日本音響学会誌 68 巻 6 号 (2012), pp. 305–310

解説

雑音下音声認識評価基盤 CENSREC*

北岡教英 (名古屋大学)**・中村 哲 (奈良先端科学技術大学院大学)***

43.72.Ne

1. はじめに

雑音下における音声認識の高精度化は常に大きな課題として多くの研究が行われてきている。こうして開発された多くの手法を客観的に性能比較できれば，手法を利用することを考えているシステム開発者にはよい指針となる。また，80 年代の DARPA プロジェクトのように，共通評価環境を用意することで客観的な比較を可能にし，効率的・効果的な研究開発や競争を促すこともできる。

本解説では，このような考えの元で組織された雑音下音声認識評価ワーキンググループ (WG) の活動として構築・配布されてきた一連の雑音下音声認識評価基盤を紹介し，それぞれの内容を概説する。そしてその他の雑音下音声認識のための共通データベースについても触れ，CENSREC 間やその他のデータベースとの関係を概観する。

2. CENSREC シリーズ

本節では，雑音下音声認識の標準評価基盤として一般に配布され利用できる，CENSREC (Corpora and Environments for Noisy Speech RECognition) シリーズの概要を紹介する。

CENSREC は，当初，欧州の AURORA プロジェクトによる評価基盤と連動して開発された。AURORA-2 [1] と呼ばれる英語連続数字認識評価基盤を日本語訳した，加法性雑音下日本語連続数字認識タスク AURORA-2J (後に CENSREC-1 と改名) が最初である。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，孤立単語認識タスクの CENSREC-3 と続く。これらは，様々なデータで個々に行われてきた音声認識システム，特に雑音下音声認識手法を，評価項目を集中して客観的に性能評価する [2] ことで研究開発にも利用者 (システム開発者) にも有用な情報を提供できることを目的に設計されている。

2.1 CENSREC-1/AURORA-2J

加法性雑音下の音声認識評価に絞って設計されている連続数字認識タスクの評価基盤である [3]。欧州で作成された AURORA-2 データベースを日本語訳した音響モデル学習データと評価データ，学習・認識スクリプトとベースライン性能を提供している。

2003 年 7 月の配布開始から約 180 部 (2012 年 3 月現在) が配布されて，現在も多くの研究機関で利用されている。

2.1.1 データ構成

1 から 7 桁の連続数字音声に AURORA-2 と同一の数種の雑音を様々な SNR レベルに調整して重畳したデータからなる。内容を AURORA-2 とできるだけ一致させるため，学習及びテストデータの構成 (男女構成，話者数など) は，AURORA-2 のものをそのまま採用している。発話内容は AURORA-2 を日本語訳したものとなっており，英語の “ゼロ” と “オー” と同様に 0 に対して “Z (ぜろ)” と “O (まる)” の二つの読みを与えている。

学習環境は 2 種類ある。一つはクリーン音声による学習 (clean training)，もう一つは雑音重畳音声による学習 (multicondition training) と呼ばれ，それぞれ 110 名 (男女各 55 名) による 8,440 発話で音響モデルを学習する。種々の雑音を想定した学習が Multicondition 学習と呼ばれ，本データは 4 種類の雑音 (Subway, Babble, Car, Exhibition) を 5 種類の SNR レベル (clean, 20 dB, 15 dB, 10 dB, 5 dB) で重畳した音声からなる。

テストデータは，Multicondition 学習では雑音が既知の場合と未知の場合を比較できるように雑音

* Corpora and Environments for Noisy Speech RECognition: CENSREC.

** Norihide Kitaoka (Nagoya University, Nagoya, 464–8603)

*** Satoshi Nakamura (Nara Institute of Science and Technology, Ikoma, 630–0101)

表–1 CENSREC-1 のテストセットの雑音環境

	雑音	フィルタ
セット A	地下鉄，ざわめき，車内，展示場	G.712
セット B	レストラン，市街，空港，駅	G.712
セット C	地下鉄，市街	MIRS

%Acc				
	A	B	C	Overall
Clean Training	79.20	77.81	75.87	77.98
Multicondition training	93.23	88.98	90.63	91.01
Average	86.22	83.39	83.25	84.49

Relative performance				
	A	B	C	Overall
Clean Training	61.12%	60.39%	51.84%	59.09%
Multicondition training	20.09%	43.79%	33.88%	36.08%
Average	40.61%	52.09%	42.86%	47.59%

図–1 CENSREC-1 の音声認識実験結果（上）とベースライン性能からの改善（下）の例

が重畳されたセットが用意されている（表–1）。発声は男女 104 名による 4,004 発話で，それらを 4 つに分けた 1,001 発話に，いずれかの雑音が 7 種類の SNR レベル（clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, −5 dB）で重畳されている。電話回線を通じた音声認識を意識しているため，フィルタ特性の違いによる影響も考慮したデータとなっている。

2.1.2 評価方法

音声認識の標準的なツールである HTK（Hidden Markov model Tool Kit）を利用して音響モデルを学習し，認識実験が行える。単語単位の HMM と 2 種類の無音（sil, sp）を作成する。特徴パラメータには MFCC（Mel-frequency Cepstral Coefficients）の 12 次元と対数パワー，その 1 次及び 2 次微分の計 39 次元である。

これらは HTK が動作する環境であれば，添付されたスクリプトで簡単に実行できる。更に結果を添付された Excel シートに入力すれば，ベースライン性能からの改善が容易に分かる（図–1）。

2.2 CENSREC-2

2005 年 12 月に配布が開始された，自動車内の連続数字発声データである [4]。雑音下で実際に発声されたものである点に特徴がある。

2.2.1 データ構成

名古屋大学の実験車両で実際に走行して，車室内で収録された [5]。3 種類の走行（アイドリング，市街地走行，高速走行）と 4 種類の車内環境（通常走行，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，ヘッドセットによる接話マイクロホンと，天井に取り付けられた遠隔マイクロホンで録音した音声を用いた（文献 [5] の 1 番と 6 番）。学習データ，評価データはそれぞれ 73 名及び 31 名による発話で，話者に重なりはない。総発話数は 17,651 となっている。

2.2.2 評価方法

学習時とテスト時の収録環境の一致/不一致によって 4 条件が設定されている。雑音抑圧手法のターゲットなどによって適宜使い分けることができる。学習/評価スクリプトで容易に評価でき，Excel シートも用意されている。

Condition 1: マイク種別，収録環境共に一致
Condition 2: マイク種別一致，収録環境相違
Condition 3: マイク種別相違，収録環境一致
Condition 4: マイク種別，収録環境共に相違

2.3 CENSREC-3

2005 年 2 月から配布されている CENSREC-3 は，実走行車内での孤立単語音声認識の評価環境である [6]。CENSREC-2 とは音声認識タスクが異なり，両者を用いることでタスクの違いによる影響を評価できる。

CENSREC-2 と同様に収録されており，男女各 202 名，91 名による音素バランス文 14,050 発話から音素 HMM を学習する。そして 3 種類の走行速度（アイドリング，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 [7] で，カーナビゲーションシステムへの入力を想定した 50 語を男性 8 名女性 10 名が発話した，総数 14,216 発話で評価する。

学習データと評価データの収録環境の一致/不一致により 6 条件が設定されている。

Condition 1, 2, 3 マイク種別，収録環境共に一致（マイク種別により 1，2，3 と分類）
Condition 4 マイク種別一致，収録環境相違
Condition 5, 6 マイク種別，収録環境共相違

2.4 CENSREC-1-C

これまでは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であった。しかし実環境では，それ以外に独立に手法として評価すべきこともある。

そのひとつに音声区間検出（Voice Activity Detection; VAD）がある。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。そこで VAD 評価環境として CENSREC-1-C (CENSREC-1-Concatenated) [8] を構築した。

2.4.1 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 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。

雑音加算によるシミュレーションデータは CENSREC-1 の同一話者による 9 ないし 10 の音声データを接続することにより，連続的な発話のデータを計 164 データ作成した。雑音環境も CENSREC-1 のセット A, B と同様である。ただし，セット C は提供しない。

実環境データは，二つの雑音環境（学生食堂，高速道路付近）及び二つの 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）。男女 5 名，計 10 名が各環境 4 回ずつ，計 160 のデータとなっている。（ただし 1 名分は意図どおりに発話できていないデータとして評価から除外）。

2.4.2 評価方法

音声分析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。

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。以下の二つの 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

また，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。

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易いが，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，ある一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すべて誤検出とする。

ベースライン VAD として，音声パワーに基づく方法による結果を公開している。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。

2.5 CENSREC-4

CENSREC-4 は，様々な残響環境における遠隔発話の音声認識の評価を目的とする評価環境であり [9]，2008 年 3 月に配布が開始された。CENSREC-4 は，これまでの CENSREC-1 と同様に連続数字発話で構成されており，大きく二つのサブセットに分けることができる。一つは「基本データ」であり，もう一つが「追加データ」である。基本データに対してのみ，HTK による HMM 学習/認識スクリプトが提供される。

2.5.1 基本データ

種々の残響環境をシミュレーションするために，幾つかのインパルス応答を収録した。8 種類（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）の環境で収録したインパルス応答を用いている。

多くの環境ではマイクロホンとマウスシミュレータの間の距離を 0.5 m（車環境では 0.4 m，浴室環境では 0.3 m）とした。各環境の収録時における残響時間（T_{60}）は 0.05 秒〜0.75 秒である。

CENSREC-1 作成時に収録された防音室内での連続数字発話に上記のインパルス応答を畳み込むことによって残響環境をシミュレーションしたデータセットを提供する。ただし，CENSREC-1 とは異なり 16 kHz サンプリングである。

基本データと呼ばれるテストセットはセット A（オフィス，エレベータホール，車内，リビングルーム）とセット B（ラウンジ，和室，会議室，浴室）の二つのサブセットからなる。CENSREC-1 と同様に，CENSREC-4 では，セット A で用いたインパルス応答は提供される学習データの作成にも用いられるが，セット B で用いられるものは学習データには用いない。

2.5.2 追加データ

この他，追加データと呼ばれるデータがある。追加データには，残響のインパルス応答の畳み込みと加算性雑音の重畳の両方が施されたデータが収録されている。まず室内のインパルス応答を畳み込んだ後，そのインパルス応答を収録した場所で収録された雑音を，SNR を調整して加算することで作成されたものがセット C と呼ばれ，これにはセット A 中の 2 環境とセット B 中の 2 環境が含まれている。

更に，実際に残響及び加算性雑音が存在する環境において発声されたデータも収録されている。

実環境データは，マウスシミュレータを用いずに人間により発声された音声が，近接マイク及び遠隔マイクの二つのマイクロホンで同時に録音されている。このデータセットをセット D と呼ぶ。収録環境はセット C のものと同一であり，各環境 10 名（男性 5 名，女性 5 名）が連続数字を発話している。発声されたデータはテストデータ（49 ないし 50 発話）と学習データ（11 データ，適応用）からなり，全部で 2,536 発話（2,536 ファイル）となっている。

2.5.3 評価方法

これまでの CENSREC と同様に，学習/認識/評価ツールを提供しており，残響未対応の認識結果をベースラインとしてそれからの改善が容易に分かるようになっている。ただし，CENSREC では音声認識への影響要因の個別評価を目的としていることもあり，現時点では複合的な要因を持つセット C とセット D については評価未対応であることには注意されたい。

2.6 CENSREC-1-AV

音響的に雑音が重畳された場合に，他の信号の情報を用いる「マルチモーダル音声認識」が有効である。中でも発声時の口唇動画像を用いる「オーディオビジュアル音声認識」が近年研究されており，その評価基盤として CENSREC-1-AV [10] は 2011 年 4 月に配布が開始された。

2.6.1 バイモーダル音声認識

よく用いられる方法として，音声と時系列画像それぞれの特徴量系列 O_A, O_V の認識用 HMM を用い，以下のようにしてそれらの重み付き尤度を求めるマルチストリーム HMM による認識がある。

$$b_{AV}(O_{AV}) = \lambda_A b_A(O_A) + \lambda_V b_V(O_V) \quad (1)$$

ここで，$\lambda_A + \lambda_V = 1$ とする。

2.6.2 データ構成

CENSREC-1 に準じた連続数字読み上げデータベースである。発話者はブルースクリーンを背景に椅子に座り，襟元にピンマイクをつけて発話する。その音声は 2 台のディジタルビデオカメラに入力されている。カメラは 1 台がカラー映像を，もう 1 台は近赤外線映像を撮影した。収録時は，音声は 48 kHz，画像は 29.97 fps インタレース映像（720×480 ピクセル）である。音声は最終的に 16 kHz にダウンサンプリングされている。また，映像は，インタレースを解除し時系列画像にした上で，発話ごとに座標を固定して 81×55 ピクセルの切出し窓を設けて口唇付近の画像を切り出した。

学習データとテストデータはそれぞれ 42 名（男性 22 名，女性 20 名）による 3,234 発話及び 52 名（男性 25 名，女性 26 名）による 1,963 発話となっており，テストデータの音声にのみ市街地道路及び高速道路走行時の雑音を CENSREC-1 と同様の SNR で重畳している。画像には乗用車内の明度値のガンマ補正で映像雑音をシミュレートした。

2.6.3 評価方法

ベースラインとして，音声のみ，映像のみ，及び式 (1) で統合した場合の認識精度が与えられている。用いた特徴量は，音声は MFCC に基づく 39 次元，画像は固有顔特徴量 [11] 10 次元とその 1 次及び 2 次微分の計 30 次元である。これらからの改善が容易に分かるようになっている。

3. 音声認識標準データベース/評価基盤

先にもふれたように，CENSREC は AURORA プロジェクトと連動して開発が始まったものである。当時，欧州の AURORA に対して米国で作成された，軍事関係の環境雑音を付加したシミュレーションデータベースである SPINE [12] も多く用いられた。それ以前から，米国では DARPA が主導して共通の評価用データベースを提供した競争型の研究開発が活発であり，読み上げ音声の TIMIT [13] や電話回線上の会話コーパス SWITCHBOARD [14] は現在でもよく用いられる。これらの中でも AURORA は，雑音下音声認識の評価に焦点を絞り，提供されたツールや整備されたデータベースによる導入と比較の手軽さが特徴である。

以前に本会誌の解説 [15] において，(広義での) 雑音下音声認識に対する様々な手法を紹介したが，その際にそれらを幾つかの音声認識への影響要因と，雑音への対処へのアプローチとにより分類した。これまでのデータベースでは，こうした分類の下で評価対象が明確にされてはいなかった。

この観点では，SPINE や AURORA，そして CENSREC-1〜4 は環境に着目した基盤である。CENSREC-1 は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。これに対し，CENSREC-2 は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，実環境とシミュレーションの差が比較できる。更に，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。また，加法性の雑音とは異なる乗法性の残響に着目したのが CESNREC-4 である。すなわち，雑音下音声認識の困難さを「雑音 (残響) の違い」「シミュレーション/実環境の違い」及び「タスクの違い」という軸で互いに比較できる環境となっている。

AURORA-3 と呼ばれるデータベースは CENSREC-2 と同様の自動車内実環境発話連続数字データである。欧州の言語資源配布機関である ELDA から，SpeechDat-Car というデータベースを基に作成され配布されている[1]。更に，AURORA-4 [16] は読み上げ音声データベース Wall Street Journal コーパスに雑音を付加したデータである。これを見ると，AURORA プロジェクトも同様な考えで開発が進められたものと推察される。しかし AURORA-4 はタスクが難しいわりに，解くべき問題が AURORA-2 と変わらないことから，実際には AURORA-2 が圧倒的に多く用いられてきている。

一方，CENSREC はアプローチに目を向けた。CENSREC-1-C は，一般的な音声認識手法の改良とは異なる，しかし現実に，実用的な雑音下音声認識で大きな役割を果たしているアプローチに着目し評価できる基盤となっている。更に CENSREC-1-AV は，もう一歩踏込み，認識手法への別のアプローチにも着目している。これらの基盤は，音響環境における観測音声信号が

$$y = h * x + n$$

(ただし x は発話の原信号，n は加法性の雑音，h は残響を含む伝達特性) と書かれるとした場合に，個々のオペレータが個別に見れば何等かの領域で線形に扱えることから，これらを分離することで個々の対処を考え，組み合わせることが有効であると考えてきた。また，CENSREC-1-C は x に対する前処理，CESNREC-1-AV は特徴量 y の拡張と考えることもできる。

ここで，実際のアプリケーションでの利用に目を向けると，加法性や乗法性とは異なる非線型な信号の変形 (ディジタル伝送のパケットロスなど) や，人間の発話時の歪 (個人の違いやロンバード効果など) も存在することが分かる。例えば中村 [17] でも，「音源」「話者及び雑音源からの音響パス」「雑音」「回線特性」に分類している。これは，以下のように書ける。

$$y = c(h * l_n(x) + n)$$

ただし，c は (ディジタル) 伝送路における非線型変換，l_n は発話者による非線型な発話変形を表す。こうした実環境の様々な影響を同時に受けた音声が最終的なターゲットになることは言うまでもない。AURORA では，複合的な雑音をシミュレートした評価基盤が AURORA-5 [18] として作成された[2]。ただ，現時点までに多く利用されているわけではない。それは，複合的な雑音にはそれぞれの雑音への対処手法の組み合わせで対応することになり，結局どの手法を評価しているのか分からなくなってしまう，といった問題があるのではないかと考えている。

そこで一つの方向として，個別要因を更に抽出して集中した対処手法の研究を促進することが考えられる。例えばロンバード効果は，音声認識への悪影響を指摘されながらそれのみを研究することが難しい対象であり，それが分離され整備されたデータは有用であろう [19]。

一方で，近年の信号処理技術と音声認識技術双方の進展に基づいて現実的なアプリケーションを指向した評価課題も設定されてきている。例えばマルチマイクロホンを入力とした室内ハンズフリー入力タスクである CHiME [20] は，これまで別々に研究が進んだ両技術を融合させるよい機会であ

[1] 欧州言語が各種含まれているが，英語がないため普及が遅れている感がある。

[2] シミュレーションであるため l_n は考慮されていない。

ると思われる。CENSREC-4でも収録はマルチマイクロホンで行っており，更に実環境で直接収録している貴重なデータがある。これらが標準的な仕様を設定して公開できると日本語での研究に役立つものとなろう。

いずれにせよ，評価する対象が明確である，あるいは技術的に狙いを持った，シンプルなデータベース/評価基盤が有効であり，またそれらの存在が技術発展の加速に大きく貢献すると考える。

4. ま と め

本解説では，雑音下音声認識評価基盤 CENSREC について，その内容を概説し，それらの設計経緯について触れた。本基盤は，音声認識技術の評価を目的に設計・作成されているものの，雑音対策は様々な音声処理で必要であるし，例えば VAD は音声通信でも重要である。また，バイモーダルは最新の画像・映像技術の導入でさらなる高度化が望める。このように，音声認識関係のみでなく他の関係各所においても利用され，貢献できることを望む。

謝 辞

本稿は，CENSREC 構築に尽力した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（H20 年度まで活動）の活動成果である。同メンバの皆様に感謝します。また，CENSREC の作成には国立情報学研究所音声資源コンソーシアム（NII-SRC）からの支援を受け，また NII-SRC から無償提供されている。NII-SRC に感謝いたします。

文 献

[1] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next," *Proc. Eurospeech 2001* (2001).
[2] 中村 哲, "音声認識の動向 [II]—音声認識性能の客観的評価に向けて—," 信学会誌, **89**, 830–835 (2006).
[3] 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 Trans. Inf. Syst.*, **E88-D**, 535–544 (2005).
[4] 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).
[5] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara and F. Itakura, "Construction and evaluation of a large in-car speech corpus," *IEICE Trans. Inf. Syst.*, **E88-D**, 553–561 (2005).
[6] M. Fujimoto, K. Takeda and S. Nakamura, "CENSREC-3: An evaluation framework for Japanese speech recognition in real driving-car environments," *IEICE Trans. Inf. Syst.*, **E89-D**, 2783–2793 (2006).
[7] 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とその予備評価," 音講論集, 3-P-10, pp. 185–186 (2002.3).
[8] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoust. Sci. & Tech.*, 30, 363–371 (2009).
[9] T. Fukumori, T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, T. Yamada, K. Yamamoto, S. Tsuge, M. Fujimoto, T. Takiguchi, C. Miyajima, S. Tamura, T. Ogawa, S. Matsuda, S. Kuroiwa, K. Takeda and S. Nakamura, "CENSREC-4: An evaluation framework for distant-talking speech recognition under reverberant environments," *Acoust. Sci. & Tech.*, 32, 201–210 (2011).
[10] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda and S. Nakamura, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," *2010 Int. Conf. Auditory and Visual Speech Processing* (2010).
[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, 3, 71–86 (1991).
[12] T.H. Crystal, A.S. Nielsen and E. Marsh, "Speech in noisy environments (spine) adds new dimension to speech recognition R&D," *Proc. HLT 2002*, pp. 212–216 (2002).
[13] P. Price, W.M. Fisher, J. Bernstein and D.S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," *Proc. ICASSP-88*, Vol. 1, pp. 6751–654 (1988).
[14] J.J. Godfrey, E.C. Holliman and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. ICASSP 92*, pp. 517–520 (1992).
[15] 北岡教英, "音声認識におけるロバストネス," 小特集—自動音声認識研究の動向と展望—," 音響学会誌, 66, 23–27 (2010).
[16] N. Parihar, J. Picone, D. Pearce and H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," *Proc. Eurospeech'03*, pp. 337–340 (2003).
[17] 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 (2002).
[18] H.G. Hirsh and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," *Proc. Interspeech2005*, pp. 2697–2700 (2005).
[19] T. Ogawa and T. Kobayashi, "Influence of Lombard effect: Accuracy analysis of simulation-based assessment of noisy speech recognition systems for various recognition conditions," *IEICE Trans. Inf. Syst.*, **E92-D**, 11, 2244–2252 (2009).
[20] H. Christensen, J. Barker, N. Ma and P. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," *Proc. INTERSPEECH2010*, pp. 1918–1921 (2010).