単語正解精度のヒストグラム

Fig. 2　Histogram of the word accuracy per speaker.

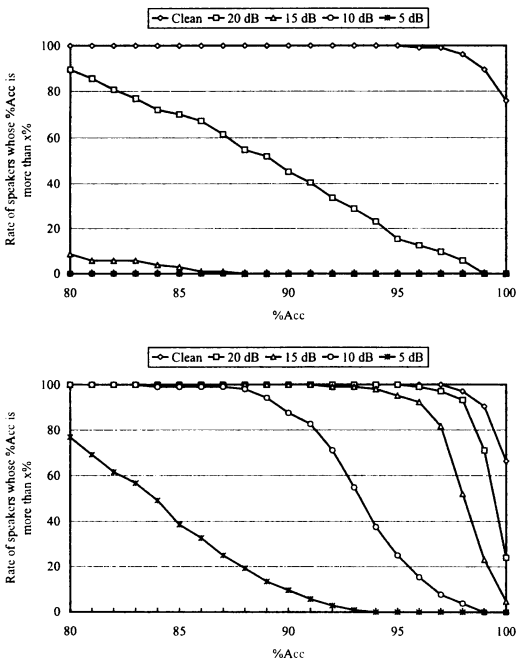

図 3　単語正解精度が x%以上の話者の割合（ベースライン）（上：Clean training，下：Multicondition training）

Fig. 3　Rate of speakers whose word accuracy is more than x% for the baseline front-end. The upper is in the clean training and the lower is in the multicondition training.

ES 202 050 の分布は，ベースラインと比べて単語正解精度が高い方に移動していることが分かる．その一方で，ES 202 050 の分布は，中心付近から左側への広がりが大きく，単語正解精度が依然として低い話者が相当数存在していることが見て取れる．

3.2.4　認識率が x%以上である話者の割合

単語正解精度が x%以上の話者の割合を図 3（ベースライン）と図 4（ES 202 050）に示す．

単語正解精度の目標値 x を高くするほど，その単語正解精度が得られる話者の割合が減っていることが分かる．特に，目標値が 100%に近いときの落ち込みが大きい．この割合は，音

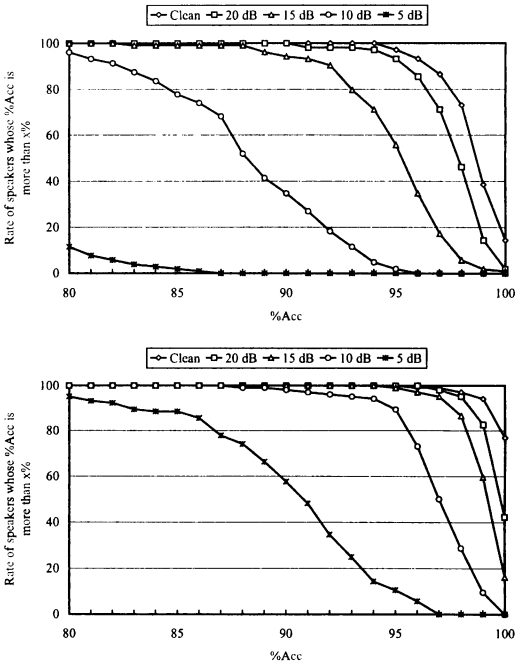

図 4　単語正解精度が x%以上の話者の割合（ES 202 050）（上：Clean training，下：Multicondition training）

Fig. 4　Rate of speakers whose word accuracy is more than x% for ES 202 050. The upper is in the clean training and the lower is in the multicondition training.

声認識サービスの適用可否の判断に用いることができる．例えば，$x = 90$ のときに 90%の話者を確保できる SNR に着目すると，ベースラインでは，Clean training の場合は Clean のみ，Multicondition training の場合は SNR15dB 以上である．一方，ES 202 050 では，Clean training の場合は SNR15dB 以上，Multicondition training の場合は SNR10dB 以上であり，適用可能な SNR の範囲が広がっていることが分かる．

3.2.5　認識率の推定

音声認識サービスを提供する際には，サービス品質を保証するために，対象とする雑音環境でどの程度の認識率が得られるのかを事前に知る必要がある．

そのための一手法として，音声のひずみ値により認識率を推定する方法がある．最近，ひずみ尺度として ITU-T 勧告 P.862 の PESQ (Perceptual Evaluation of Speech Quality) [15]，大量の音声データの代わりに ITU-T 勧告 P.50 の擬似音声 [16] を用いて認識率を推定する方法が提案されており，まずまずの精度で認識率を推定できることが示されている [17]．しかし，この手法ではひずみの計算に原信号（クリーンな信号）が必要であることから，著者らは認識対象の音声信号のみから，SNR や残響特性を測定し，そこから認識率を推定する方法について研究を進めている．

3.2.6　残響の影響

これまで音声認識評価における残響の尺度については，残響時間および音源とマイクロホンまでの距離が一般的であった．しかしながら，これらの尺度だけでは音声認識時の残響の影響

を十分に表現できていないという問題があった．問題点を列挙すると下記のとおりである．

（1） 同じ室内（残響時間が同じ）でも場所によって残響の度合いは異なる．

（2） 残響時間では「初期残響が大きい」のか「残存時間が長い」のか正確に判断できない．

（3） 音声認識は大きな初期残響よりも残存時間が長いほうが認識率は劣化する可能性がある．

この問題を解決するために，現在我々は上記問題点を踏まえて新しい残響尺度の検討を行っている．具体的には，部屋，スピーカ，マイクロホンの位置情報から直接音，1次反射音，2次反射音などを同定し，その結果に基づく信号対残響比（Signal to Reflection Noise Ratio: SRR）を算出することで新しい残響の尺度を模索中である．現在は系のインパルス応答を用いて残響尺度の算出を模索中であるが，将来的には受音信号（残響あり音声）を入力すれば，自動的に残響尺度が算出されるような評価尺度を目指す予定である．

3.2.7 評価雑音セットの選定

実際のアプリケーションを考える際には，アプリケーション毎にタスクが異なるので，発話内容を規定することが難しい．そこで，代表的な雑音セットとSNRを規定して評価することにより，性能を測定することを考えている．自動車の10/15モード燃費のようないろいろなモードの雑音で認識性能を測定し，平均するような形で測定する．現在，AURORA-2J等のデータベースで規定された雑音環境以外の，様々な環境下での音声認識評価を容易にすることを目的とした，評価用雑音データベースの作成を予定している．評価用雑音データベースの作成にあたり，実環境で収録された種々の雑音[18]より，選定した10種類程度の雑音を1セットとして構成することを検討している．

3.2.8 評価カテゴリー

これまでにAURORA-2/3を使用した研究結果が数多く公表されているが，元来音声認識フロントエンドの改善・評価を目的としたプロジェクトであるにもかかわらず，これらの発表の中にはバックエンド（提供されている学習／テスト用スクリプト）の変更を伴うものが多く存在し（例えば，HMMの混合数を増加する，モデル単位を変更してコンテキスト依存モデルを導入する，AURORA-2データベースに含まれていない雑音データを使用するなど），これらを同一に比較評価することが難しくなってきている．

そこで，CENSRECでは，バックエンドの変更に対して，その度合に応じたカテゴリーを設定する．バックエンドを変更した結果を発表する場合，提示するカテゴリーから一つを選び，発表でそれを示すよう規定する．カテゴリー内で性能比較を行うことで，各手法の性能比較をより適切に行うことができると考えられる．

4. まとめ

本稿では，実環境における音声認識技術の進展を促進するための種々の活動について紹介した．特に，著者らが情報処理学会 音声言語情報処理研究会 雑音下音声認識評価WGにおいて進める音声認識の評価の枠組みの現状と今後の予定について述べた．今後さらに関係各所との意見交換を行い，検討，活動を進める予定である．

[謝辞]

本研究の一部は独立行政法人 情報通信研究機構の研究委託により実施したものである．

文献

[1] 中村 哲, “実音響環境に頑健な音声認識を目指して,” 電子情報通信学会 技術報告, SP2002-12, pp.31-36, Apr. 2002.
[2] http://elazar.itd.nrl.navy.mil/spine/
[3] http://eurospeech2001.org/ese/NoiseRobust/
[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.
[5] 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.
[6] ETSI standard document, “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm,” ETSI ES 201 108 v1.1.2, Apr. 2000.
[7] H. G. Hirsh and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” ISCA ITRW ASR2000, pp.18-20, Apr. 2000.
[8] D. Pearce, “Developing the ETSI AURORA advanced distributed speech recognition front-end & What next,” Proc. Eurospeech 2001, 2001.
[9] http://www.elda.fr/
[10] Aurora document no. AU/337/01, “Experimental framework for the performance evaluation of speech recognition fron-ends on a large vocabulary task: Version 1.0,” Ericsson, June 2001.
[11] Aurora document no. AU/345/01, “Large vocabulary evaluation of front-ends- baseline recognition system description,” Mississipppi State University, Aug. 2001.
[12] http://htk.eng.cam.ac.uk/
[13] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” IEICE Transactions on Information and Systems, Vol.E88-D, No.3, Mar. 2005. (to appear)
[14] ETSI ES 202 050 V1.1.1, “Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,” 2002.
[15] 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.
[16] ITU-T Recommendation P.50, “Artificial voices,” Sept. 1999.
[17] Takeshi Yamada, Nobuhiko Kitawaki, “A PESQ-based performance prediction method for noisy speech recognition,” Proc. International Congress on Acoustics, ICA2004, Vol.II, pp.1695–1698, Apr. 2004.
[18] 遠藤俊樹, 中村 哲, “実環境騒音 DB の収集及び DSR フロントエンドによる音声認識実験,” 音講論集, 1-P-13, pp. 187-188, Sept. 2004.
[19] 藤本雅清, 中村 哲, 武田一哉, 黒岩眞吾, 山田武志, 北岡教英, 山本一公, 水町光徳, 西浦敬信, 佐宗 晃, 宮島千代美, 遠藤俊樹, “CENSREC-3: 実走行車内単語音声データベースと評価環境の構築,” 信学技報, Dec. 2004. (to appear)
[20] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent advances in the automatic recognition of audio-visual speech,” Proceedings of IEEE, Vol.91, no.9, pp.1306–1326, Sept. 2003.
[21] 根木大介, 前野俊樹, 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, “映像付き雑音環境下音声認識評価用共通データベース AURORA-2J-AV/AURORA-3J-AV の構築,” 信学技報, PRMU, May. 2004.