2005－SLP－57（8）
2005／7／15

ATR実環境雑音DB -ATRANS- を用いた雑音重畳音声認識実験

遠藤 俊樹，　堀内 俊治，　清水 徹，　中村 哲

ATR 音声言語コミュニケーション研究所
〒 619-0288 京都府相楽郡精華町光台 2-2-2
Tel: 0774-95-1301　　Fax: 0774-95-1308
E-mail: {toshiki.endo,toshiharu.horiuchi,tohru.shimizu,satoshi.nakamura}@atr.jp

あらまし　　本稿では，ATR により収録した実環境雑音 DB を日本語連続数字タスクである AURORA-2J の音声信号に計算機上で重畳させ，音声認識実験を行った結果について報告する．音声と雑音を重畳する際に問題となるマイクロホンの違いによる周波数特性 (特に有効周波数帯域) の差異を回避する手法として，各々のマイクロホンのインパルス応答によって特性補正フィルタを作成し適用した．分散型音声認識システム (DSR) と雑音除去機能を有する拡張分散型音声認識システム (ADSR) を介した音声認識実験結果から，ADSR による雑音除去は特に車内雑音のような定常雑音に有効であること，テストデータと同じ環境の雑音を重畳した学習データから音響モデルを作成するマッチド学習は，定常雑音だけでなく一部の非定常雑音に対しても有効であることが明らかになった．

キーワード:　ATR 実環境雑音 DB，雑音重畳音声認識実験，AURORA-2J

Speech Recognition Experiments with ATR Ambient Noise Sound DataBase -ATRANS-

Toshiki Endo,　Toshiharu Horiuchi,　Tohru Shimizu and Satoshi Nakamura

ATR Spoken Language Communication Research Laboratories
2-2-2, Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0288 Japan
Tel: 0774-95-1301　　Fax: 0774-95-1308
E-mail: {tōshiki.endo,toshiharu.horiuchi,tohru.shimizu,satoshi.nakamura}@atr.jp

Abstract　　In this paper, we describe method for noise additive speech recognition experiment with ATR Ambient Noise Sound database (ATRANS) and AURORA-2J database. In this method, channel characteristic (especially efficient frequency band) difference between microphons used for speech data recording and noise data recording is compensated. The speech recognition results using the Distibuted Speech Recognition (DSR) and the Advanced DSR (ADSR) shows that ADSR is effective for notably stationary noise such as in-car noise, and matched trained acoustic model is effective not only for stationary noise, but also for a kind of non-stationary noise.

Keywords:　ATR Ambient Noise Sound Database, Noise additive speech recognition experiment, AURORA-2J

1　はじめに

実環境での音声認識システムの利用では，周囲雑音や残響による性能劣化が大きな課題であり，これに対し頑健な音声認識手法の研究が盛んになされている．

耐雑音認識手法を評価する場合，SPINE (Speech recognition In Noisy Environments) や AURORA プロジェクトなどが用意した共通のデータ，評価法が用いられている．日本でも，情報処理学会 音声言語情報処理研究会内の雑音下音声認識評価ワーキンググループ [1] により，雑音重畳音声データや評価ツールが配布され [2, 3]，音声認識の研究に使用されている．

また，耐雑音認識手法の研究に活用できる雑音 DB としては，電子協雑音 DB[4]，NoiseX92[5]，JIS 生活環境音データベース [6]，RWCP 実環境音響音声 DB[7]，環境騒音 DB[8] がこれまでリリースされている．以下各々の特徴を記す．

- **NoiseX92**：工業，軍事関連の特殊な雑音 DB.
- **JIS 生活環境データベース**：家庭内の生活音雑音 DB.
- **電子協雑音 DB**：17 種類の長時間収録の雑音 DB．ただし雑音種別毎に収録装置構成が異なる．
- **RWCP 実環境音響音声 DB**：非定常で短い環境音や様々な部屋のインパルス応答を含む DB.
- **環境騒音 DB**：移動体通信端末機の利用環境と想定される非常に多くの環境雑音を，単一指向性マイクロホンを 4ch 組合せて収録されている．

表 1: 収録雑音種別分類と収録雑音例

	屋外	屋内
交通関連	バスターミナル，空港ロータリ，駅改札口，街道，産業用道路，飛行場 (7 種類)	駅ホーム，電車内，車内，機内，空港ロビー駅地下通路等 (25 種類)
商業関連	駅前広場，市場 (2 種類)	デパート食品売場，マーケット，エレベータホール，地下道，ホテルのロビー，展示会場飲食店，電話ボックス等 (13 種類)
オフィス関連	-	受付，居室，マシンルーム等 (4 種類)
工業関連	道路工事，建築工事等 (3 種類)	板金工場，物流センタ，ボイラー室 (3 種類)
その他	競技場，田圃，森林，サイレン等 (7 種類)	体育館，ジム，ボーリング場等 (5 種類)

これらの雑音 DB は，各々の目的と雑音種別が異なる上，使用された収録装置が異なるため，利用目的に応じて使用方法に注意を要する．

筆者らは，既存雑音 DB の収録雑音を網羅するばかりでなく，さらに細かく条件の異なる環境を含む身近な広範囲の実環境雑音を収録した (ATR Ambient Noise Sound database : ATRANS)．たとえば，車内雑音では，車種，速度，窓の開閉，天候による条件の異なる雑音を収録することや，既存の DB にない自然の音やサイレン関連雑音などを新たに収録した．また，本 DB を用いた簡易な音声認識評価結果を [9] で報告した．

本稿では，ATRANS と日本語連続数字タスクの AURORA-2J の音声信号を用いた雑音重畳音声認識実験に関して記述する．AURORA-2J の音声信号は音声認識を，ATRANS の雑音信号は音声認識以外の用途にも使用することを目的としているため，異なるマイクロホンにより収録されている．そこで，AURORA-2J の音声信号と ATRANS の雑音信号間の周波数特性 (特に有効周波数帯域）の差異を補正し，計算機上で雑音重畳した音声に対して，DSR，ADSR フロントエンド処理を施した場合の音声認識実験結果を求めた．その結果，ADSR による雑音除去では，特に定常雑音に有効であること，マッチド学習は一部の非定常雑音にも有効であることがわかった．

2 ATR 実環境雑音 DB -ATRANS-

ATRANS は，身近な日常環境の雑音を収録した DB である．筆者らは，収録において以下の 3 つの点を考慮した．

1. 多様な目的に利用可能であること．
2. より多種類の環境雑音を収録すること．
3. 各雑音種別内で，状況変化を考慮すること．

これらの点を念頭に置き，収録機器構成，収録する雑音種別，収録方法等の決定を行った．以下詳細を記述する．

2.1 機器構成と録音方法

ATRANS は音声認識のみならず，音声端末機の周囲環境を含めた品質評価や，ATRANS を含めた音声による聴覚主観評価など，多様な目的に利用可能とするように配慮した．そのため，収録に用いたマイクロホンは，少なくとも，音声認識システムで広く使用される Sennheiser 社製 HMD410-6 マイクロホンより広い有効周波数帯域を持ち，またその帯域内で受音レベルに片寄りの少ないフラットな周波数特性と高い感度を持つものとして，DPA 社製 4060 コンデンサマイクロホンを使用した．また，DAT 録音機 (Sony 社製 TCD-D10 ProII) を用い，16 ビット 48kHz サンプリングで収録した．雑音種別により，雑音レベルが大きく異なるため，収録の際には，オーバーフローを避ける範囲で振幅をなるべく大きくなるよう録音レベル調整した．録音レベルを各雑音種別毎に変更するため，騒音計により録音中の騒音レベルを測定した．

2.2 収録雑音種別

我々は，表 1 に示す環境雑音例を含む計 69 種類の環境で計 50 時間以上収録した．収録雑音種別として，日常生活に関連のある交通関連，商業関連，オフィス関連の雑音と，工業関連，その他，アミューズメント，自然に関連したものを屋内，屋外問わず，幅広い範囲で収録した．

2.3 収録時間と条件

各雑音種別内での状況変化に関しては、雑音環境でおきる雑音イベントが短いものは 1 秒以内ものから，長いものでは数分に及ぶことが考えられ，それよりも長く収録する必要があることから，最低 30 分以上連続で収録した．また，同一雑音種別において，条件が異なる場合の収録を行った．たとえば，車内雑音に関して，天候，経路，窓の開閉，車種 (乗用車，トラック，バスなど) の条件が異なる場合を，展示会場ではブース内，通路と場所毎の収録，道路工事では工事行程毎の収録、サイレンは消防車，パトカー，救急車と異なる種類のサイレンの収録を行っている．

3 雑音重畳認識評価フレームワーク

本節では，収録した ATRANS を雑音重畳認識評価に利用する方法を記述する．

3.1 周波数特性の変更

本稿では，ATRANS を AURORA-2J のクリーン音声に計算機上で重畳させて音声認識実験を行う．ATRANS は 2.2 節で述べたように，多様な目的へ利用可能にするため，コンデンサマイクロホン DPA 社製 4060 を用いて収録されている．一方，AURORA-2J は，雑音下連続日本語数字

音声認識タスクの共通フレームワークであり，音声認識性能評価を目的としているため，音声認識システムでよく利用されるヘッドセットタイプのダイナミックマイクロホン (Sennheiser 社製 HMD410-6) を用いて音声収録されている．従って，AURORA-2J の音声信号と ATRANS の雑音信号は，各々のマイクロホンの周波数特性，特に有効周波数帯域に依存したものとなっている．異なる特性を持つ信号を重畳すると，信号レベルを計算する周波数帯域幅の違いから，有効周波数帯域の広い雑音信号レベルは，音声信号のレベルより大きくなり，その結果所望の SNR に対して，小さなレベルで重畳してしまう問題が生じる．

筆者らは，雑音信号が音声信号と同等の特性となることを目的とした特性補正フィルタ $W(z)$ を作成した．具体的には，2^{17} の長さの Swept sine 信号を用いて，各々のマイクロホンの含んだ音響システム伝達関数 $H_s(z)$，$H_n(z)$ を測定し，これらの伝達関数から，特性補正フィルタ $W(z)$ は式 (1) で求めた．

$$W(z) = \frac{z^{-m} \cdot H_s(z)}{H_n(z)} \tag{1}$$

ここで，z^{-m} は因果性，安定性を満たすための遅延である．なお，$H_s(z)$，$H_n(z)$ はアンプ，ラウドスピーカの特性および空間特性も含むが，それらはキャンセルされ，特性補正フィルタ $W(z)$ には，マイクロホンの周波数特性差のみ含まれる．特性補正フィルタを図 1 上に，周波数振幅特性を図 1 下に示す．一例として，オフィスの居室雑音を DPA 社製 4060 マイクロホンと Sennheiser 社製 HMD410-6 マイクロホンで同時録音したスペクトルと，前者のスペクトルの特性を補正した雑音信号のスペクトルを図 2 に示す．図 2 より，特性補正された雑音スペクトルが，Sennheiser 社製 HMD410-6 マイクロホンで収録した雑音スペクトルに，音声認識に用いる周波数帯域内で非常に近づいていることが分かる．

3.2 雑音重畳

図 3 に，雑音重畳の機能ブロック図を示す．音声信号の無音区間の長さによって重畳する雑音信号のレベルが変化しないように，レベル計算では，音声信号の有音，無音ラベル情報を用い，音声信号の有音区間のみのレベルを計算する．また，雑音信号のレベル計算では，重畳する音声信号の有音区間と重なる区間のみのレベル計算を行う．そして，所望の SNR になるように雑音信号の重畳係数 α を決定し，加算する．

4 雑音重畳音声認識実験

4.1 音声認識実験概要

ATRANS の雑音信号にから，たとえば，駅ホーム雑音では，電車が発車停車している部分，アナウンスの流れている

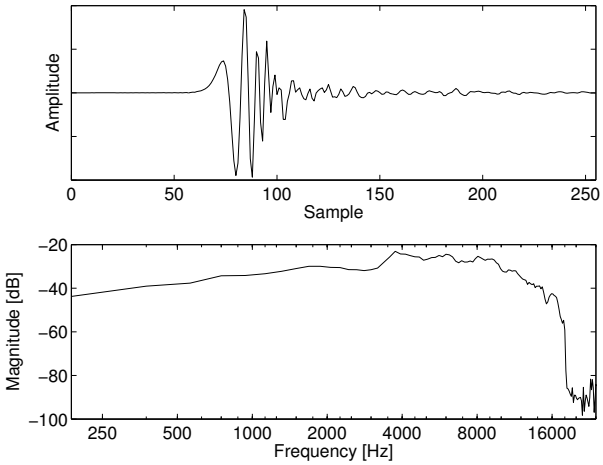


図 1: 特性補正フィルタ

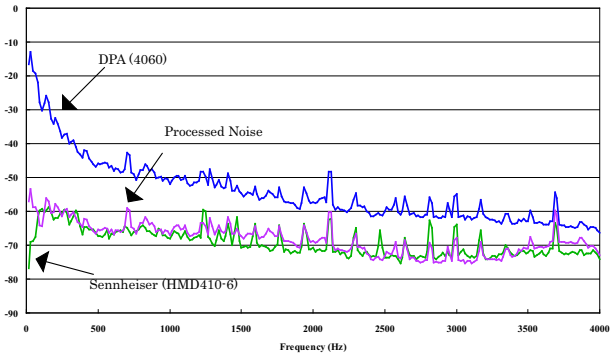


図 2: オフィスの居室雑音スペクトラム

部分，静かな時のように状況を細かく分け，20 秒の長さの雑音信号を作成した．その後，切り出した雑音信号に 3 章で示した処理を施し，日本語連続数字 DB である AURORA-2J のうち，地下鉄雑音セットに用いられたクリーンな発話データに重畳し，音声認識実験を行った．

音声認識エンジンのフロントエンド部は，分散型音声認識フロントエンド [10] (以下 DSR と記述する) と 2 段ウイナーフィルタによる雑音除去機能を有する拡張分散型音声認識フロントエンド [11] (以下 ADSR と記述する) を用い，8kHz サンプリング，16kHz サンプリングの両方のモードで行った．また，クリーン音声信号により音響モデルを学習するクリーン学習の他，認識性能の上限値の目安となるよう，テストデータと同一種別でかつ同じ時間長の雑音が重畳された音声信号により音響モデルを学習するマッチド

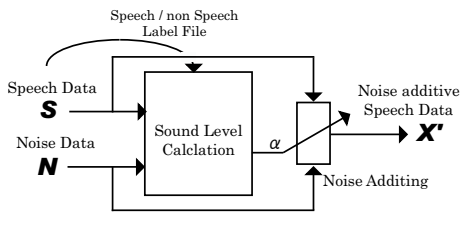


図 3: 雑音重畳機能ブロック図

表 2: 音声認識実験に使用した雑音種別

駅改札口，駅ホーム (アナウンス時，電車発車時)，駅前広場，街道，空港 (アナウンス時，ロビー，付近)，在来線車内，車内 (高速道路 (晴，雨，車種を変えて)，一般道 (晴，雨，車種を変えて)，トンネル走行，アイドリング)，産業用道路，新幹線車内，地下鉄 (車内，ホーム)，トラック (窓 (開，閉) で加速時，高速走行時)，バス (加速時)，バスターミナル，飛行機内，リムジンバス，居酒屋，市場，駅前広場，エレベータホール (デパート，病院)，展示会場 (ブース内，通路)，電話ボックス内 (車道付近，人混み付近)，地下通路，デパート食品売場，ファーストフード店，ホテルのロビー，マーケット (レジ付近)，レストラン，オフィス (受付，居室，エレベータホール，マシンルーム)，建築工事，板金工場 (金属打撃音，金属切断音)，道路工事 (切断，破砕，舗装)，物流センタ，ボイラー室，競技場 (ラグビー)，ゲームセンタ，サイレン (救急車，消防車，パトカー)，スポーツジム，田圃，体育館 (バスケット)，ビリヤード場，ボウリング場，祭り，森 (蝉)

学習音響モデルを使用した場合の実験を行った．その他の実験条件は，参考文献 [2] と同様である．また，音声認識実験で使用した雑音種別は，67 種類で表 2 に示した．

4.2 平均認識率

図 4，図 5 に，それぞれ 8kHz，16kHz サンプリングの場合の平均認識率 (縦棒) と標準偏差 (縦線) を示す．ここで平均と標準偏差は，表 2 に記した雑音種別を重畳して得た認識結果の平均と標準偏差を表す．

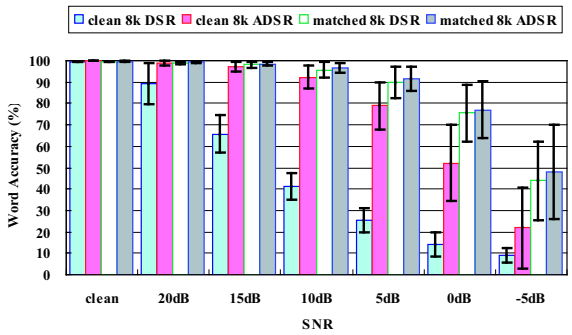


図 4: 平均認識率と標準偏差 (8kHz サンプリング)

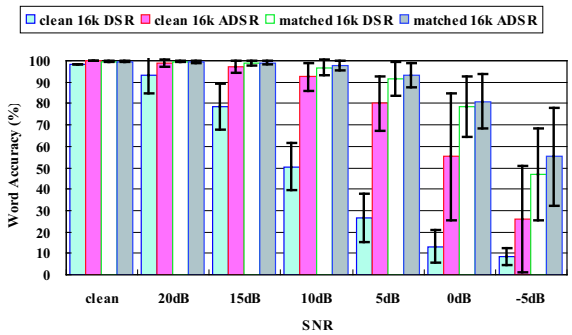


図 5: 平均認識率と標準偏差 (16kHz サンプリング)

DSR と ADSR の場合の認識結果の比較：

- ADSR の場合 DSR の場合と比較し，雑音除去機能により平均認識率が高い．
- 特に，クリーン学習の場合に ADSR の認識率が大きく向上する．但し，SNR が低い場合には標準偏差も大きくなり，雑音除去の有効性が雑音種別により偏っていることが分かる．詳細は 4.2.2 に記す．

クリーン学習とマッチド学習の認識結果：

- DSR の場合，マッチド学習の認識率が非常に大きくなる．SNR が低い場合には標準偏差も大きくなるものの．すべての雑音種別で概ね認識率が高くなった．詳細は 4.2.4 に記す．
- ADSR の場合も，特に SNR が 5dB 以下でマッチド学習の認識率が非常に高くなる．
- マッチド学習の場合，DSR と ADSR，8kHz サンプリングと 16kHz サンプリングの場合の平均認識率の差は小さくなる．

8kHz サンプリング音声と 16kHz サンプリング音声による認識結果の比較：

- 8kHz サンプリング音声に比べ 16kHz サンプリング音声の場合の方が，認識率は高い．
- クリーン学習の DSR の場合は，SNR が 5〜15dB の時に特に標準偏差が大きくなる．各雑音種別とも，認識率は高くなるものの，程度にばらつきがあった．詳細は，4.2.1 に記す．
- クリーン学習の ADSR の場合，SNR が 5dB 以下の時に特に標準偏差が大きくなり，非定常雑音の一部の認識率が下がった．これは，4k〜8kHz の周波数帯域において，音声信号に比べ雑音信号のレベルが非常に大きかったことが考えられる．具体的な雑音種別の結果は，4.2.2 で記す．

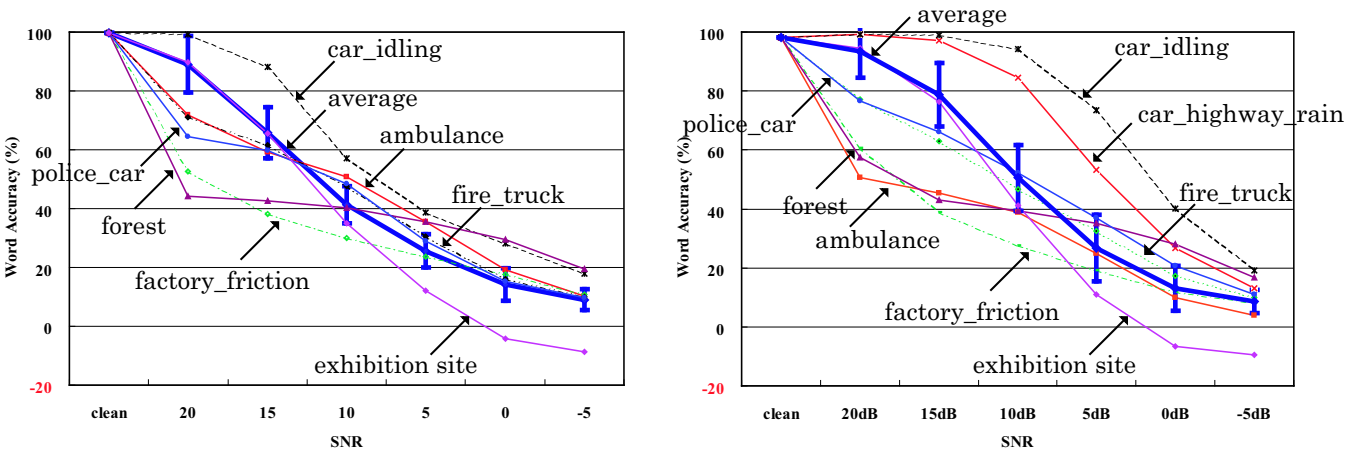
以下，クリーン学習 DSR，クリーン学習 ADSR，マッチド学習 DSR，マッチド学習 ADSR の場合の認識結果のうち平均認識率から大きく異なる結果のものを詳細に分析した．

4.2.1 クリーン学習モデル DSR の音声認識結果

図 6 に，クリーン学習モデル DSR の認識結果を示す．図 6 の中で，太線は表 2 に示した雑音種別の認識結果の平均と標準偏差である．8kHz，16kHz サンプリングどちらの場合も雑音種別毎の傾向は似ているが，後者の方がよりばらつく傾向がある．SNR が高い場合でも認識率が低いものとしては，非定常雑音である消防車 (fire_truck)，救急車 (ambulance)，パトカー (police_car)，そして比較的定常的な森 (forest)，金属切断音 (factory_feiction) がある．SNR が低い時に認識率が低いものとして，展示会場 (exhibition_site) のような背景発話があり，逆に高くなるものとして，定常的な雑音である車内騒音 (car_idling など)，森 (forest) がある．

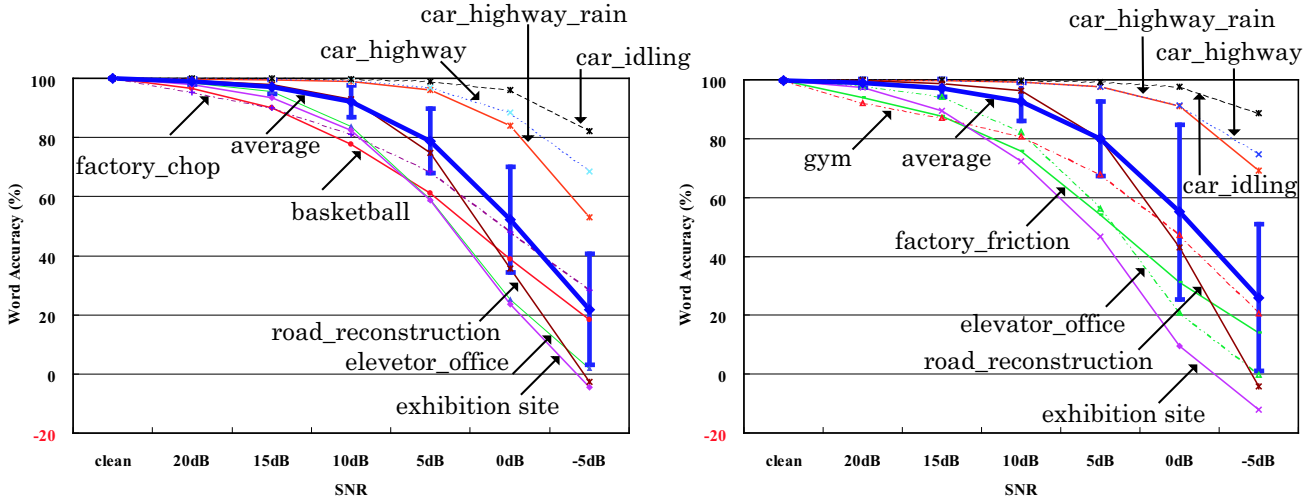
4.2.2 クリーン学習モデル ADSR の音声認識結果

図 7 にクリーン学習モデル ADSR の認識結果を示す．同じく太線は平均認識率と標準偏差である．比較的定常な車内騒音 (car_idling) 等は認識率が特に高い．逆に，体育館 (bascket)，金属打撃音 (factory_chop)，展示会場 (exhibition_site)，オフィスのエレベータホール (elevator_office)，



(a) 8kHzサンプリング　　(b) 16kHzサンプリング

図 6: 単語認識率 (クリーン学習モデル，DSR)



(a) 8kHzサンプリング　　(b) 16kHzサンプリング

図 7: 単語認識率 (クリーン学習モデル，ADSR)

道路舗装工事 (road_reconstruction) などの非定常雑音の認識率が低い．よって，ADSR は一部の非定常な雑音に対してまだ不十分であるものの，定常雑音の除去には有効であると考えられる．

4.2.3　マッチド学習モデル DSR の音声認識結果

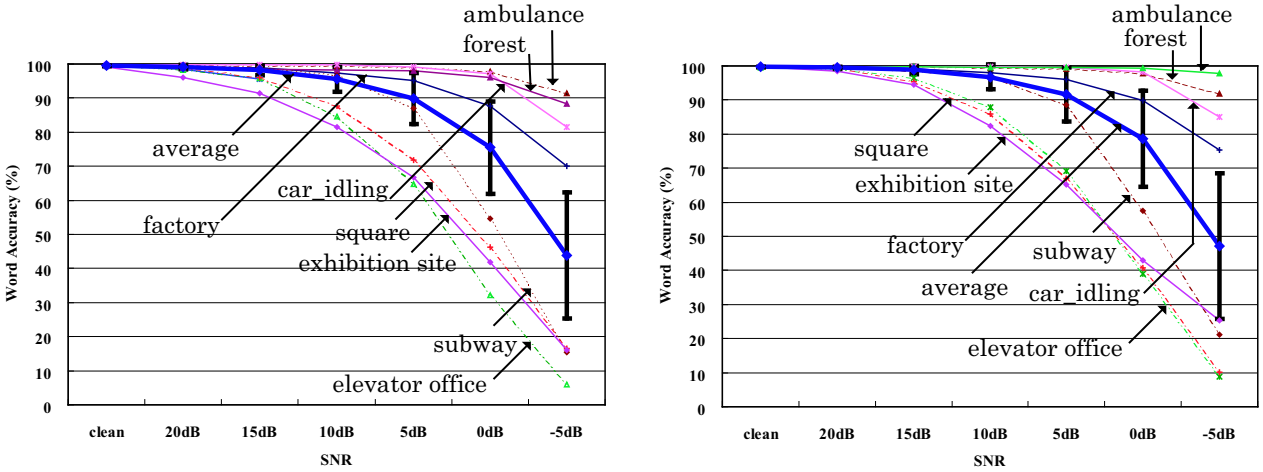
図 8 にマッチド学習モデル DSR の認識結果を示す．同じく太線は平均認識率と標準偏差である．ADSR による雑音除去の場合と認識結果の傾向は異なり，消防車 (fire_truck)，救急車 (ambulance)，パトカー (police_car)，金属打撃音，打撃音 (factory)，森 (forest) のような一部の非定常雑音の認識率が高く，逆に展示会場 (exhibition_site)，駅前広場 (square) のような背景発話雑音と，オフィスのエレベータホール (elevator_office)，地下鉄雑音 (subway) のような非定常雑音の認識率が低い．

4.2.4　マッチド学習モデル ADSR の音声認識結果

図 9 にマッチド学習モデル ADSR の認識結果を示す．同じく太線は平均認識率と標準偏差である．マッチド学習モデル DSR の場合と比較して，車内雑音 (car_idling 等) や森 (forest) のような定常的な雑音の認識率も高くなり，全体として平均認識率が上がっている．

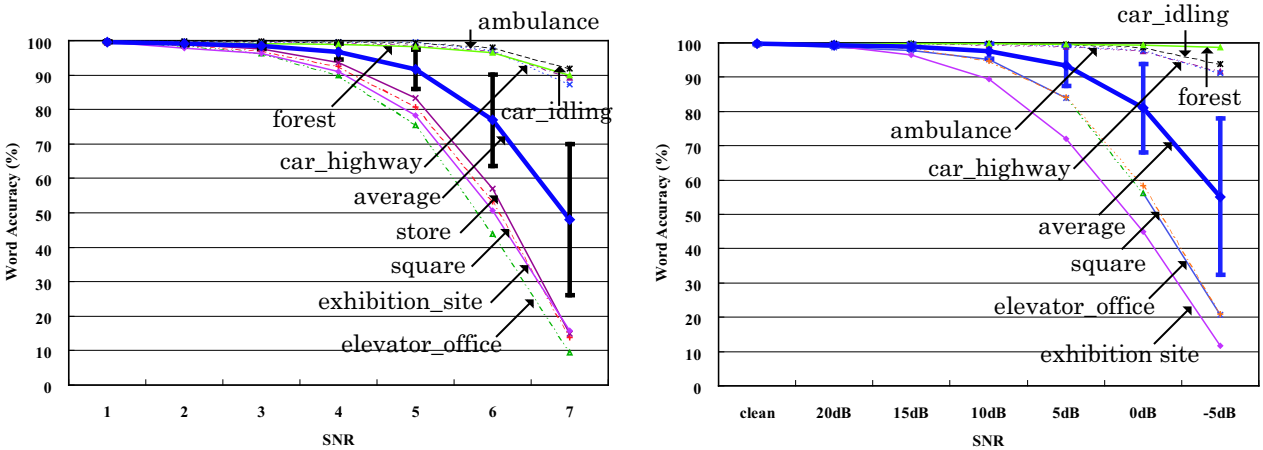
5　まとめ

本稿は，実環境雑音 DB である ATRANS と日本語連続数字タスクの AURORA-2J の音声信号のように，重畳する雑音信号と音声信号の収録時のマイクロホンが異なる場合に，周波数特性 (特に有効周波数帯域) の差異を補正する手法を示した．その上で，計算機上で雑音重畳した音声に DSR，ADSR フロントエンド処理を施し，音声認識実験を行った．その結果 ADSR による雑音除去では，特に車内雑音のような定常雑音に有効であることと，マッチド学習がサイレンのような非定常雑音にも有効であることを示した．逆に，エレベータホールや地下鉄車内のような非定常雑音と駅前広場や展示会場のような背景発話の問題が残っていることを示した．



(a) 8kHzサンプリング　(b) 16kHzサンプリング

図 8: 単語認識率 (マッチド学習モデル，DSR)



(a) 8kHzサンプリング　(b) 16kHzサンプリング

図 9: 単語認識率 (マッチド学習モデル，ADSR)

謝辞　本研究は，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．

参考文献

[1] 情報処理学会 雑音下音声認識評価ワーキンググループ Web site，
http://sp.shinshu-u.ac.jp/AURORA-J/

[2] 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, pp.535-544, Mar. 2005.

[3] 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，“実走行車内単語音声データベース CENSREC-3 と共通評価環境の構築，”情処研報 Vol.2005, No.12, 2005-SLP-55 (8), pp.41-46, Feb. 2005.

[4] 電子協騒音データベース Web site，
http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html

[5] NoiseX92 Wab site,
http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[6] Kenji Kurakata, Kazuma Matsushita and Yasuo Kuchinomachi, “Database of Domestic Sounds for Evaluation of Auditory-signal Audibility : JIS/TR S 0001, ”ASJ, Vol.24 (2003), No. 1, pp.23-26, Jan. 2003.

[7] RWCP 実環境音響音声 DB Web.site,
http://tosa.mri.co.jp/sounddb/index.htm

[8] 小川 峰義，高橋 玲，“環境要因の評価に用いる騒音データベースの構築，”ASJ 秋期講演論文集, 3-6-23, pp.371-372, Sep. 1996.

[9] 遠藤 俊樹，中村 哲，“実環境騒音 DB の収集および DSR フロントエンドによる音声認識実験，”ASJ 秋期講演論文集, 1-P-13, pp.187-188, Sep. 2004.

[10] ETSI ES 201 108 V1.1.3, “Speech processing, Transmission and Quality aspects (STQ), Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithm,” Sep. 2003.

[11] ETSI ES 202 050 V1.1.3, “Speech processing, Transmission and Quality aspects (STQ), Distributed Speech Recognition; Advanced front-end feature extraction algorithm; Compression algorithm,” Nov. 2003.