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

2005－ＳＬＰ－59 (24)
2005／12／22

SLP雑音下音声認識評価WG活動報告

－評価用データと評価手法について－

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] 滝口 哲也*

[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 * 神戸大学

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，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のワーキンググループを2001年10月に組織し、評価用標準コーパス、標準バックエンドの作成、配布を行ってきた。本稿では，本活動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

A Report of SLP Speech Recognition Evaluation WG

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

[1] ATR Spoken Language Communication 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 * Kobe University

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. To ovecome these problems, we organized a working group under Information Processing Society of Japan. This paper introduces current activities and a future road-map of a common standardized framework for noisy speech recognition by the working group organized by the authors.

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

1. はじめに

音声認識の性能の客観的な比較評価を可能にすることは研究開発の効率を促進する上で不可欠である．たとえば、これまで、多くの研究発表がなされてきたが、評価実験の実験条件や評価データが異なるなど、客観的な比較評価が困難であった．さらに、評価データは、通常、提案手法を確認することを目的に設計された小規模データであることが多く、実世界に存在する課題のどの範囲、程度を扱っているのかなどの関連が不十分な状況があった[1]．

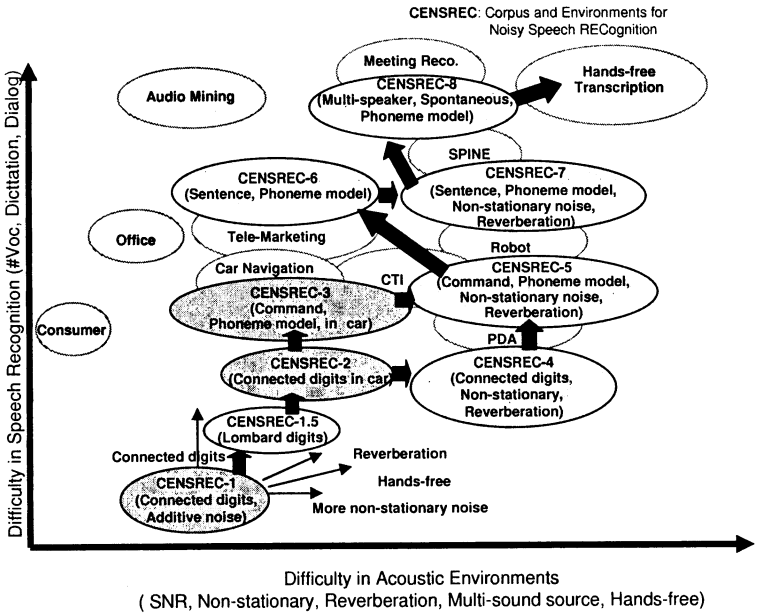

図 1　雑音下音声認識評価ロードマップ

客観的な比較評価が研究開発に大きな進歩をもたらした例として米国の国防総省の DARPA プロジェクトがある. このプロジェクトでは、DAPRA の研究予算のもと、学習データ、評価データを揃えることで、客観的な比較評価を可能にし、実用的な音声認識技術の発展を導いた. このプロジェクトにより、現在、音声認識の主流になっている HMM や N-gram 技術が実用的なレベルまで成長した.

本報告で対象としている雑音下の音声認識は、雑音の種類など要因が複雑で客観的な比較評価がより困難であるが、カーナビ, 携帯電話などの音声認識技術の主要な適用領域において必要となる要素技術である. このような雑音下の音声認識の問題に対して，欧州では AURORA [2] 研究プロジェクトが進められた. 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. さらに、このデータを用いた研究発表のスペシャルセッションを Eurospeech，ICSLP で企画し共通データを利用した比較評価を促進した [3]. 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、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 をそのスクリプトと共に配布している [4], [5]. 特に、この AURORA-2,3 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今日では比較評価の標準的な実験環境として利用されている.

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 [6], [7]. 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，これまで IPSJ-SLP-NOISEWG の活動と現状，今後のロードマップについて述べる.

2. IPSJ SLP 雑音下音声認識評価 WG

本ワーキンググループ（WG) 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. これまで、議論を重ねながら、AURORA の日本語版といえる雑音下音声認識評価のための共通学習データ、評価データ、バックエンドスクリプト、評価用のスプレッドシートを、評価フレームワークを CENSREC(Corpus and ENvironments for Noisy Speech RECognition) と呼び、これまで配布を行ってきた.

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. CENSREC-1(AURORA-2J) が AURORA-2、CENSREC-2 が AURORA-3 に対応している.WG では、これまで、AURORA-2 の日本語版である CENSREC-1 を作成・配布した後，自動車内単語実発話の CENSREC-3 を配布し、今回、AURORA-3 の日本語版である CENSREC-2 を配布するに至っている. 今後もさらに、

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．以下，これまで作成・配布してきたCENSREC-1,2,3について述べる．

2.1 CENSREC-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，すでにこれまで日本語連続数字コーパス，評価スクリプトとして100set以上配布した．本節では，CENSREC-1コーパスの収録と，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．

2.1.1 収録する数字列

CENSREC-1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．

CENSREC-1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，"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．

2.1.2 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．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．

2.1.3 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を用いてHMMの学習および認識実験を行なうよう，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）．

2.1.4 ベースライン性能と認識性能比較

提供されている認識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，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．

図2にベースラインとなる認識性能を示す．この結果は，認識結果を集計するため，共通のMicrosoft Excel Spread Sheetとして配布される．

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．

%Acc				
	A	B	C	Overall
Clean Training	46.51	43.98	49.90	46.17
Multicondition training	91.53	80.39	85.83	85.93
Average	69.02	62.18	67.86	66.05

図2 CENSREC-1のベースライン性能(%)

2.2 CENSREC-2

CENSREC-2 は、実際に自動車内で発話した連続数字データを対象とした評価タスクである.CENSREC-1 が連続数字発話に種々の雑音を計算機上で加算したデータであったのに対し、実環境発話のデータを対象とすることにより、シミュレーションと実際の差を確認することができる.

2.2.1 音声データの収録環境

自動車内音声の収録は、名古屋大学で特別に設計された実験車両を用いて行っている. 実験車両には, 運転席周辺に 5 本のマイクロホンが図 3 に示すような位置に設置されており, 3, 4 番はダッシュボード上, 5, 6, 7 番は天井に設置されている. また, 1 番は接話マイクロホンである. これらのマイクロホンの内, CENSREC-2 では, 1 番の接話マイクロホンと 6 番の遠隔マイクロホンで収録された音声を用いる [8]. それぞれのマイクロホンには, SONY ECM77B を用いている. データの収録条件は、表 1 に示す 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 種類の環境で行う. 評価データの発話者数は 104 名であり、収録音声の総数は 17,651 発話である. このうち、73 名分を学習データとし、31 名を評価用とした.

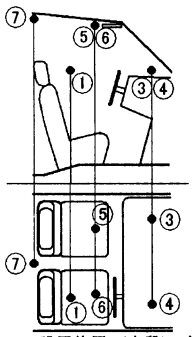

図 3 マイクロホンの設置位置 (上段) : 側面, (下段) : 真上

表 1 CENSREC-2 音声データ収録環境

走行速度	車内環境
アイドリング	通常走行、エアコン On、オーディオ On、窓あけ
低速走行	通常走行、エアコン On、オーディオ On、窓あけ
高速走行	通常走行、エアコン On、オーディオ On、窓あけ
アイドリング	通常走行、エアコン On、オーディオ On、窓あけ

2.2.2 評価環境の設計

CENSREC-2 では、様々な環境で収録された音声データを用いて、4 種類の音声認識票か環境 (Cond. 1〜4) を構成する.

Cond. 1: 学習と評価でマイク種別、収録環境がともに一致
Cond. 2: 学習と評価でマイク種別が一致、収録環境が相違
Cond. 3: 学習と評価でマイク種別が相違、収録環境が一致
Cond. 4: 学習と評価でマイク種別、収録環境がともに相違

図 2.2.2 に CENSREC-2 のベースライン性能を示す. CENSREC-2 に関する詳細は、文献 [9] を参照されたい. 本評価環境は、2005 年 12 月より配布を開始している.

表 2 CENSREC-2: 各評価環境で用いられる学習データ

評価環境	Condition 1		Condition 2		Condition 3		Condition 4	
マイクロホン	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔
アイドリング	—	○	—	○	○	—	○	—
低速走行	—	○	—	—	○	—	—	—
高速走行	—	○	—	—	○	—	—	—

表 3 CENSREC-2: 各評価環境で用いられる評価データ

評価環境	Condition 1	Condition 2	Condition 3	Condition 4
アイドリング	○	—	—	—
低速走行	○	○	○	○
高速走行	○	○	○	○

CENSREC-2 Baseline Results (%)				
Condition 1	Condition 2	Condition 3	Condition 4	Average
80.58	74.49	61.46	48.87	66.35

図 4 CENSREC-2 のベースライン性能 (%)

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

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

表 4 CENSREC-3: 評価データの収録環境

走行速度	車内環境
アイドリング	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
低速走行	通常走行, エアコン (Low), エアコン (High), オーディオ On, 窓開
高速走行	通常走行, エアコン (Low), エアコン (High), オーディオ On, 窓開

表 5 CENSREC-3: 各評価環境で用いられる学習データ

評価環境	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
マイクロホン	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔
アイドリング	○	○	○	—	—	○	—	○	○	—	○	—
低速走行	○	○	○	—	—	○	—	—	○	—	—	—

表 6 CENSREC-3: 各評価環境で用いられる評価データ

評価環境	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
マイクロホン	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔	接話	遠隔
アイドリング	○	○	○	—	—	○	—	—	—	—	—	—
低速走行	○	○	○	—	—	○	—	○	—	○	—	○
高速走行	○	○	○	—	—	○	—	○	—	○	—	○

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．

2.4 音声データの収録環境

自動車内音声の収録は，CENSREC-2 と同じ特別に設計された実験車両を用いて行う．使用したマイクロフォンも CENSREC-2 と同条件である．

評価データの収録は，表 4 に示す，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 種類の環境で行う [10]．評価データの発話者数は 18 名（男性 8 名，女性 10 名）であり，収録音声の総数は，マイクロホン一本あたり 14,216 発話である．

図 2.4 に CENSREC-2 のベースライン性能を示す．CENSREC-3 の詳細は，文献 [11] を参照されたい．

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	83.48

図 5 CENSREC-3 のベースライン性能 (%)

2.5 CENSREC-1,2-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．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 [12]．

3. 性能評価手法

3.1 各種認識率による評価方法

現在のところ，音声認識の性能評価には認識率を用いるのが一般的である．しかし，認識率は「平均的な」性能を示しているにすぎず，認識誤りの傾向を伺うことはできない．例えば，認識性能に話者依存性があることは良く知られており，それを評価する方法が必要である．そこで，我々は，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を用いて話者依存性を評価する方法を考案し，その有効性を示した [13]．これらの評価ツールは，[14] で公開している．

認識誤りの傾向としては，他にも様々なものが考えられる．例えば，音声対話システムの場合には，人間同士の会話でも起こり得るような違和感のない誤りが望ましいであろうし，音声ワープロの場合には，逆に違和感のある誤りの方が誤りの発見が容易という意味で望ましいのかもしれない．また，アプリケーションによっては，認識誤りがバースト的なのか，散発的なのかが重要な違いを持つことも考えられる．今後はこのようにアプリケーション側の視点から評価方法の検討を行っていく予定である．

3.2 認識率の推定

耐雑音手法の開発や音声認識サービスの品質保証のためには，対象とする雑音環境でどの程度の認識性能が得られるのかを調査する必要がある．しかし，実際に現場で認識実験を行う，あるいは大量の音声データを収集するのは，コストの面から非現実的である．最近，音声のひずみ尺度である PESQ [15] と擬似音声 [16] を用いて，容易かつ短時間に認識率を推定する方法が提案されている [17]．しかし，このアプローチは，雑音抑圧手法を前処理として用いる場合にしか適用できないという問題がある．今後は，実環境音声認識の性能に影響を与える要因を洗い出して定量化し，各々の値から総合的に認識性能を推定するモデルを開発する予定である．

4. 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．

- 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．
- 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．

- 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、室内環境、発話者、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た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. 最新の活動状況としては、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 [18] を行っている. 調査結果から初期の反射音については音声認識性能を劣化させるのではなく、性能を改善できる可能性があることを確認した. 今後継続して詳しく調査を行い残響時間、初期残響と音声認識率の関係に基づく音声認識における新たな残響の尺度化を目指す.

5. 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種々の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. 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 [19] 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.

6. ま と め

本稿では, 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.

現在、WG で把握している範囲では, 2003 年の配布開始から 2005 年秋音響までの発表件数 (CENSREC-1 を使用しているもの. ベースライン関係の発表を除く) は 39 件で, そのうち評価フレームワークを使用しているものが 27 件, 残りの 12 件は単にデータベースとして使用している. 各研究機関における耐雑音手法の評価だけでなく, データベースとして別目的に使用されているケースがかなりあり, このことからデータベースとして考えても有用なものとなって いると考えられる.

また、データ配布先にアンケートを配布し使用内容の調査も行っている. 現在までに、回収したデータでは、CENSREC-1 が CENSREC-2 より利用されており、CENSREC-1 の結果は殆どが学会発表に使用されたという回答を得ている. また、今後必要な標準化について、SNR の計算法、音声と雑音の加算法、残響の加算法、さらに認識率以外の主観性能を反映した評価尺度の策定などの要望が大きいことが明らかとなった.

これまでの活動を総括して、耐雑音手法の評価フレームワークの目的である様々な研究機関の結果をつき合わせて比較・検討できるような環境を作ることにはある程度成功していると考えられる. しかし, さらに広い利用の啓蒙、実用的な客観評価データベース、評価バックエンド、標準ツール、などの比較しやすい "比較し易い土台作り" に以前大きな期待と要望があり、学界全体として今後も活動を盛り上げていく必要がある.

[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://eurospeech2001.org/ese/NoiseRobust/
[3] 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.
[4] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next," Proc. Eurospeech 2001, 2001.
[5] Aurora document no. AU/345/01, "Large vocabulary evaluation of front-ends- baseline recognition system description," Mississipppi State University, Aug. 2001.
[6] 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.
[7] 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.
[8] 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.
[9] 藤本雅清、武田一哉、中村 哲、"自動車内における連続数字音声コーパス CENSREC-2 の設計と評価"、信学技報, Dec. 2005. (to appear)
[10] 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.
[11] 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. SP
[12] 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.
[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.
[14] http://sp.shinshu-u.ac.jp/AURORA-J/
[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] 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 年度春季研究発表会、Mar, 2006. (to appear)
[19] 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.