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

2003－ＳＬＰ－49 (18)
2003／12／18

AURORA-2Jを用いた ETSI STQ Aurora WI008 Advanced DSR Frontendの評価

山田 武志[1] 武田 一哉[2] 北岡 教英[3] 藤本 雅清[4] 黒岩 眞吾[5]
山本 一公[6] 西浦 敬信[7] 佐宗 晃[8] 水町 光徳[9] 遠藤 俊樹[9]
中村 哲[9]

[1]筑波大学 〒305-8573 茨城県つくば市天王台1-1-1
[2]名古屋大学 [3]豊橋技術科学大学 [4]龍谷大学 [5]徳島大学 [6]信州大学 [7]和歌山大学 [8]産総研 [9]ATR-SLT
E-mail: [1]takeshi@is.tsukuba.ac.jp

あらまし 本稿では，分散型音声認識のための標準フロントエンドであるETSI ES201 (WI007) とETSI ES202 (WI008) の性能をAURORA-2Jを用いて比較評価する．その際，AURORA2やAURORA-2Jで採用している平均的な認識性能を表す評価指標に加えて，話者毎の認識性能を表す評価指標を用いる．具体的には，話者毎の単語正解精度の最大値，最小値，平均値，標準偏差，話者毎の単語正解精度のヒストグラム，単語正解精度がx%以上の話者の割合である．その結果，WI008の認識性能は，WI007を大幅に上回っていることが確認できた．その一方で，話者毎の認識性能には，さらなる改善が必要であることが分かった．

キーワード 雑音下音声認識，AURORA-2J，ETSI STQ Aurora WI008 Advanced DSR Frontend，評価指標

Evaluation of ETSI STQ Aurora WI008 Advanced DSR Frontend Using AURORA-2J

Takeshi YAMADA[1], Kazuya TAKEDA[2], Norihide KITAOKA[3], Masakiyo FUJIMOTO[4], Shingo KUROIWA[5], Kazumasa YAMAMOTO[6], Takanobu NISHIURA[7], Akira SASOU[8], Mitsunori MIZUMACHI[9], Toshiki ENDO[9], and Satoshi NAKAMURA[9]

[1]University of Tsukuba 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
[2]Nagoya University [3]Toyohashi University of Technology [4]Ryukoku University [5]University of Tokushima [6]Shinshu University [7]Wakayama University [8]AIST [9]ATR-SLT
E-mail: [1]takeshi@is.tsukuba.ac.jp

Abstract This paper describes the results of comparative evaluation of ETSI ES201 (WI007) and ETSI ES202 (WI008), which are the standard frontends for distributed speech recognition. As the evaluation index, the word accuracy for each speaker is used in addition to the overall word accuracy. The experimental results using AURORA-2J confirmed that the WI008 achieves much better recognition performance than the WI007, and the WI008 still has the problem of speaker dependency.

Key words noisy speech recognition, AURORA-2J, ETSI STQ Aurora WI008 advanced DSR frontend, evaluation index

1. まえがき

携帯電話やPDAに代表される携帯端末の急速な普及に伴い，ワイヤレス・モバイル環境がますます一般的なものとなりつつある．このような中，携帯端末に音声認識を組込むことにより，使い勝手を向上させるという試みがなされている．しかし，携帯端末の計算資源は限られているので，現状では中～大語彙の音声認識を行うことは難しい．

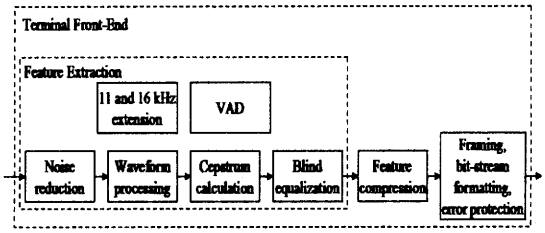

図1 WI008の端末側の処理フロー
Fig. 1 Process flow of the WI008 in the terminal side.

このような問題を解決するために，携帯端末側では特徴量抽出などの音響分析を行い，サーバ側ではデコードなどの認識処理を行うという分散型音声認識（DSR：Distributed Speech Recognition）が提案されている[1]．この提案に沿って，欧州電気通信標準化機構（ETSI：European Telecommunications Standards Institute）では，音声認識のフロントエンドの標準化が進められている．その成果として，2000年4月にメルケプストラムからなる特徴量を抽出するETSI ES201（WI007）[2]，2002年10月にWI007に雑音抑圧処理を組込んだETSI ES202（WI008）[3]が勧告された．

一方，このような標準化のための活動の一環として，ETSI傘下のAURORAグループでは，フロントエンドの開発評価用コーパス（AURORA2，AURORA3など）を配布している．これらのコーパスでは，比較評価のためのベースラインフロントエンドとして，WI007，WI008を順次採用している．日本においても，IPSJ SIG-SLP傘下の雑音下音声認識評価ワーキンググループ[4]が同様の活動を進めており，2003年7月にはAURORA2の日本語版であるAURORA-2J[5]の配布を開始している．

現在のところ，AURORA-2JではベースラインフロントエンドとしてWI007を採用しており，AURORA-2JにおけるWI008の有効性を評価していない．そこで，本稿では，AURORA-2Jを用いてWI007とWI008の性能を比較評価する．その際，AURORA2やAURORA-2Jで採用している平均的な認識性能を表す評価指標に加えて，話者毎の認識性能を表す評価指標を用いる．

2. ETSI STQ Aurora WI008 Advanced DSR Frontend

WI008の端末側の処理フローを図1に示す．端末側では，入力音声に対して，雑音抑圧，波形処理，特徴抽出，Blind Equalization，圧縮，エンコードが順番に行われる．以下では，雑音抑圧部について述べる．

WI008では，ウィーナーフィルタ理論に基づく2段階の雑音抑圧がフレーム同期で実行される．雑音抑圧部の処理フローを図2に示す．第1段階では，VAD（音声区間検出）の結果を用いて音声と雑音のパワースペクトルを推定し，ウィーナーフィルタ係数を求める．ここで，図中の「PSD mean」は，パワースペクトルを時間軸方向に平滑化する処理を表す．そして，

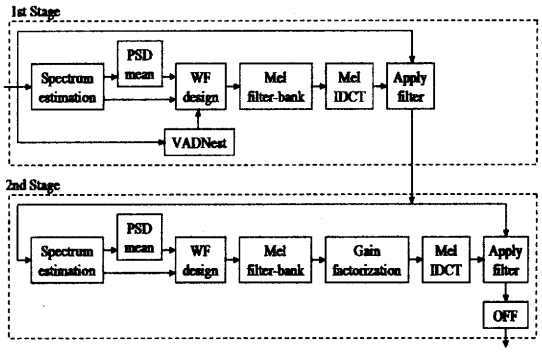

図2 雑音抑圧部の処理フロー
Fig. 2 Process flow of the noise reduction part.

表2 認識実験の条件
Table 2 Conditions of the recognition experiments.

窓関数	ハミング窓
フレーム長	25msec
フレーム周期	10msec
特徴量	メルケプストラム係数（12次元）
	＋対数パワー（1次元）
	＋Δ係数（13次元）
	＋ΔΔ係数（13次元）
HMM（数字）	16状態，混合分布数20
HMM（sil）	3状態，混合分布数36
HMM（sp）	1状態（silの第2状態と共有）

ウィーナーフィルタ係数をメル周波数領域に変換し，入力信号に適用して雑音抑圧された音声信号を得る．

第2段階では，第1段階の出力信号を入力とし，再度雑音抑圧を行う．まず，VADの結果を用いるのではなく，入力信号の開始10フレームを雑音区間とみなし，雑音のパワースペクトルを推定する．次に，入力信号のパワースペクトルと雑音のパワースペクトルの比により，逐次的に雑音のパワースペクトルを更新し，ウィーナーフィルタ係数を求める．さらに，ウィーナーフィルタ係数をメル周波数領域に変換し，別途推定したSNRを用いてウィーナーフィルタ係数を調整した後，入力信号に適用して雑音抑圧された音声信号を得る．最終的に，この音声信号から直流成分を除去したものを特徴抽出部に送る．

3. 認識実験

3.1 実験条件

認識実験には，雑音下連続数字認識タスクであるAURORA-2J[5]を用いる．AURORA-2Jの学習セットとテストセットを表1，認識実験の条件を表2に示しておく．

本実験では，学習データとテストデータの両方に対して，WI008による雑音抑圧を行っている．そして，雑音抑圧後のデータを用いて，AURORA-2Jの標準スクリプトにより学習と認識を行っている．よって，評価カテゴリ[5]は0（バックエンドの変更なし）である．なお，AURORA-2Jのベースラインフロントエンドは WI007であり，雑音抑圧は行われていない．

表1　AURORA-2Jの学習セットとテストセット
Table 1　Training and test sets of the AURORA-2J.

学習・テストセット	音声	雑音	チャネル	SNR
Clean training	110名，8,440発話	なし	G.712	Clean
Multicondition training	同上	Subway, Babble, Car, Exhibition	G.712	Clean, 20, 15, 10, 5
テストセットA	104名，4,004発話	Subway, Babble, Car, Exhibition	G.712	Clean, 20, 15, 10, 5, 0, −5
テストセットB	同上	Restaurant, Street, Airport, Station	G.712	同上
テストセットC	104名，2,002発話	Subway, Street	MIRS	同上

表3　単語正解精度（WI008）
Table 3　Word accuracy (WI008).

Clean Training (%Acc)

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.43	98.37	98.57	98.40	98.44	98.43	98.37	98.57	98.40	98.44	98.62	98.64	98.63	98.48
20 dB	97.61	98.43	98.69	97.72	98.11	96.04	96.43	97.94	98.03	97.11	97.64	96.67	97.16	97.52
15 dB	94.17	96.07	97.94	95.80	96.00	92.17	94.23	96.00	96.17	94.64	93.80	93.65	93.73	95.00
10 dB	86.18	89.18	95.17	90.34	90.22	81.46	88.27	89.02	89.36	87.03	85.69	86.91	86.30	88.16
5 dB	66.53	69.20	83.21	73.19	73.03	59.90	72.46	73.22	77.78	70.84	62.57	68.83	65.70	70.69
0 dB	36.97	31.92	48.05	37.70	38.66	23.43	44.01	40.98	49.24	39.42	32.70	40.24	36.47	38.52
-5 dB	9.06	-2.15	12.65	6.26	6.46	-6.91	13.42	8.56	11.97	6.76	9.40	16.17	12.79	7.84
Average	76.29	76.96	84.61	78.95	79.20	70.60	79.08	79.43	82.12	77.81	74.48	77.26	75.87	77.98

Multicondition Training (%Acc)

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.79	99.73	99.79	99.75	99.77	99.79	99.73	99.79	99.75	99.77	99.88	99.70	99.79	99.77
20 dB	99.63	99.58	99.76	99.38	99.59	99.32	99.43	99.31	99.38	99.36	99.60	99.58	99.59	99.50
15 dB	99.26	99.40	99.61	99.01	99.32	98.56	98.94	98.39	98.40	98.57	99.39	99.06	99.23	99.00
10 dB	98.28	98.31	98.30	97.13	98.01	94.53	96.58	94.90	95.25	95.32	97.97	96.95	97.46	96.82
5 dB	94.14	92.14	94.72	92.01	93.25	82.16	88.85	86.22	88.77	86.50	91.80	88.63	90.22	89.94
0 dB	78.94	68.77	80.14	76.09	75.99	52.13	69.07	67.97	71.34	65.13	70.13	63.18	66.66	69.78
-5 dB	44.15	25.67	43.13	43.20	39.04	8.35	33.74	28.30	37.67	27.02	33.80	29.29	31.55	32.73
Average	94.05	91.64	94.51	92.72	93.23	85.34	90.57	89.36	90.63	88.98	91.78	89.48	90.63	91.01

表4　ベースラインに対する誤り改善率（WI008）
Table 4　Relative performance (WI008).

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%

3.2　平均的な認識性能の比較評価

WI008の単語正解精度を表3に示す．ここで，表中のSNR方向の平均値（Average）は，CleanとSNR−5dBを除いて算出されている．WI008のベースライン（WI007）に対する誤り改善率を表4に示す．単語正解精度（%Acc）と誤り改善率の定義は次の通りである．

$$\%\mathrm{Acc} = \frac{H - I}{N} \times 100 \qquad (1)$$

$$\text{誤り改善率} = \frac{\%\mathrm{Acc} - \%\mathrm{Acc\ of\ Baseline}}{100 - \%\mathrm{Acc\ of\ Baseline}} \times 100 \qquad (2)$$

ここで，Nは総単語数，Hは正解単語数，Iは挿入誤り数である．単語正解精度と誤り改善率は，認識性能の評価指標としてよく用いられており，AURORA2とAURORA-2Jの評価指標としても採用されている．

表3より，Clean conditionの場合，SNR15dB以上のときは90%以上，SNR10dBのときでも90%に近い単語正解精度が得られている．また，Multicondition trainingの場合は，SNR5dBのときでもほぼ90%の単語正解精度を達成している．これは，フロントエンドでのみ雑音対策を施すという条件では，現状の最高値に近いものである．表4のベースラインに対する誤り改善率からも，WI008の性能の高さが伺える．

3.3　話者毎の認識性能の比較評価

音声認識の評価の際には，多数の話者のデータを使うのが一般的である．しかし，式（1）の単語正解精度では，話者を区別していないために，平均的な認識性能しか知ることができない．音声認識の性能は話者によって異なることが知られており，これは音声認識を用いたサービスの品質に大きな影響を与えるものである．以下では，話者毎の認識性能を比較評価する．

表5と表6に，話者毎の単語正解精度の最大値，最小値，平均値，標準偏差を示す．ここで，表5はWI007，表6はWI008のものである．表中の最大値（最小値）は，表3の各セルの単語正解精度を，話者毎の単語正解精度の中から最高（最低）のものに置き換えて得られたものである．同様に，平均値と標準偏差も，話者毎の単語正解精度の平均値と標準偏差に置き換え

表5 話者毎の単語正解精度の最大値，最小値，平均値，標準偏差（WI007）

Table 5 Maximam/minimum/average value and standard deviation of the word accuracy for each speaker (WI007).

		A	B	C	Overall
Clean training	最大値	69.53	68.75	70.17	69.35
	最小値	17.10	9.13	20.44	14.58
	平均値	45.96	43.10	49.44	45.51
	標準偏差	10.18	11.60	10.20	10.75
Multicondition training	最大値	99.05	94.92	97.53	97.09
	最小値	74.16	54.14	62.36	63.79
	平均値	91.55	80.12	85.82	85.83
	標準偏差	5.04	8.42	6.67	6.72
Average	最大値	84.29	81.84	83.85	83.22
	最小値	45.63	31.63	41.40	39.19
	平均値	68.75	61.61	67.63	65.67
	標準偏差	7.61	10.01	8.44	8.73

表6 話者毎の単語正解精度の最大値，最小値，平均値，標準偏差（WI008）

Table 6 Maximam/minimum/average value and standard deviation of the word accuracy for each speaker (WI008).

		A	B	C	Overall
Clean training	最大値	93.10	93.55	90.83	92.83
	最小値	52.50	52.89	49.21	52.00
	平均値	79.10	77.56	75.76	77.82
	標準偏差	8.24	8.38	8.32	8.31
Multicondition training	最大値	99.73	98.51	98.64	99.02
	最小値	76.47	67.27	73.90	72.27
	平均値	93.26	88.89	90.68	90.99
	標準偏差	4.80	5.94	5.08	5.31
Average	最大値	96.41	96.03	94.73	95.92
	最小値	64.49	60.08	61.56	62.14
	平均値	86.18	83.22	83.22	84.40
	標準偏差	6.52	7.16	6.70	6.81

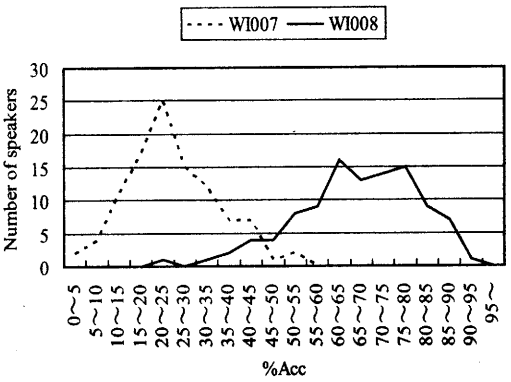

図3 話者毎の単語正解精度のヒストグラム（Clean training，テストセットA，Subway，SNR5dB）

Fig. 3 Histogram of the word accuracy for each speaker (Clean training, Testset A, Subway, SNR5dB).

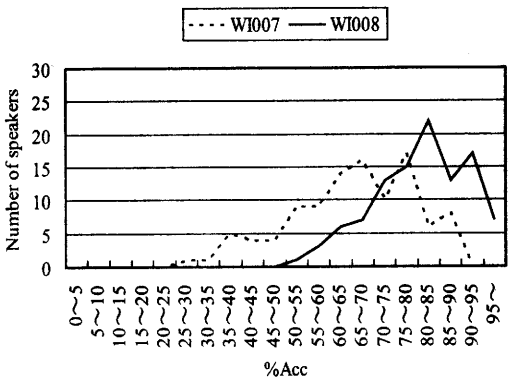

図4 話者毎の単語正解精度のヒストグラム（Multicondition training，テストセットA，Car，SNR0dB）

Fig. 4 Histogram of the word accuracy for each speaker (Multicondition training, Testset A, Car, SNR0dB).

て得られたものである．

表5と表6より，以下のことが分かる．

- WI008の最大値を見ると，Clean trainingの場合は90%以上，Multicondition trainingの場合は100%に近い値を得ており，理想的な条件の下では相当高い認識性能を達成している．一方，WI007，WI008共に，最大値と最小値の差はかなり大きく，数十%にも及んでいる．
- AverageのOverallに着目すると，WI007の最大値と最小値の差は44.03%（＝83.22－39.19）である．一方，WI008の最大値と最小値の差は33.7%（＝95.92－62.14）であり，WI007よりも差が小さくなっている．
- 表3と表6より，話者を区別せずに求めた単語正解精度と話者毎の単語正解精度の平均値には，さほど差がないことが分かる．
- WI008の標準偏差は，WI007よりも小さくなっている．WI008の標準偏差をさらに詳しく調べたところ，標準偏差はSNRが低くなるほど大きく，雑音の種類によってもその大きさにばらつきが見られた．

次に，話者毎の単語正解精度のヒストグラムを図3と図4に示す．ここで，図3は，Clean training，テストセットA，Subway，SNR5dBの場合である．また，図4は，Multicondition training，テストセットA，Car，SNR0dBの場合である．図中の横軸は単語正解精度の範囲（5%刻み），縦軸は話者の数である．

図3と図4より，WI008の分布は，WI007と比べて単語正解精度が高い方に比重を移していることが分かる．その一方で，WI008の分布は，中心付近から左側への広がりが大きく，単語正解精度が依然として低い話者が相当数存在していることが見て取れる．

音声認識の運用時には，単語認識精度にある目標値を設定した

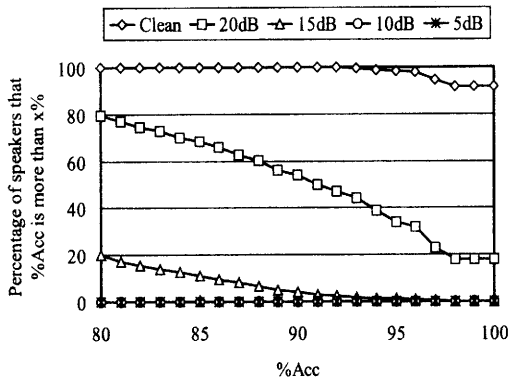

図5 単語正解精度がx%以上の話者の割合 (Clean training, WI007)

Fig. 5 Percentage of speakers that the word accuracy is more than x% (Clean training, WI007).

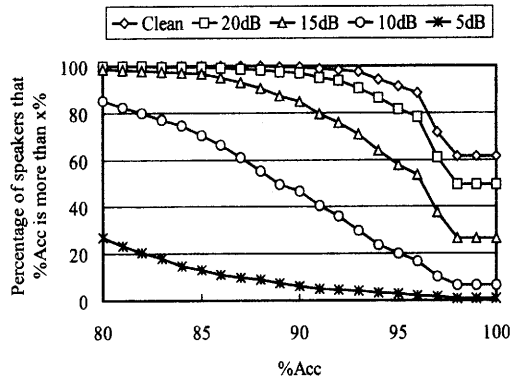

図7 単語正解精度がx%以上の話者の割合 (Clean training, WI008)

Fig. 7 Percentage of speakers that the word accuracy is more than x% (Clean training, WI008).

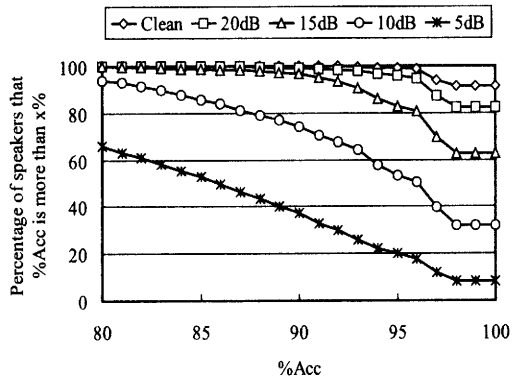

図6 単語正解精度がx%以上の話者の割合 (Multicondition training, WI007)

Fig. 6 Percentage of speakers that the word accuracy is more than x% (Multicondition training, WI007).

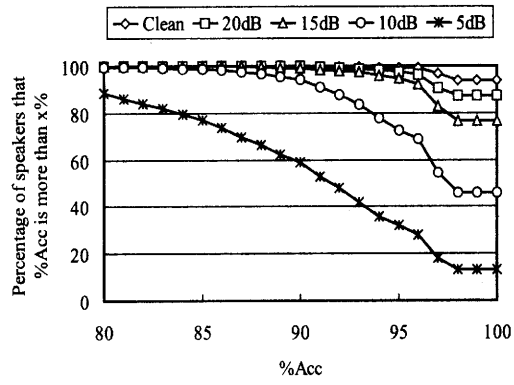

図8 単語正解精度がx%以上の話者の割合 (Multicondition training, WI008)

Fig. 8 Percentage of speakers that the word accuracy is more than x% (Multicondition training, WI008).

とき，それをどのくらいの割合のユーザに対して確保できるかが重要な指標となる．図5と図6 (WI007)，図7と図8 (WI008) に，単語正解精度がx%以上の話者の割合を示す．ここで，図5と図7はClean trainingの場合，図6と図8はMulticondition trainingの場合である．

図より，以下のことが分かる．

- 単語正解精度の目標値xを高くするほど，その単語正解精度が得られる話者の割合が減っている．特に，目標値が100%に近いときの落ち込みが大きい．
- 一例として，$x = 90$のときに90%の話者を確保できるSNRに着目すると，WI007では，Clean trainingの場合はCleanのみ，Multicondition trainingの場合はSNR15dB以上である．一方，WI008では，Clean trainingの場合はSNR20dB以上，Multicondition trainingの場合はSNR10dB以上であり，適用可能なSNRの範囲が広がっている．
- Clean trainingの場合のCleanのとき，WI008では，WI007と比べて単語正解精度の目標値xを高くしたときの話者の割合が落ち込んでいる．これは，雑音抑圧処理によってクリーンな音声にひずみが生じ，その結果単語正解精度が僅かに低下しているからであると考えられる．

4. むすび

本稿では，AURORA-2Jを用いてWI007とWI008の性能を比較評価した．その際，AURORA2やAURORA-2Jで採用している平均的な認識性能を表す評価指標に加えて，話者毎の認識性能を表す評価指標を用いた．具体的には，話者毎の単語正解精度の最大値，最小値，平均値，標準偏差，話者毎の単語正解精度のヒストグラム，単語正解精度がx%以上の話者の割合である．その結果，WI008の認識性能は，WI007を大幅に上回っていることが確認できた．その一方で，話者毎の認識性能

には，さらなる改善が必要であることが分かった．

雑音環境下で音声認識を用いたサービスの品質を確保するには，雑音抑圧手法を含む雑音対策にさらなる研究開発が必要である．そして，その際には，様々な指標を用いて多角的に評価することが重要である．今後は，ユーザの主観による認識性能を反映する客観的な評価指標の研究を進める予定である．

なお，本稿で述べた評価指標をAURORA-2Jの評価に用いるためのプログラムを，IPSJ SIG-SLP雑音下音声認識評価ワーキンググループのホームページ[4]にて配布する予定である．

謝辞

本研究の一部は，総務省戦略的情報通信研究開発推進制度，及び通信・放送機構の研究委託による．本研究では，IPSJ SIG-SLP雑音下音声認識評価WGの雑音下音声認識評価環境（AURORA-2J）を利用した．

文　献

[1] H.-G. Hirsch, D. Pearce, "The AURORA Experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, 2000.

[2] ETSI ES 201 108 v1.1.2, "Distributed speech recognition; front-end feature extraction algorithm; compression algorithm," 2000.

[3] ETSI ES 202 050 V1.1.1, "Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[4] http://sp.shinshu-u.ac.jp/AURORA-J/

[5] 山本一公，中村哲，武田一哉，黒岩眞吾，北岡教英，山田武志，水町光徳，西浦敬信，藤本雅清，"AURORA-2J/AURORA-3Jデータベースとその評価ベースライン，" 情報処理学会研究報告，SLP-47-19, 2003.