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

2006－SLP－63 (1)
2006／10／20

CENSREC-1-C: 雑音下音声区間検出評価基盤の構築

北岡 教英[1,2] 山田 武志[3] 柘植 覚[4] 宮島千代美[2] 西浦 敬信[5]
中山 雅人[5] 傅田 遊亀[5] 藤本 雅清[6] 山本 一公[7] 滝口 哲也[8]
黒岩 眞吾[4] 武田 一哉[2] 中村 哲[9]

[1] 豊橋技術科学大学 [2] 名古屋大学 [3] 筑波大学 [4] 徳島大学 [5] 立命館大学
[6] NTT CS 基礎研 [7] 信州大学 [8] 神戸大学 [9] NiCT/ATR
E-mail: kitaoka@slp.ics.tut.ac.jp

あらまし 雑音下における音声認識，音声強調，音声符号化などの音声処理で重要な役割を果たす音声区間検出 (Voice Activity Detection; VAD) 手法を評価するための基盤として CENSREC-1-C を構築した．これは，雑音下で発声された連続数字音声データと VAD 結果の評価を行うツール群からなる．評価方法としては一般的なフレームベースの検出性能評価尺度と音声認識を指向した発話単位の評価尺度を定義した．そして，音声パワーに基づくベースライン手法による VAD の結果をこれら２つの評価尺度で評価した結果を示した．

キーワード 雑音下音声区間検出，評価基盤，CENSREC-1-C，フレーム単位の評価尺度，発話単位の評価尺度

CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment

Norihide KITAOKA[1,2], Takeshi YAMADA[3], Satoru TSUGE[4], Chiyomi MIYAJIMA[2], Takanobu NISHIURA[5], Masato NAKAYAMA[5], Yuki DENDA[5], Masakiyo FUJIMOTO[6], Kazumasa YAMAMOTO[7], Tetsuya TAKIGUCHI[8], Shingo KUROIWA[4], Kazuya TAKEDA[2], and Satoshi NAKAMURA[9]

[1] Toyohashi Univ. of Tech. [2] Nagoya Univ.
[3] Univ. of Tsukuba [4] Univ. of Tokushima [5] Ritsumeikan Univ. [6] NTT CS research Lab. [7] Shinshu Univ. [8] Kobe Univ. [9] NiCT/ATR
E-mail: kitaoka@slp.ics.tut.ac.jp

Abstract Voice activity detection (VAD) plays an important role in speech processing including speech recognition, speech enhancement, and speech coding under noisy environment. We developed a evaluation framework for VAD under noisy environments, named CENSREC-1-C. This framework consists of noisy continuous digit utterances and evaluation tools for VAD results. We defined two evaluation measures, one for frame-level detection performance and the other for utterance-level detection perfromance. We showed the evaluation results of a baseline power-based VAD method.

Key words Voice activity detection under noisy environments, evaluation framework, measure for frame-level detection, measure for utterance-level detection

1. はじめに

近年，大規模な音声データと統計的手法により音声認識の基本性能は飛躍的に進歩した．最近では，実環境，特に雑音環境下での認識性能向上が盛んに研究されている．筆者らは，2001 年 10 月から，情報処理学会音声言語情報処理研究会傘下のワーキンググループ [1] として，主に雑音下日本語音声認識の評価方法の議論および評価基盤の構築・提供を行ってきており，これまでに CENSREC-1(Corpus and Environment for Noisy Speech RECognition 1; AURORA-2J) [2]，CENSREC-2 [3]，および CENSREC-3 [4] の作成・配布を行った．現在は，これまで配布した評価基盤を含めて，雑音下音声認識の性能に影響を与える様々な要因を分離・整理し，個々に評価・比較できる評価基盤群の提供を目指している．本稿では，それらの要因の一つとして最近特に研究が盛んになっている雑音下の音声区間検出手法 (Voice Activity Detection, VAD) の評価環境として構築した **CENSREC-1-C** (CENSREC-1-Concatenated) について述べる．雑音下連続数字発声データと評価ツールからなり，VAD を行った結果を 2 種類の評価基準に基づいて評価できるものである．

2. データ構成

本データは，連続数字を間隔をあけて発声したものからなる．個々の発声内容は CENSREC-1 [2] に準じている．雑音加算によるシミュレーションデータに対する評価と，実際の雑音環境下で発声された実環境データに対する評価が行えるよう，以下の 2 種類のデータ群を用意している．全データに対し，目視により作成された音声区間の始端・終端データが与えられている．

2.1 雑音加算によるシミュレーションデータ

クリーン環境で収録された音声に雑音を付加したデータセットであり，詳細な音声収録条件，発話内容，発音スタイル等は，CENSREC-1 [2] と同様である．語彙は数字 11 種類（1〜9, 0（まる），Z（ぜろ）），無音（sil）の 12 種類で，各発話は 1〜7 桁の連続数字音声となっている．音声データの収録は，ヘッドセットマイクロフォン (Sennheiser MHD25) を用いて防音室にて行っており，データのフォーマットは，サンプリング周波数 8kHz，量子化ビット数 16bit である．

本評価環境では，1 発話（1 連続数字）毎に収録された CENSREC-1 の音声データを複数接続することにより，連続的な発話のデータを作成している．接続する発話は同一話者のものであり，CENSREC-1 の 1 つの雑音環境で同一話者により発話されている 9 もしくは 10 の発話を 1 つの音声データとして接続した．接続の際には，CENSREC-1 から 1 秒の無音区間データを切り取り，発話の間に挟んでいる．また，CENSREC-1 の 1 つの雑音環境の音声データの話者数は 104 名（男女各 52 名）であり，話者毎に接続データを作成するため，1 環境におけるデータ数は 104 となる．

表 1　データベースの雑音環境

	加法性雑音	フィルタ特性
Set A	Subway, Babble, Car, Exhibition	G.712
Set B	Restaurant, Street, Airport, Station	G.712

表 2　実環境データの収録機材と収録条件

マイクロホン	コンデンサマイクロホン Sony ECM-77(近接・遠隔とも)
マイクロホンアンプ	ポータブルマルチミキサ Audio-technica AT-PMX5P
レコーダ	リニア PCM レコーダ Sony PCM-D1
サンプリング周波数	8kHz (収録は 48kHz)
量子化ビット数	16 ビット

本評価環境における雑音環境もまた，CENSREC-1 と同様であり，表 1 に示すような 2 種類のテストセットを設定した [2]．

表 1 において，Set A，Set B ではそれぞれ 4 種類の加法性雑音を用い，SNR は −5〜20 dB（5 dB 刻み）及びクリーン環境を用意している．全ての音声データには，ITU-T G.712 で勧告された電話回線を模擬したフィルタが畳み込まれている [2]．

本評価環境では，CENSREC-1 に習って Set A，Set B という名称を付けているが，評価の際には両者は一括して扱われる（一方が学習用といった区別はない）．

なお，本データベースでは CENSREC-1 の Set C（雑音は Set A, B からの 1 種類ずつ，フィルタ特性が MIRS）は提供していない．

2.2 実環境データ

実環境データの収録は，表 2 の機材および条件により 2 つの雑音環境（学生食堂，高速道路付近）および 2 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（ヘッドセット）と遠隔位置（話者の口元からの距離：50cm）に設置して，同期収録を行った．ただし近接マイクのデータは評価対象としていない．各 SNR 環境の設定は，学生食堂環境では，混雑時を低 SNR (平均雑音パワー 69.7dBA)，閑散時を高 SNR (同 53.4dBA) とし，また高速道路付近環境では，高速道路の本線付近を低 SNR (同 69.2dBA)，側道付近を高 SNR (同 58.4dBA) とした．音声データを収録した被験者は，男女 5 名，合わせて 10 名（被験者の年齢内訳は，20 歳前後男女各 3 名，30 歳前後男女各 1 名，40 歳以上男女各 1 名）とした．1 名の被験者に対して収録した音声は，各雑音環境および各 SNR 環境につき 1〜12 桁の連続数字を 8〜10 回，約 2 秒間の間隔で発声した音声を 1 つのファイルとして，計 4 ファイル（総発話数：38〜39 発話）である．

被験者のうち 1 名の収録データは，1 つの連続数字において数字間のインターバルが非常に長いという特徴があった．本評価環境では，1 つの連続数字を 1 つの音声区間と定義している

表 3 収録データの全容

シミュレーションデータ	
各ファイル中の発声	CENSREC-1 の 9〜10 発話を無音を挟んで接続
雑音	CENSREC-1 と同じ表 1 の雑音を −5〜20dB(5dB 刻み) で加算，またはクリーン
データ量	各条件 104 ファイル，全 5824 ファイル
実環境データ	
各ファイル中の発声	連続数字を約 2 秒間隔で 8〜10 発話
雑音	学生食堂，高速道路付近でそれぞれ低 SNR 条件および高 SNR 条件
データ量	各条件 4 ファイル，全 160 ファイル（評価対象は 144 ファイル）

ため，このようなデータに対して VAD を行った結果は正解セグメントと合わない恐れがあるため，今回は評価対象から外している．ただし，考えながら発声する場合などには起こりうる現象であると考えられるので，参考データとして収録してある．

2 種類の雑音環境および 2 種類の SNR 環境において，評価対象である被験者 9 名に対して収録した音声データの総発話数は 1380 発話（総ファイル数：144）である．また，上述した評価対象外の被験者 1 名を加えた被験者 10 名に対して収録した音声データの総発話数は 1532 発話（総ファイル数：160）である．

以上をまとめると，表 3 のようになる．なお，これらのデータは CENSREC-1 に準じた内容であるため，CENSREC-1 のデータで学習した HMM を用いて音声認識実験を行うことが可能である．ただし，本評価環境の配布物として認識スクリプトは含めていない．

2.3 正解セグメント

本評価環境では音声データに加えて，音声区間検出の評価を行う際に必要となる，音声区間の時間情報を記述した正解セグメントを同時に配布する．シミュレーションデータに対しては雑音重畳前のクリーンデータに対して，また実環境データに対しては近接マイク収録データに対して，目視により音声区間の始端・終端を検出した．実環境下の遠隔マイクに対する正解セグメントは近接のものに固定遅延を一律 12 サンプル付与することにより作成した．

3. 音声区間検出の性能評価尺度

評価尺度としては，VAD の性能評価に広く用いられているフレームベースの評価尺度と，音声認識のための VAD を指向した発話単位の評価尺度を定義する．2. 節のデータに対して実行した VAD の結果を，本節の尺度により評価するツールを配布する．評価者は，VAD の結果を定められた形式でファイルに保存し，評価ツールに与えることにより容易に評価結果を得ることができる．

表 4 SNR 別の評価に用いる平均値の計算条件

	平均に用いるデータ	
条件	シミュレーションデータ	実環境データ
Clean	Clean	該当なし
High SNR	SNR 20，15，10 dB	高 SNR 環境
Low SNR	SNR 5，0，−5 dB	低 SNR 環境
Average	Clean，SNR 20，15，10，5，0，−5 dB	高 SNR 環境と低 SNR 環境

3.1 フレームベースの評価尺度

音声区間検出の性能を測るためのフレームベースの評価尺度として，FRR（False Rejection Rate）と FAR（False Acceptance Rate）を用いる．

$$\mathrm{FRR} = \frac{N_{FR}}{N_s} \times 100\ [\%] \tag{1}$$

$$\mathrm{FAR} = \frac{N_{FA}}{N_{ns}} \times 100\ [\%] \tag{2}$$

ここで，N_s は音声フレーム数，N_{FR} は音声を非音声と検出したフレーム数，N_{ns} は非音声フレーム数，N_{FA} は非音声を音声と検出したフレーム数である．複数のデータを対象とする場合には，データ毎に FRR と FAR を求め，その平均値を用いて評価する．

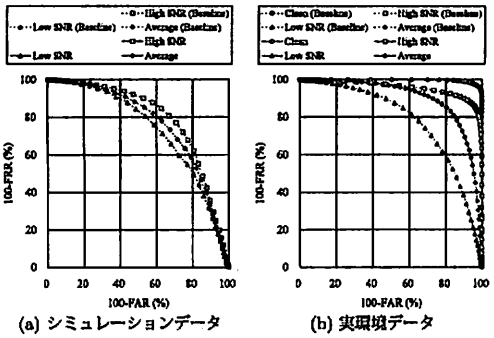
$$\overline{\mathrm{FRR}} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{FRR}_m \tag{3}$$

$$\overline{\mathrm{FAR}} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{FAR}_m \tag{4}$$

なお，正解の音声区間はサンプル単位で与えられているため，1 フレーム内に音声区間と非音声区間が混在することがある．このようなフレームについては，1 サンプルでも音声区間が含まれていれば音声とみなすことにしている．また，フレーム長は任意である(注1)．

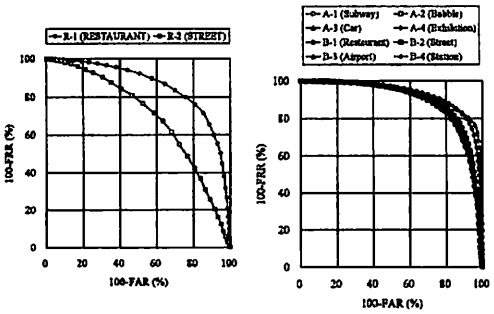
一般に FRR と FAR はトレードオフの関係にあり，どちらの性能を重視するかは対象とするアプリケーションによって決まる．よって，音声区間検出のアルゴリズムの多くは，しきい値によって FRR と FAR を調整できるようにしている．このようなアルゴリズムについては，ROC 曲線（x 軸：100 − FAR，y 軸：100 − FRR）を示すことが望ましい．本評価環境では，シミュレーションデータについては，雑音が 8 種類，SNR が 7 種類，実環境データについては，雑音が 2 種類，SNR が 2 種類ある．雑音と SNR の組合せの各々に対して ROC 曲線を求め，比較評価することは煩雑であることから，次の 2 種類の ROC

(注1)：フレーム長は VAD の性能を左右する重要なパラメータであることから，その値を規定することは適切ではない．なお，音声処理のアプリケーション（音声認識，音声強調，音声符号化など）では，数ミリ秒から数十ミリ秒周期の音声区間情報を必要とすることが多く，この範囲内であればフレーム長が異なっても FRR と FAR にさほどの違いは生じないと考えられる．



(a) シミュレーションデータ　(b) 実環境データ

図 1　ベースラインの SNR 別の ROC 曲線(注2)



(a) シミュレーションデータ　(b) 実環境データ

図 2　ベースラインの雑音別の ROC 曲線

曲線を用いて VAD の性能を評価する．

(1)　SNR 別の ROC 曲線

(2)　雑音別の ROC 曲線

(1) については，ベースラインとの比較評価を目的とする．具体的には，表 4 に示す 4 つの条件での平均値に対して ROC 曲線を求め，ベースラインとの性能比較を行う．ここでは，雑音の違いは考慮せず，全ての雑音に対する結果を平均する．

(2) については，ベースラインとの比較評価ではなく，評価対象である VAD アルゴリズムの特性評価（雑音の違いに対するロバスト性の評価）を目的とする．ここでは，SNR としては表 4 の Average に相当する平均のみを対象とする．いずれの場合も，ROC 曲線は x 軸を 100 − FAR（非音声区間検出率），y 軸を 100 − FRR（音声区間検出率）とする．後述するベースラインに対する結果を図 1 と図 2 に示す．

3.2　発話単位の評価尺度

一般に音声認識のための VAD は発話単位 (孤立単語，連続単語，文など) で行う．本評価環境では，1 つの「連続数字」が

(注2)：本来は 4.2 節で説明するベースライン手法と新たに評価対象とする手法の比較が目的であるため，全ての凡例が含まれている (図中の点線はベースライン，実線は比較対象手法)．ただし，ここではベースライン手法のみ示している．

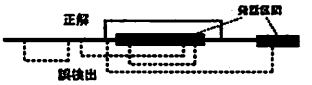


図 3　発話区間検出の正解/不正解

「発話」の単位である．発話単位の VAD の結果を評価する尺度として，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．

$$\mathrm{Corr} = \frac{N_c}{N} \times 100\ [\%] \tag{5}$$

$$\mathrm{Acc} = \frac{N_c - N_f}{N} \times 100\ [\%] \tag{6}$$

ここで，N は音声データに含まれる総発話数，N_c は正解検出数，N_f は誤検出数である．Corr は，どれだけ多く発話区間を検出できたかを評価する尺度であるのに対し，Acc は，どれだけ余分に発話区間を検出してしまったかを見るための尺度である．これは，検出に成功した発話は完全に認識できるという条件の下での認識性能におおよそ対応する．

また，複数のデータを対象とする場合には，全データに対して N, N_c, N_f の値を累積し，次式に基づいて平均の Corr, Acc を求める．

$$\overline{\mathrm{Corr}} = \frac{\sum_{m=1}^{M} N_{c,m}}{\sum_{m=1}^{M} N_m} \times 100 \tag{7}$$

$$\overline{\mathrm{Acc}} = \frac{\sum_{m=1}^{M} N_{c,m} - \sum_{m=1}^{M} N_{f,m}}{\sum_{m=1}^{M} N_m} \times 100 \tag{8}$$

ここで，M はデータの総数，N_m, $N_{c,m}$, $N_{f,m}$ は，それぞれ m 番目のデータの N, N_c, N_f を表す．

音声認識では，実際の発話区間より短く検出すると，端点付近の音素の情報が失われ誤認識の原因となるが，発話区間の前後に少しの無音を含めて検出しても，認識結果への影響は少ないと考えられる．そこで，検出区間がある 1 つの発話区間全体を含み，その前後の発話区間と重なりがなければ正解候補とする．検出結果に長い無音が含まれた場合も誤認識の原因となるが，本データベースでは，発話と発話の間の無音区間の長さが約 2, 3 秒程度と比較的短いため，ここでは検出結果が複数の発話にまたがらなければ，正解候補と考えることとした．

本評価環境では，複数の検出区間がオーバーラップすることを許すため，1 つの発話に対して複数の正解候補が該当する場合がある．そのような場合，正解候補のうち正解の発話区間との誤差 (始・終端点におけるずれのポイント数の合計) が最小の正解候補のみを正解 (N_c) として数え，残りの候補は全て誤り (N_f) として数える．1 つの発話に対して複数検出しても Corr の値に影響はないが，Acc の値は低くなる．

その他，無音区間のみを含む検出結果や，始・終端点が発話の途中にある検出結果，複数の発話にまたがる検出結果は，全

表 5　シミュレーションデータにおけるベースラインに対する発話単位の評価結果 (上: 正解率, 下: 精度)

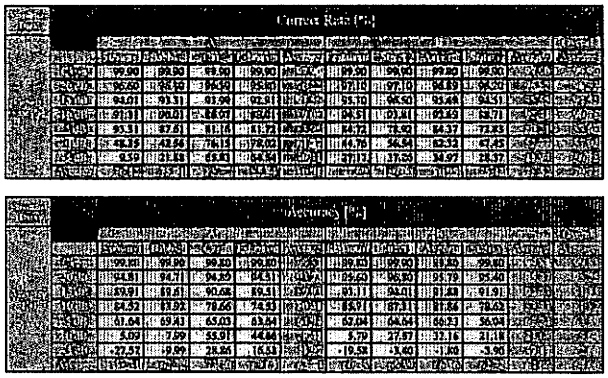
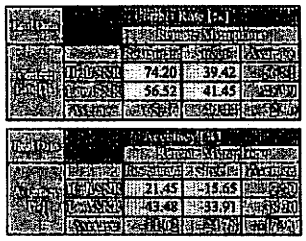


表 6　実環境データにおけるベースラインに対する発話単位の評価結果 (上: 正解率, 下: 精度)



て誤り (N_f) として数える (図 3).

なお，本評価基準では，検出区間が少し欠けた場合でも不正解と見なされるため，評価の際には，検出された区間の前後に適度なマージン (例えば，数百 ms 程度) を与えた結果を出力する VAD 手法がよい評価を得ることとなる．

最終的な結果は，複数の条件 (しきい値) における結果の中で，平均の Corr が最大となるしきい値におけるものとする．つまり，いかに発話区間を多く検出できたか (Corr) を基準にしきい値を選ぶ．ここで，平均の Corr とは，シミュレーションデータの場合，全雑音，全 SNR に対する Corr の平均，実環境データの場合，全雑音，全 SNR 環境における Corr の平均を指す．また，しきい値はシミュレーションデータと実環境データとの間で異なってもよいが，それぞれのデータの中では共通としなければならない．後述するベースラインに対する結果を表 5 と表 6 に示す．なお，本評価環境にはこれらの表とともに評価対象手法の結果を入力するための表も用意されている．これにより，ベースラインからの相対的な性能評価ができる．

4. ベースライン VAD

本評価環境でベースラインとする VAD のアルゴリズムとその性能について述べる．現在，VAD では一般的に音声のパワーに基づいた手法が多く用いられる．本ベースライン VAD も，音声のパワーに着目したものである．ただし，雑音環境化での使用のために特に調整したものではない．

4.1 手　　法

ベースライン VAD では，以下の手順で音声区間を決定する．

（1） 音声信号のフレーム化

音声信号をフレームに分割する．フレーム幅は 5ms，フレーム周期は 2ms である．

（2） 各フレームのパワーの計算

分割されたフレームごとに対数パワーの平均 (以下パワーと呼ぶ) を以下の式で計算する．

$$\mathrm{POW}_i = 10\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} s_i^2(n)\right), \quad (i = 0, 1, \ldots, M-1) \tag{9}$$

ここで，POW_i は i フレーム目の対数パワーの平均，$s_i(n)$ はフレーム i に属する音声信号，N はフレームに含まれる音声信号のサンプル数を示す．

（3） 音声，非音声を決定するしきい値の計算

フレームごとのパワーを用い，各音声データごとの初期しきい値 (THR_{int}) を決定する．初期しきい値は文献 [5] の方法を用い自動的に決定する．この方法は，フレームごとのパワーをしきい値により 2 クラス (C_1 (音声)，C_2 (非音声)) に分類するものである．クラスを分離する最適なしきい値は判別基準 ($\eta(s)$)

$$\eta(s) = \frac{\sigma_B^2(s)}{\sigma_T^2} \tag{10}$$

$$\sigma_T^2 = \sigma_W^2(s) + \sigma_B^2(s) \tag{11}$$

を最大とするしきい値 (s) を求めることにより決定する．ここで，$\sigma_W^2(s)$ は平均クラス内分散，$\sigma_B^2(s)$ は平均クラス間分散を

示す．実際には，σ_T^2 はしきい値 (s) に依存しないため，平均クラス間分散 ($\sigma_B^2(s)$) を最大とするしきい値を求めることとなる．アルゴリズムの詳細は文献 [5] を参照にされたい．この手法で求めた最適なしきい値を初期しきい値 (THR_{int}) として用いる．その後，初期しきい値を以下の式で変動し，各しきい値 (THR) とする．

$$\mathrm{THR} = \mathrm{THR}_{int} + k \cdot \alpha \tag{12}$$

$$\alpha = \frac{(\mathrm{POW}_h - \mathrm{POW}_l)}{K} \tag{13}$$

ここで，POW_l は初期しきい値未満のフレームの平均パワー，POW_h は初期しきい値以上のフレームの平均パワーを示す．今回は，$K = 40$，$k = -40, -39, \ldots, 39, 40$ とし，81 通りのしきい値に対し音声区間検出を行った．

（4） 音声区間検出

音声区間検出は以下の方法で行う．

（a） 音声開始フレームの検出

フレームのパワーがしきい値以上となった場合，当該フレームを音声開始フレーム候補とし，次の音声終了フレーム検出を行う．

（b） 音声終了フレームの検出

音声開始フレームの候補が決定された後，フレームのパワーがしきい値未満となりかつそのフレーム以降のある一定区間 (ベースライン実験では 500ms) 以上のフレームのパワーがしきい値未満の場合，音声終了フレーム候補とする．一定区間以上連続してフレームのパワーがしきい値未満でない場合，再度音声終了フレームを検出する．

（c） 音声区間の決定

音声開始フレームと音声終了フレームにより決定される音声区間が一定区間 (ベースラインの実験では 100ms) 以上の場合その区間を音声区間として決定し，未満の場合にはその音声区間は棄却する．その後，次の音声区間を検出するため，再度音声開始フレームの検出を行う．

（5） 結果の出力

4. で決定したフレーム単位の音声区間を音声信号のポイント単位の音声区間に換算する．

音声開始ポイント = 音声開始フレーム
× 標本化周波数 (Hz) × フレーム周期 (秒)

音声終了ポイント = 音声終了フレーム
× 標本化周波数 (Hz) × フレーム周期 (秒) − 1

4.2 ベースラインの性能

本節では，4.1 節で述べた VAD の性能を示す．実験条件および評価方法は 2. 節および 3. 節の通りである．

4.2.1 フレームベースの性能評価

SNR 別，雑音別の ROC 曲線を図 1 と図 2 に示す．(a)，(b) はそれぞれシミュレーションデータと実環境データを示す．

SNR 別の ROC 曲線 (図 1) より，SNR が高い場合には実環境データ，シミュレーションデータに関わらず高い音声区間検出が可能であるが，SNR の低下とともに急速に性能が劣化していることがわかる．実環境データにおいては高 SNR (High SNR) と低 SNR (Low SNR) との性能の差は比較的少ない．雑音別の ROC 曲線 (図 2) からは，シミュレーションデータでは大きな差が見られないが，実環境データでは雑音の違いが音声区間検出精度に大きく影響を与えており，本手法では両者の雑音には対応仕切れていないことといえる．

4.2.2 発話単位の性能評価

シミュレーションデータおよび実環境データの発話区間検出正解率 (Corr) と発話区間検出正解精度 (Acc) を表 5 および表 6 に示す．これは，最も発話区間検出正解率 (Corr) が高かったしきい値 (シミュレーション: $k = 10$，実環境: $k = 17$) の結果である．なお，4.1 節で述べた手法で発話区間を決定した後，その発話区間の前後に 300ms のマージンをとり発話区間として用いた．

これらの結果より，フレームベースの評価と同様に，シミュレーション，実環境に関わらず，SNR が高い場合は発話区間検出性能が高いが SNR の低下とともに急速に劣化していることがわかる．

5. ま と め

本稿では，雑音下の音声区間検出手法の評価基盤として構築した CENSREC-1-C について報告した．VAD は用途により，時間遅れがあまり許されない場合や，低演算量，低メモリ消費量が必須となる場合がある．今後，このような要因についての評価も検討する．また，本環境を用いた音声認識性能による評価方法なども提案していく．

謝 辞

本評価基盤の構築にあたり，国立情報学研究所 音声資源コンソーシアムに御協力頂いた．ここに感謝致します．

文 献

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

[2] S. Nakamura *et al.*, "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition," *IEICE Transactions on Information and Systems*, Vol. E88–D, No. 3, pp. 535–544, 2005.

[3] 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.

[4] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-3: An Evaluation Framework for Japanese Speech Recognition in Real Driving-Car Environments," *IEICE Transactions on Information and Systems*, (accepted).

[5] 大津展之, "判別および最小 2 乗基準に基づく自動しきい値選定法," 信学論, Vol.J63-D, No. 4, pp. 349–356, 1980.