2008－SLP－73 (8)
2008／10／24

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 (3)

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[7] 傳田遊亀[8] 藤本雅清[9] 田村哲嗣[10] 松田繁樹[11]
小川哲司[12] 黒岩眞吾[13] 武田一哉[1] 中村 哲[11]

[1]名古屋大学 [2]筑波大学 [3]神戸大学 [4]徳島大学 [5]豊橋技術科学大学 [6]立命館大学 [7]近畿大学
[8]村田機械 [9]NTT CS 研 [10]岐阜大学 [11]NiCT/ATR [12]早稲田大学 [13]千葉大学

[1]E-mail: kitaoka@nagoya-u.jp

概要 我々雑音下音声認識評価ワーキンググループは，2001年10月から情報処理学会音声言語情報処理研究会の下に組織され，数多く研究されている雑音下の音声認識手法を容易に評価・比較可能な標準評価基盤 CENSREC シリーズの開発・配布を行ってきた．本稿ではその CENSREC シリーズを概観し，また主な音声認識研究の発表の場である日本音響学会全国大会および IEEE ICASSP の発表件数調査を踏まえて，その位置づけを確認する．最後に，今後の展望について述べる．

Progress Report of SLP Noisy Speech Recognition Evaluation WG: Individual evaluation framework for each factor affecting recognition performance (3)

Norihide Kitaoka[1], Takeshi Yamada[2], Tetsuya Takiguchi[3], Satoru Tsuge[4], Kazumasa Yamamoto[5], Chiyomi Miyajima[1], Takanobu Nishiura[6], Masato Nakayama[7], Yuki Denda[8], Masakiyo Fujimoto[9], Satoshi Tamura[10], Shigeki Matsuda[11], Tetsuji Ogawa[12], Shingo Kuroiwa[13], Kazuya Takeda[1], and Satoshi Nakamura[11]

[1]Nagoya Univ. [2]Univ. of Tsukuba [3]Kobe Univ. [4]Univ. of Tokushima [5]Toyohashi Univ. of Tech. [6]Ritsumeikan Univ. [7]Kinki Univ. [8]Murata Machinery [9]NTT CS Lab. [10]Gifu Univ. [11]NiCT/ATR [12]Waseda Univ. [13]Chiba Univ.

[1]E-mail: kitaoka@nagoya-u.jp

Abstract We organized a working group under Special Interest Group of Spoken Language Processing in Information Processing Society of Japan have developed evaluation frameworks of noisy speech recognition (CENSREC series) with which one can evaluate his/her own noise-robust speech recognition method and compare it with the others. In this report, we introduce the series and then review the history of the noisy speech recognition researches in ASJ and ICASSP and view the roles of our works in the history. Finally we discuss the future directions.

1 はじめに

音声認識技術の真の実用化のためには，特に実環境に存在する音響的外乱（加法性雑音，乗法性雑音）の問題に対処する必要がある．単純なコマンド語の認識であっても，雑音下で安定して高い認識性能が得られるのであれば，様々なアプリケーションへの応用が可能であり，音声認識の普及に弾みがつくと期待できる．

従来，雑音下音声認識のための手法が数多く提案されてきた [1]．しかし，これらの手法の性能を評価する際には，独自に生成・収録したデータ，独自に構築した認識システムを用いていることが多く，手法間の性能比較をすることは容易ではなかった．80年代の DARPA プロジェクトの例からも分かるように，効率的かつ効果的な研究開発を実施するためには，種々の手法を客観的に比較評価し，また競争を促すための共通評価環境が必要不可欠である．こうした問題意識のもと，我々は 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，7 年間に渡って様々な評価環境の構築・配布を行ってきた [2, 3, 4, 5, 6, 7, 8]．

当初，我々は欧州の AURORA プロジェクトと連携する形で活動を開始した．まず加法性雑音を対象とする連続数字認識タスクの AURORA-2J（後に CENSREC-1 と改名），次に自動車内の実発話を対象とする連続数字認識タスクの CENSREC-2 と孤立単語認識タスクの CENSREC-3 を構築・配布した．これらは AURORA プロジェクトの AURORA-2，AURORA-3 に対応する [9]．AURORA プロジェクトはその後，大語彙連続音声認識を対象とする AURORA-4 [9] を発表したものの，AURORA-2 と比べて認識タスクが難しくなったのみであり，解くべき問題は本質的に同じであった．このことから我々はその採用を見送ることに決めた．認識性能を劣化させる要因は複数あり，最終的にはそれらの問題を同時に解決することが求められる．しかし，加法性雑音の問題さえ完全に解決できていない現状では，これは問題設定として適切ではない恐れがある．将来性のある新しいアイデアがあっても，何らかの欠点があるがために公表を差し控えるということにもなりかねない．このような議論のもと，CENSREC-3 以降は，認識性能の劣化要因を個別に取り上げ，各々に対する重点的な研究開発の実施を促すという独自の方針で活動を進めてきた．

本稿では，まず我々が雑音下音声認識のための共通評価環境として構築・配布している，CENSREC（Corpus and Environment for Noisy Speech RECognition）シリーズの概要を述べる．次に，CENSREC シリーズの利用状況を報告すると共に，雑音下音声認識の最近の研究動向について概観する．最後に，今後の展開について述べる．

2 CENSREC シリーズの概要

2.1 CENSREC-1（AURORA-2J）

CENSREC-1 [10] は，AURORA-2 の日本語版に相当するものであり，加法性雑音を対象とする評価環境である．2003 年 7 月より 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．

本評価環境は，学習データ，テストデータ，学習・認識用スクリプト，評価用ツール，評価用 Excel スプレッドシートからなり，利用者はすぐに実験を開始することができる．このことは他の CENSREC シリーズについても共通の特徴である．認識タスクは連続数字認識である．音響モデルの学習方法としては，Clean training（クリーン音声による学習）と Multicondition training（雑音重畳音声による学習）の 2 通りを用意している．後者の場合，4 種類の雑音（Subway，Babble，Car，Exhibition）を 5 通りの SNR レベル（Clean，20dB，15dB，10dB，5dB）で重畳した音声データを用いている．テストデータは 8 種類の雑音を 7 通りの SNR レベル（Clean，20dB，15dB，10dB，5dB，0dB，−5dB）で重畳しており，Multicondition training の場合は半分が既知雑音，残りが未知雑音となる．

2.2 CENSREC-2

CENSREC-2 [11] は，AURORA-3 の日本語版に対応するものであり，自動車内の実発話を対象とする評価環境である．本評価環境は 2005 年 12 月より配布を行っている．

認識タスクは CENSREC-1 と同じ連続数字認識である．自動車内で接話マイクロホンと天井に設置した遠隔マイクロホンにより収録した音声データを用いている [12]．収録条件は，3 種類の走行速度（アイドリング，市街地低速走行，高速走行），4 種類の車内環境（通常走行，エアコン On，オーディオ On，窓開）を組合せた 11 種類である．認識条件は，学習データとテストデータの収録条件（マイク，収録環境）の一致の程度に基づき，以下の 4 種類を設定している．

Cond. 1: マイク，収録環境が共に一致

Cond. 2: マイクが一致，収録環境が相違

Cond. 3: マイクが相違，収録環境が一致

Cond. 4: マイク，収録環境が共に相違

2.3 CENSREC-3

CENSREC-3 [13] も自動車内の実発話を対象とする評価環境であり，2005年2月より配布を行っている．

認識タスクは孤立単語認識である．収録条件は上述したCENSREC-2と同じであり，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，市街地低速走行，高速走行），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種類である [14]．学習データとテストデータの収録条件の一致の程度に基づき，AURORA-3のWell-matched condition（WM），Moderate-mismatched condition（MM），High-mismatched condition（HM）に準じた以下の6種類の評価環境を設定している．

Cond. 1, 2, 3: マイク，収録環境が共に一致（WM）

Cond. 4: マイクが一致，収録環境が相違（MM）

Cond. 5, 6: マイク，収録環境が共に相違（HM）

2.4 CENSREC-4

CENSREC-4 [15] は，認識性能の劣化要因のうち，特に残響を対象とする評価環境であり，2008年3月より配布を行っている．

認識タスクはCENSREC-1と同じ連続数字認識である．音声データは，基本データセットと追加データセットに大別される．基本データセットは，残響の影響のみを受けた音声データからなる．具体的には，CENSREC-1の音声データ（ただしサンプリングレートは16kHzである）に，8種類の環境（オフィス，エレベータホール，車内，リビングルーム，ラウンジ，和室，会議室，浴室）で測定したインパルス応答を畳み込むことにより生成した．残響時間は，車内の0.05秒からエレベータホールの0.75秒までをカバーしている．本評価環境では基本データセットに対する認識性能の比較評価を主眼としており，実験・評価用のツールは基本データセットに対してのみ提供される．一方，追加データセットは，残響と加法性雑音の影響を同時に受けた音声データからなり，上記の基本データセットに各環境で収録した背景雑音を重畳して生成したシミュレーションデータセット，各環境で人間が実際に発話した音声を収録した実環境データセットを含む．開発した残響対策手法が加法性雑音に対してどの程度ロバストであるのかを検証することなどに利用できる．

2.5 CENSREC-1-C

ここまで述べた評価環境では，認識対象となる個々の発話が正確に切り出されていることを前提としている．しかし，実際には認識対象となる発話を自動的に抽出する必要があり，その精度によって認識性能は大きく変動する．CENSREC-1-C [16] は，このような音声区間検出技術を対象とする評価環境である．本評価環境は2007年9月より配布を行っている．

本評価環境は，テストデータ，評価用ツール，評価用Excelスプレッドシートからなる．認識処理を対象としていないので学習用データや学習・認識用スクリプトは含まれていない．音声データとしては，計算機上で雑音を加算したシミュレーションデータと実環境で収録した実発話データの両方を用意している．いずれも連続数字を繰返し発話している状況を想定している．シミュレーションデータでは，CENSREC-1における同一話者の9〜10個の発話（音声データ）を接続している．接続の際には，CENSREC-1の音声データ中の1秒の無音を発話間に挟んでいる．実環境データでは，2種類の環境（レストラン，高速道路脇），2種類のSNRレベル（雑音レベルの高い時間帯と低い時間帯に相当）を設定し，10名の話者が50cm離れたマイクロホンに向かって発話したものを収録した．各話者は連続数字を2秒程度の無音を挟んで8〜10回断続的に発話しており，これを1つの音声データとして収録している．話者毎の音声データ数は1条件あたり38〜39個である．1名の話者は発話の仕方が意図しないものであったので，9名による全1,380発話をテストデータとした．性能評価のために，一般的なフレーム単位での検出性能を測る尺度，音声認識を指向した発話単位での検出性能を測る尺度を定義した．

3 研究動向から見たCENSREC

上記のように，CENSREC-1〜4のような(広義の)雑音環境を模擬あるいは実収録したデータに基づく音声認識そのものの評価環境と，CENSREC-1-Cの

ような，違ったアプローチからの雑音環境への対応を指向したもの，両者を視野にいれてこれまで開発を行ってきた．これについては「環境」と「対処のアプローチの面で検討」と一昨年度の報告 [7] において述べていることである．そして，シミュレーションと実環境の違い，タスクの違いについての比較も可能であることも念頭に入れたトータル設計としている．これらの設計指針を，これまでの実際の研究動向を踏まえて振り返ってみる．

3.1 日本音響学会全国大会およびICASSPに見る雑音下音声認識研究動向とCENSREC

図1に，日本音響学会全国大会の音声Aセッションの，また図4に，ICASSPの音声部門の，過去5年分の雑音下音声認識で対象とされている雑音の推移を示す[1]．常に加法性の雑音が多く対象とされていることが見て取れる．図2に，音響学会でのCENSRECシリーズの使用された発表数を示す．実際にCENSREC-1が発表されたのはさらに1年ほど前になるが，安定して使用されてきている．

また，図1，4いずれからも，すこしずつ加法性から，その他の対象へ比率が移りつつある傾向も見受けられる．乗法性・残響や，あるいはVADによる(加法性を主とした)対処が増加傾向にある．この傾向は，2005，6年あたりから見られるように見受けられるが，CENSREC-1-Cの発表は2006年であり，また利用件数も増加傾向にあることからもよいタイミングでのリリースであったと考えている．また，CENSREC-4のリリースも，研究の増加に合わせたものとなり，タイムリーであると自負している．INTERSPEECH[15]でも興味を引いた手応えがあった．

すこし変わった傾向として，加法性と乗法性の雑音の両者に対応した手法が2006，7年頃に一旦減少していることがある．それ以前には，両者にうまく対処しようとする手法が多く提案され，一定の効果を得た(例えば[17])．しかし，我々は，本稿を含め，常々，個別要因の対処ができずにトータルの対処は難しいと述べてきた[7, 8]が，当時は一般的にもそういった認識がされていたのではないかと想像される．ただ，最近は(特にICASSPを見ると)トータル性能を求める手法への回帰が見られる．かねてより，加法性雑音による音声の変形を対数領域あるいはケプストラム領域でモデル化し，それを補正する手法が増加傾向にある(例えば[18, 19])．これらの手法はモデル化手法から考えても乗法性雑音に相性がよい．従って今後，これらの方法の成功などからも両者を同時に対処する手法が再度成功する可能性もあるとも考えられよう．

図3および図5に，音声認識を対象とした使用チャネルの推移を示す．意外なことに，あまり複数チャネルを用いた音声認識手法は増加していない．マイクロフォンアレイなどの複数センサ技術はかなり進歩したとの認識であるが，音声認識分野との融合が課題であるように感じる．Audio-visualのマルチモーダル音声認識に関しては，現在でも海外で活発に研究が行われているが，日本では比較的少ない．我々は現在も基盤整備中であるが，これがなんらかの影響を与えることができれば幸甚である．

3.2 今後の展開

前節では，外的要因が分離されて対処されるようになってきた傾向，およびまた統合されつつあるのではないかという見通しを，学会発表状況から述べた．しかし，ここまで外的な要因による音声信号への影響は加算性雑音と乗算性歪みに集約されると考えてきたが，そのほかに，加算性や乗算性とは異なる非線型な信号の変形(デジタル伝送におけるパケットロスなど)，あるいは外的な影響を受けた信号のフィードバックによる影響が存在する．

$$y = c(h \otimes l_n(x) + n)$$

(ここでc，l_nはそれぞれ(デジタル)伝送路における非線型変換，およびロンバード効果に代表される発話者による非線型変換[8])このように，確実に音声認識に影響を与えながら，研究例の少ない多くの解決すべき要素が残されていることが容易に想像できる(例えば[20])．先に述べたAVによるアプローチや，ロンバード効果の研究は，その収録の難しさ・コストなども大きな障壁になっていると思われる．こうした障壁を取り除き，活性化するに値する分野であると考えて，これらの整備を進めているところである．

4 おわりに

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

[1] ここで，乗法性と残響は一つにまとめている．インパルス応答の長短によりある程度区別されてしかるべきと考えるが，物理的現象として同等であることとあまり詳細に分類すると分かりにくくなることからここでは敢えて一つとした．

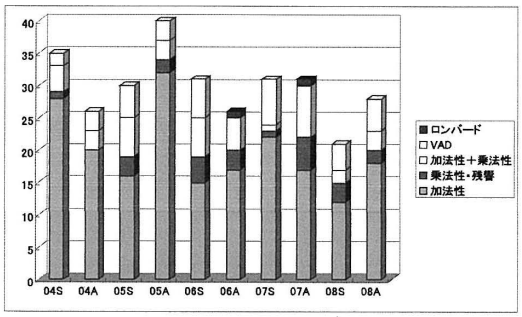

図 1: 日本音響学会における対象雑音環境の推移 (縦軸が件数，横軸は西暦に対応し，S が春季，A が秋季．)

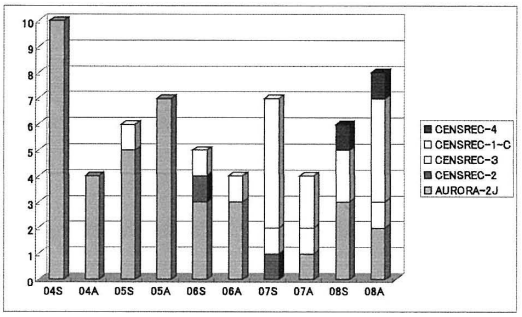

図 2: 日本音響学会における CENSREC シリーズの利用件数 (縦軸が件数，横軸は西暦に対応し，S が春季，A が秋季．)

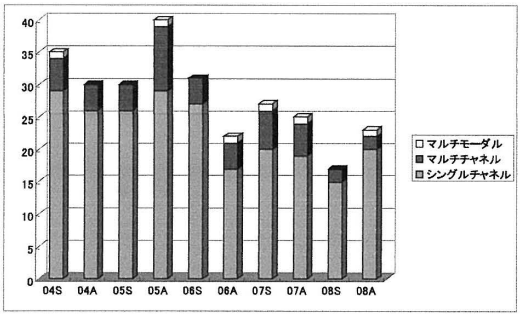

図 3: 日本音響学会における耐雑音手法の使用チャネルの推移 (縦軸が件数，横軸は西暦に対応し，S が春季，A が秋季．)

振り返り，最近の雑音下音声認識の研究動向と対比した．そして，比較的研究動向に合致したタイミング

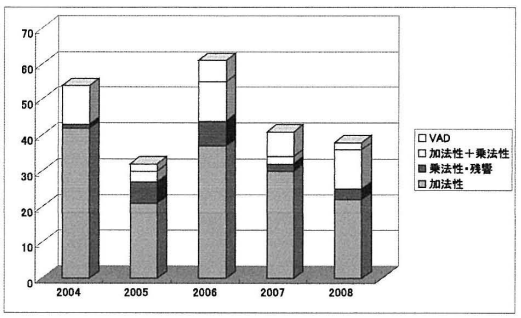

図 4: ICASSP における対象雑音環境の推移 (縦軸が件数，横軸は西暦に対応．)

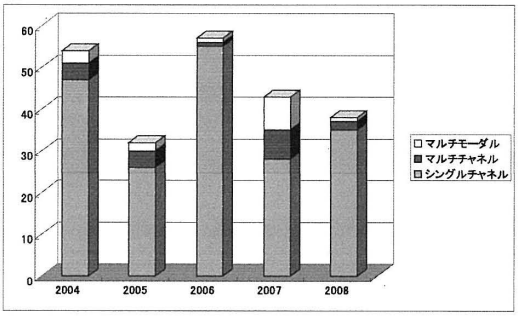

図 5: ICASSP における耐雑音手法の使用チャネルの推移 (縦軸が件数，横軸は西暦に対応．)

で開発を進め，リリースできてきたことを確認した．

我々は，雑音下音声認識への影響要因を分類して個別評価環境を提供するという方針が，その開発時点では次善であるとは考えていたが，最終的なものとして，必ずしも最善であるとは考えていない．統合的な評価環境についても考慮すべき時期がきているのかもしれないとの見解を述べた．

このような見解も視野にいれ，またこれまで/これからの CENSREC シリーズを含め，一体何が明らかになるのか，またこれから何が必要か，全体的に見直していきたい．

謝辞　本ワーキンググループの活動をご支援頂いている国立情報学研究所音声資源コンソーシアムに感謝致します．

参考文献

[1] 中村哲, "実環境に頑健な音声認識を目指して," 信学技報, SP2002-12, pp. 31–36, April 2002.

[2] AURORA-J/CENSREC Web site: `http://sp.shinshu-u.ac.jp/CENSREC/`

[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. 65–69, 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. 51–55, Feb. 2003.

[5] 中村哲, 武田一哉, 黒岩眞吾, 北岡教英, 山田武志, 山本一公, 西浦敬信, 佐宗晃, 水町光徳, 宮島千代美, 藤本雅清, 遠藤俊樹, "実環境下音声認識の評価の標準化とその動向", 音声言語シンポジウム, 2004-SLP-54-24, pp. 139–144, Dec. 2004.

[6] 中村哲, 武田一哉, 黒岩眞吾, 北岡教英, 山田武志, 山本一公, 西浦敬信, 佐宗晃, 水町光徳, 宮島千代美, 藤本雅清, 遠藤俊樹, 滝口哲也, "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—," 音声言語シンポジウム, 2005-SLP-59-24, pp. 139–144, Dec. 2005.

[7] 北岡教英, 山田武志, 滝口哲也, 柘植覚, 山本一公, 宮島千代美, 西浦敬信, 中山雅人, 傳田遊亀, 藤本雅清, 田村哲嗣, 黒岩眞吾, 武田一哉, 中村哲, "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," 音声言語シンポジウム, 2006-SLP-64-1, pp. 1–6, Dec. 2006.

[8] 北岡教英, 山田武志, 滝口哲也, 柘植覚, 山本一公, 宮島千代美, 西浦敬信, 中山雅人, 傳田遊亀, 藤本雅清, 田村哲嗣, 松田繁樹, 小川哲司, 黒岩眞吾, 武田一哉, 中村哲, "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 (2)," 音声言語シンポジウム, 2007-SLP-69-1, pp. 1–6, Dec. 2006.

[9] D. Pearce, "Developing the ETSI Aurora advanced distributed speech recognition front-end and what next?," Proc. ASRU2001, pp. 131–134, Dec. 2001.

[10] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, 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.

[11] 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, Feb. 2006.

[12] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, F. Itakura, "Construction and evaluation a large in-car speech corpus," *IEICE Transactions on Information and Systems*, Vol. E88-D, No. 3, pp. 553–561, Mar. 2005.

[13] M. Fujimoto, K. Takeda, S. Nakamura, "CENSREC-3: An evaluation framework for Japanese speech recognition in real driving-car environments," *IEICE Transactions on Information and Systems*, Vol. E89-D, No. 11, pp. 2783–2793, Nov. 2006.

[14] 武田一哉, 河口信夫, 藤井博厚, 北勝也, 板倉文忠, "走行状況別車内音声データベース," 音講論集, 3-P-10, pp. 185–186, Mar. 2002.

[15] M. Nakayama, T. Nishiura, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, T. Ogawa, S. Matsuda, S. Kuroiwa, K. Takeda, S. Nakamura, "CENSREC-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments," Proc. Interspeech2008, pp. 968–971, Sep. 2008.

[16] 北岡教英, 山田武志, 柘植覚, 宮島千代美, 西浦敬信, 中山雅人, 傳田遊亀, 藤本雅清, 山本一公, 滝口哲也, 黒岩眞吾, 武田一哉, 中村哲, "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, Oct. 2006.

[17] T. Takiguchi, S. Nakamura, K. Shikano HMM-Separation-Based Speech Recognition for a Distant Moving Speaker IEEE Transactions on Speech and Audio Processing, Vol.9, No.2, pp. 127-140, 2001

[18] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using AURORA II database and tasks," Proc. Eurospeech '01 vol.1, pp. 221–224, 2001.

[19] M. Fujimoto, S. Nakamura, "A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging, " IEICE Trans. Inf. & Syst., Vol.E89-D No.3 pp.922-930, 2006.

[20] 小川哲司, 倉持公壮, 小林哲則, "シミュレーションに基づく騒音下音声認識システム評価におけるロンバード効果の影響の検証 —複数の認識タスク, 騒音レベルに対する評価," 音講論集, 1-P-29, pp. 195–198, Sep. 2007.