

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

藤本 雅清¹ 武田 一哉² 中村 哲¹

¹ATR 音声言語コミュニケーション研究所

E-mail: {masakiyo.fujimoto, satoshi.nakamura}@atr.jp

²名古屋大学大学院 情報科学研究科

E-mail: kazuya.takeda@nagoya-u.jp

あらまし 本稿では, SLP 雑音下音声認識評価ワーキンググループの活動成果として, 自動車内における連続数字音声認識の評価用データベース CENSREC-2 と, 標準評価スクリプトによるベースライン評価結果について述べる. 音声データの収録は, 接話マイクと遠隔マイクの 2 種類を用いて, 3 種類の走行速度と 4 種類の車内環境を組み合わせた 11 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 4 種類の評価環境を提供する.

キーワード: 雑音下音声認識, 共通評価フレームワーク, 自動車内音声データベース, CENSREC-2

CENSREC-2: Data Collection for In-Car Digit Speech Recognition and Its Common Evaluation Framework

Masakiyo Fujimoto¹ Kazuya Takeda² Satoshi Nakamura¹

¹ATR Spoken Language Communication Research Laboratories

E-mail: {masakiyo.fujimoto, satoshi.nakamura}@atr.jp

²Graduate School of Information Science, Nagoya University

E-mail: kazuya.takeda@nagoya-u.jp

Abstract This paper introduces a common database and an evaluation framework for connected digit speech recognition in real driving car environments, CENSREC-2, as an outcome of IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group. Speech data of CENSREC-2 was collected using two microphones, a close-talking microphone and a hands-free microphone, under carefully controlled 11 different driving conditions, i.e., combinations of three car speeds and four car conditions. CENSREC-2 provides four evaluation environments which are designed using speech data collected in these car conditions.

Keywords: noisy speech recognition, common evaluation framework, in-car speech database, CENSREC-2

1 はじめに

近年の音声認識技術の進歩は, 統計モデルの導入と大規模コーパスの収集によりもたらされた. しかし, 現在の音声認識を実際に利用されるような雑音のある環境で利用すると, 未だ著しい性能劣化が避けられない.

このような音声認識の音響環境に対する頑健性の問題に対しては, これまでに米国 DARPA [1] 主催の SPINE [2] と, 欧州 ETSI [3] 主導の AURORA [4]-[10] の 2 つのプロジェクトが進められた.

また, このような欧米の動きに対し筆者らは, 2001 年 10 月に情報処理学会 音声言語情報処理研究会内に, 有志によるワーキンググループを作り, 雑音下音声認識の評価

のための議論を進めてきた. ワーキンググループの活動成果として筆者らは, 2003 年 7 月に AURORA2 の数字を日本語に翻訳し, 同一の雑音データを付与した AURORA-2J [11] を作成し, 配布を行った. さらに, 2004 年 12 月にワーキンググループが独自に設計した, 雑音下音声認識の評価環境である CENSREC-3 (Corpus and Environments for Noisy Speech REcognition) [12] を作成し, 配布を行った*. CENSREC-3 にて取り扱った認識タスクは, 実走行車内で収録された孤立単語音声の認識であり, カーナビゲーションシステムの音声コマンド操作を意識した評価環境であった.

本稿では, 新たな雑音下音声認識の標準評価環境である CENSREC-2 の概要と, 標準評価スクリプトによるベース

*AURORA-2J は, CENSREC-1 に相当する.

ライン性能について述べる．CENSREC-2 の音声認識タスクは，実走行車内での連続数字音声認識であり，様々な環境したで収録された音声データを用いた 4 種類の評価環境を提供する．

2 自動車内音声データの収録

2.1 収録語彙

CENSREC-2 は，連続数字音声の認識を対象としており，語彙は数字 11 種類（1~9, 0（まる），Z（ぜろ）），無音（sil），ショートポーズ（sp）の 13 種類である．発話内容，発音スタイル等は，AURORA-2J [11] と同様である．

2.2 音声データの収録環境

自動車内音声の収録は，特別に設計された実験車両を用いて行った．実験車両には，運転座席周辺に 5 本のマイクロホンが図 1 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（ヘッドセット）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 [13]．それぞれのマイクロホンには，SONY ECM77B を用いている．

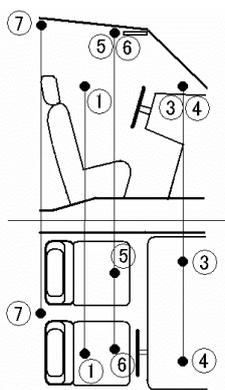


図 1: マイクロホンの設置位置（上段）：側面，（下段）：真上

音声データの収録は，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（通常走行，エアコン On，オーディオ On，窓開）を組み合わせた 11 種類の環境で行った．評価データの発話者数は 104 名（男性 52 名，女性 52 名）であり，収録音声の総数は 17,651 発話である．これらの収録音声のう

ち，73 名（男性 33 名，女性 40 名）の話者のデータを学習データとし，31 名（男性 19 名，女性 12 名）の話者のデータを評価データとした．学習データの総数は 14,687 発話（接話マイクロホン: 7,492，遠隔マイクロホン: 7,195）であり，評価データの総数は 2,964 発話（遠隔マイクロホンのデータのみ）である．

以上のような環境での収録において，収録条件は評価データ，学習データともに，標準化周波数 16kHz，語長 16bit であり，バイトオーダーはリトルエンディアンである．また，収録時に収録機器のノイズが生じたデータや，オーバーフローを起こしたデータはデータベースより削除した．

表 1: 音声データの収録環境

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

図 2 は，収録音声における各数字の出現頻度，図 3 は，桁数の出現頻度を示している（6 桁数字のデータは存在しない）．それぞれの図において，CT は接話マイクロホンのデータ，HF は遠隔マイクロホンのデータを示す．

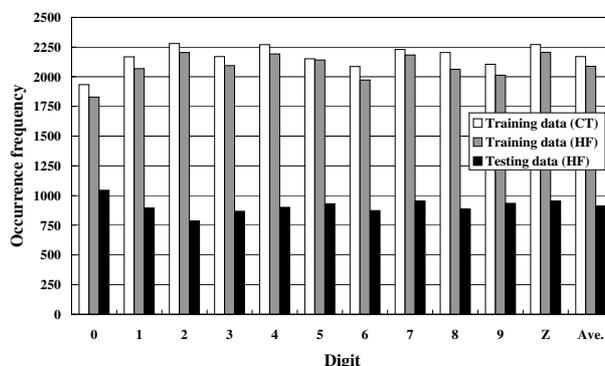


図 2: 各数字の出現頻度

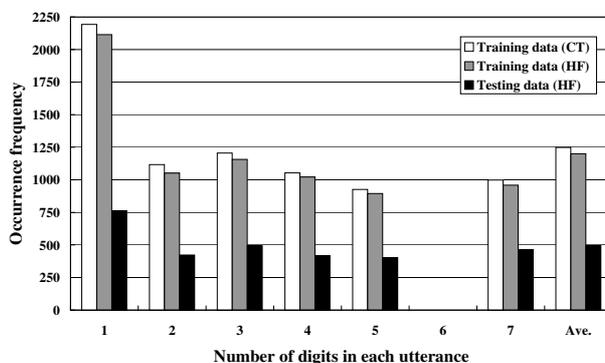


図 3: 桁数の出現頻度

表 2: 各評価環境で用いられる学習データ条件

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

表 3: 各評価環境で用いられる評価データ条件

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	○	○	○	○

前述の通り，ノイズなど問題の生じたデータを削除したため，数字及び，桁数の出現頻度に多少の不均衡が生じたが，可能な限りバランスの取れた出現頻度になるよう，学習データ及び，評価データの選定に細心の注意を払った．

3 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種類の音声認識評価環境（Condition 1～4）を構成する．

表 2, 3 の ○ 印は，各々の評価環境で用いるデータ，表 4, 5 は，使用するデータ量を示しており，それぞれの環境は，

Condition 1: 学習データと評価データでマイクロホン種別，収録環境が共に一致

Condition 2: 学習データと評価データでマイクロホン種別が一致，収録環境が相違

Condition 3: 学習データと評価データでマイクロホン種別が相違，収録環境が一致

Condition 4: 学習データと評価データでマイクロホン種別，収録環境が共に相違

という条件を設定している．

4 ベースライン認識性能

4.1 評価用スクリプト

評価用ベースラインスクリプトは，HTK [14] を用いて HMM の学習，評価実験が容易に行えるように作成し，以下のようなベースライン評価の仕様を設計した．

- 音声の認識は単語単位 HMM により行い，図 4 に示す EBNF 記法の認識文法を用いる．

- HMM のトポロジーについては，数字 HMM が 16 状態 20 混合分布，無音 HMM が 3 状態 36 混合分布，ショートポーズ HMM が 1 状態 36 混合分布（無音モデルの第 2 状態と共有）である．

- 特徴量は，HTK の HCopy により抽出された，12 次 MFCC + log-power + Δ MFCC + Δ log-power + $\Delta\Delta$ MFCC + $\Delta\Delta$ log-power の 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，24 次元のメルフィルタバンク，20ms の分析フレーム長，10ms のフレームシフトとする．また，cepstral mean subtraction は行わない．

- 自動車雑音特有の低周波成分に対処するため，メルフィルタバンク分析時に 250Hz 以下の低周波成分を取り除く．

```
$digit = one | two | three | four |
         five | six | seven | eight |
         nine | zero | oh ;
```

```
( [sil] < $digit [sp] > [sil] )
```

図 4: EBNF 記法による認識用文法

4.2 ベースライン認識結果と認識性能の比較

表 6 に，評価環境 Condition 1～4 の車内環境毎の詳細なベースライン認識性能を示す．

また，研究機関毎の認識性能比較を容易にするため，表 7 のような Microsoft Excel にて作成されたスプレッドシートを配布する．表 7 の上段は，各評価環境のベースライン

表 4: 各評価環境における学習データ数

走行速度	マイクロホン	車内環境	Condition 1	Condition 2	Condition 3	Condition 4	
アイドリング	接話	通常走行	—	—	686	686	
		エアコン On	—	—	686	686	
		オーディオ On	—	—	680	680	
		窓開	—	—	685	685	
		合計	—	—	2,737	2,737	
	遠隔	通常走行	538	538	—	—	
		エアコン On	663	663	—	—	
		オーディオ On	698	698	—	—	
		窓開	498	498	—	—	
		合計	2,397	2,397	—	—	
	合計		2,397	2,397	2,737	2,737	
	低速走行	接話	通常走行	—	—	685	—
			エアコン On	—	—	682	—
			オーディオ On	—	—	690	—
窓開			—	—	671	—	
合計			—	—	2,728	—	
遠隔		通常走行	700	—	—	—	
		エアコン On	694	—	—	—	
		オーディオ On	697	—	—	—	
		窓開	666	—	—	—	
		合計	2,757	—	—	—	
合計			2,757	—	2,728	—	
高速走行		接話	通常走行	—	—	682	—
			エアコン On	—	—	677	—
			オーディオ On	—	—	668	—
	合計		—	—	2,027	—	
	通常走行		687	—	—	—	
	遠隔	エアコン On	678	—	—	—	
		オーディオ On	676	—	—	—	
		合計	2,041	—	—	—	
		合計	2,041	—	2,027	—	
	合計		7,195	2,397	7,492	2,737	

表 5: 各評価環境における評価データ数

走行速度	車内環境	Condition 1	Condition 2	Condition 3	Condition 4
アイドリング	通常走行	198	—	—	—
	エアコン On	216	—	—	—
	オーディオ On	297	—	—	—
	窓開	195	—	—	—
	合計	906	—	—	—
低速走行	通常走行	298	298	298	298
	エアコン On	294	294	294	294
	オーディオ On	297	297	297	297
	窓開	291	291	291	291
	合計	1,180	1,180	1,180	1,180
高速走行	通常走行	293	293	293	293
	エアコン On	291	291	291	291
	オーディオ On	294	294	294	294
	合計	878	878	878	878
	合計	2,964	2,058	2,058	2,058

表 6: CENSREC-2 ベースライン認識性能の詳細 (%)

走行速度	車内環境	Condition 1	Condition 2	Condition 3	Condition 4
アイドリング	通常走行	94.06	—	—	—
	エアコン On	93.96	—	—	—
	オーディオ On	68.60	—	—	—
	窓開	96.46	—	—	—
	全環境	86.38	—	—	—
低速走行	通常走行	89.14	79.78	78.98	63.55
	エアコン On	88.09	88.60	66.70	56.41
	オーディオ On	67.04	73.27	60.30	51.26
	窓開	78.86	77.43	57.10	46.68
	合計	80.80	79.77	65.84	54.52
高速走行	通常走行	78.97	57.96	62.45	43.27
	エアコン On	79.75	77.20	52.66	39.78
	オーディオ On	63.76	67.11	51.57	40.71
	全環境	74.14	67.38	55.56	41.25
	全環境	80.58	74.49	61.46	48.87

表 7: CENSREC-2 スプレッドシート

CENSREC-2 Evaluation Results				
CENSREC-2 Baseline Results (%)				
Condition 1	Condition 2	Condition 3	Condition 4	Average
80.58	74.49	61.46	48.87	66.35
CENSREC-2 Word Accuracy (%)				
Condition 1	Condition 2	Condition 3	Condition 4	Average
CENSREC-2 Relative Improvement				
Condition 1	Condition 2	Condition 3	Condition 4	Average

性能とその平均を示しており、中段に自身の手法による認識結果 (%Acc) を入力する。中段に認識結果を入力すると、下段にベースライン性能との相対的な改善性能 (誤り改善率: Relative improvement) が自動で出力される。また、誤り改善率 (Relative improvement) は、次式により得られる。

$$\begin{aligned} \text{Relative improvement} &= \\ &= \frac{\% \text{Acc} - \% \text{Acc of baseline}}{100 - \% \text{Acc of baseline}} \times 100(\%) \end{aligned}$$

4.3 評価カテゴリー

CENSREC-2 では、バックエンドの変更 (HMM の学習方法、トポロジーの変更、特徴量の変更など) に対して、その度合に応じたカテゴリーを設定する。バックエンドを変更した結果を発表する場合、以下に示すカテゴリーから一つを選び、発表でそれを示す必要がある。バックエンドを変更しない場合は、カテゴリー 0. となる。カテゴリー内で性能比較を行なうことで、各手法の性能比較をより適切に行な

うことができる。尚、下記のカテゴリーは、AURORA-2J のカテゴリー設定に一部変更を加えたものとなっている。

カテゴリー 0. ベースラインスクリプトを全く変更しない場合。

カテゴリー 1. 標準 HMM と同じトポロジーの HMM だが、識別学習等、学習方法を変更している場合。このカテゴリーの認識時のコストは、ベースラインと全く同じである。その他の実験条件はベースラインと同じ条件に従う。

カテゴリー 2. ベースラインスクリプトと同じトポロジーの HMM で、認識時の適応技術を導入している場合。話者適応、環境適応、1 状態 1 混合の雑音 HMM を用いた PMC 等がここに含まれ、認識時に適応を行なうことによる認識コストの増加がある。その他の実験条件はベースラインと同じ条件に従う。

カテゴリー 3. 混合数や状態数等の HMM トポロジーを変

更している場合．ただし，モデル単位はベースラインと同じ (CENSREC-2 では単語単位 HMM) であることを条件とし，2 状態以上の雑音モデルを用いた PMC 等がこれに相当．その他の実験条件はベースラインと同じ条件に従う．

カテゴリ 4. 認識デコーダがベースラインスクリプトと同じ (CENSREC-2 では HVite) であることを条件にどのような処理も許される場合．モデル単位の変更や，文法・辞書の書き換え等がこれに相当．

カテゴリ 5. 規定無し．提供されるデータベース内であれば，どんな処理でも許される．認識デコーダの変更も許容．

カテゴリ B. 提供されるデータ以外のデータを使用する場合．評価データは提供されているものを用いる．

5 まとめ

本稿では，自動車内音声認識の評価用データベースである CENSREC-2 と，標準評価スクリプトによるベースライン評価結果について報告した．

CENSREC データベースが対象とする認識タスク，評価環境は，以下のような変遷を辿っている．

CENSREC-1 人工データ，連続数字認識，加法的雑音

CENSREC-2 実走行車内音声データ，連続数字認識

CENSREC-3 実走行車内音声データ，孤立単語認識

今後，非定常雑音下及び，残響下での認識，大語彙連続音声認識など，雑音環境と認識タスクを徐々に難しくした評価環境を設計し，データベースとして公開する予定である．また，標準的な雑音データベースの設計，単語正解精度以外の評価指標の検討，雑音対策用の標準ツールの開発，配布などについても検討を行う予定である．尚，AURORA-J/CENSREC に関する最新の情報は以下の URL を参照されたい．

AURORA-J/CENSREC Web site:

<http://sp.shinshu-u.ac.jp/CENSREC/>

謝辞 本研究は，情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．本研究を行うにあたり多大な助言を頂いた，SLP 雑音下音声認識評価ワーキンググループの皆様方に深く感謝致します．

参考文献

- [1] DARPA project Web site, <http://www.nist.gov/speech/publications/>
- [2] SPINE Web site, <http://elazar.itd.nrl.navy.mil/spine/>
- [3] ETSI Web site, <http://www.etsi.org/>
- [4] H.G.Hirsch and D.Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition," *Proc. ISCA ITRW ASR2000*, pp. 18-20, Paris, France, Sept. 2000.
- [5] AU/378/01, "Danish SpeechDat-Car Digits Database for ETSI STQ-Aurora Advanced DSR," *Aalborg University*, Jan. 2001.
- [6] AU/225/00, "Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation," *Nokia*, Jan. 2000
- [7] AU/273/00, "Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation," *Texas Instruments*, Dec. 2001.
- [8] AU/271/00, "Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation: Description and Baseline Results," *UPC*, Nov. 2000.
- [9] AU/337/01, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-Ends on a Large Vocabulary Task: Version 1.0," *Ericsson*, June 2001.
- [10] AU/345/01, "Large Vocabulary Evaluation of Front-ends: Baseline Recognition System Description, Final Report," *Mississippi State University*, Jan. 2002.
- [11] 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.
- [12] 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," 情報処理学会研究報告, SLP-55-8, pp.41-46, Feb. 2005.
- [13] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and 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.
- [14] HTK Web site, <http://htk.eng.cam.ac.uk/>