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

2003－ＳＬＰ－49　(19)
2003／12／18

複数の雑音抑圧手法を用いた認識結果の統合による ロバスト音声認識の検討

岡田　治郎†　　山田　武志††　　北脇　信彦††

† 筑波大学大学院システム情報工学研究科
†† 筑波大学電子・情報工学系
〒305–8573 茨城県つくば市天王台 1–1–1
E-mail: †okajiro@mmlab.is.tsukuba.ac.jp, ††{takeshi,kitawaki}@is.tsukuba.ac.jp

あらまし　様々な雑音条件に対してロバストな音声認識を実現するためには，複数の雑音抑圧手法の各々の認識結果から信頼度の高いものを選択する方法が有効であると考えられる．本稿では，フレーム正規化対数尤度に基づく信頼度を用いた認識結果の統合法を提案する．提案法では，正解と最もマッチする認識結果を得るために，各雑音抑圧手法を用いた時の N-best の認識結果を求める．そして，各認識結果に対して信頼度を付与し，信頼度が最大になるものを選択する．提案法の有効性を評価するために，雑音下連続数字認識タスクである AURORA-2J を用いて評価実験を行った，その結果，特に Multicondition training の場合に提案法の有効性を確認することができた．

キーワード　雑音下音声認識，複数の雑音抑圧手法，認識結果の統合，信頼度尺度

Integration of Recognition Results for Robust Speech Recognition

Jiro OKADA†, Takeshi YAMADA††, and Nobuhiko KITAWAKI††

† Graduate School of Systems and Information Engineering, University of Tsukuba
†† Institute of Information Sciences and Electronics, University of Tsukuba
1–1–1, Tennodai, Tsukuba, 305–8573 Japan
E-mail: †okajiro@mmlab.is.tsukuba.ac.jp, ††{takeshi,kitawaki}@is.tsukuba.ac.jp

Abstract　To achieve high recognition performance in noisy environments, we propose the integration of recognition results obtained by using four noise reduction algorithms (spectral subtraction with smoothing of time direction, temporal domain SVD-based speech enhancement, GMM-based speech estimation and KLT-based comb-filtering). In the proposed method, the log likelihood-based confidence measure is used to select the best recognition result. Experimental results on the AURORA-2J connected digit recognition task confirmed that the proposed algorithm achieves high recognition performance especially in the multicondition training.

Key words　noisy speech recognition, multiple noise reduction algorithms, integration of recognition results, confidence measure

1. まえがき

近年，統計的手法を用いることにより，音声認識性能は飛躍的に向上している．しかし，音声認識の利用が想定される環境には数多くの周囲雑音が存在し，特にマイクから離れて発話する場合には認識性能が著しく低下することが問題となっている．

従来，雑音に対してロバストな音声認識を実現するために，様々な雑音抑圧手法が提案されている．しかし，これらの手法の性能は雑音条件に依存することが多く，様々な種類の雑音を広範囲の SNR に渡って抑圧することは，極めて難しいことが知られている．このような現状において，幅広い雑音条件の下で最高の認識性能を得るためには，複数の雑音抑圧手法を統合することが有効であると考えられる．統合による認識性能の改善方法には，雑音抑圧手法の出力信号を手がかりにして音声認識の前に手法を選択する方法，各雑音抑圧手法の複数の認識結果から信頼度の高いものを選択する方法などがある．文献 [1] [2] では，前者の方法に関し，雑音条件毎に最適な雑音抑圧手法を選択することにより，大幅な性能改善を達成できることが報告されている．

本稿では，後者の方法に着目し，フレーム正規化対数尤度に

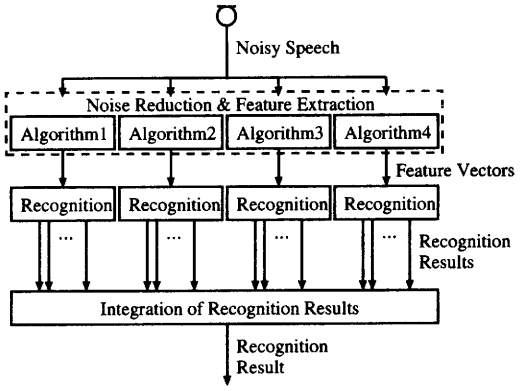

図 1　提案法の処理フロー

Fig. 1　Process flow of the proposed method.

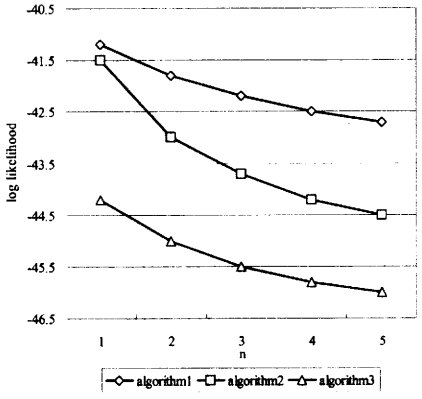

図 2　各雑音抑圧手法における順位 n とフレーム正規化対数尤度の関係

Fig. 2　*n-loglikelihood* relationship for each algorithm.

基づく信頼度を用いた認識結果の統合法を提案する．提案法では，正解と最もマッチする認識結果を得るために，各雑音抑圧手法を用いた時の N-best の認識結果を求める．そして，各認識結果に対して信頼度を付与し，信頼度が最大になるものを選択する．提案法の有効性を評価するために，雑音下連続数字認識タスクである AURORA-2J [3] を用いて評価実験を行ったので，その結果を報告する．

2. 提 案 法

2.1 提案法の概要

提案法の処理フローを図 1 に示す．提案法では，まず，音声波形を各雑音抑圧手法の入力として与え，雑音抑圧と特徴量抽出を行う．そして，各々の特徴量を用いて音声認識を行う．ここで，各認識部は，フレーム正規化対数尤度の高い順に N 個（N-best）の認識結果を出力する．統合部では，フレーム正規化対数尤度に基づいて，雑音抑圧手法 m の順位 n における認識結果 $U(m,n)$ の信頼度 $S(m,n)$ を計算しておく．複数の雑音抑圧手法を用いるので，同じ認識結果が複数の手法に渡って幾つかの順位に出現することがある．そこで，認識結果 U の信頼度を次式で定義する．

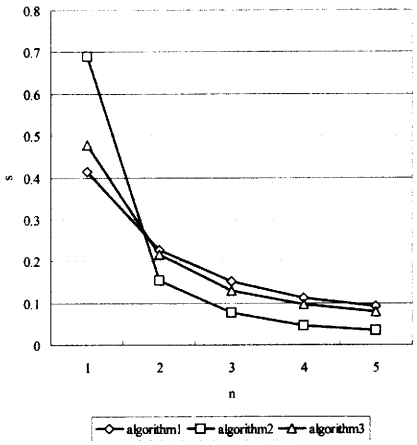

図 3　各雑音抑圧手法における順位 n と s の関係

Fig. 3　n-s relationship for each algorithm.

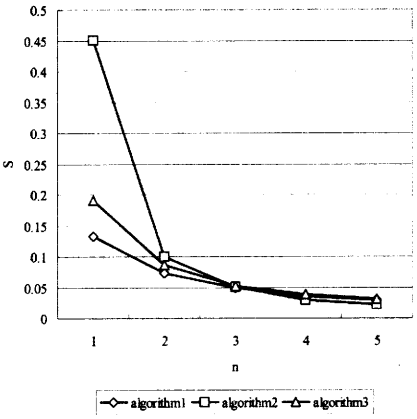

図 4　各雑音抑圧手法における順位 n と S の関係

Fig. 4　n-S relationship for each algorithm.

$$Score(U) = \sum_{m,n\,:\,U(m,n)=U} S(m,n),$$

$$1 \leqq m \leqq M,\ 1 \leqq n \leqq N \tag{1}$$

最終的な認識結果は，次式のように最も信頼度の高いものとして決定される．

$$U_{result} = \underset{U}{\operatorname{argmax}}\, Score(U) \tag{2}$$

提案法においては，信頼度をどのように設定するかが重要である．雑音抑圧手法を用いる場合，雑音抑圧されたデータを用いて学習を行うことが多く，手法毎に音響モデルが異なるため，フレーム正規化対数尤度を手法間で正規化する必要がある．次節では，信頼度の算出方法について述べる．

2.2 信 頼 度

雑音抑圧手法 m の順位 n における信頼度 $S(m,n)$ を，次式で定義する．

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

$$S(m,n) = s(m,n) \times \{s(m,1) - s(m,N)\},$$

$$1 \leqq m \leqq M, 1 \leqq n \leqq N \tag{3}$$

$$s(m,n) = \frac{\exp\{l(m,n)\}}{\sum_{i=1}^{N} \exp\{l(m,i)\}},$$

$$1 \leqq m \leqq M, 1 \leqq n \leqq N \tag{4}$$

ここで，$l(m,n)$ は，雑音抑圧手法 m の順位 n におけるフレーム正規化対数尤度である．

まず，式 (4) の意味を説明する．式 (4) は，フレーム正規化対数尤度を手法間で正規化することに相当する．具体的には，各雑音抑圧手法における N 個の認識結果において，順位 n の認識結果が全体においてどのくらい信頼できるかを表す値である．図 2 は，$N = 5$ のときのフレーム正規化対数尤度 $l(m,n)$ の模式図である．横軸は各手法内の順位を表し，縦軸はフレーム正規化対数尤度である．図 2 のような状況において，式 (4) を適用した結果を図 3 に示す．

次に，式 (3) の意味を説明する．式 (3) は，各手法の中で，下位の認識結果の信頼度に，重み付けを行うことに相当する．具体的には，上位の信頼度が高い場合は上位と下位の信頼度の関係を保持し，上位の信頼度が低い場合に関しては，相対的に下位の信頼度を上げている．図 3 のような状況において，式 (3) を適用した結果を図 4 に示す．

3. 認識実験

3.1 実験条件

本実験では，提案法において以下の 4 つの雑音抑圧手法を用いる．

(S)　時間方向スムージングを用いたスペクトルサブトラクション法 [4]
(T)　時間領域 SVD に基づく音声強調 [5]
(G)　GMM に基づく音声信号推定 [5]
(K)　ピッチ同期 KLT [6]

認識実験には，雑音下連続数字認識タスクである AURORA-2J [3] を用いる．AURORA-2J の学習セットとテストセットを表 1 に，認識実験の条件を表 2 に示しておく．学習と認識には，AURORA-2J に添付されている標準スクリプトを用いている．ベースラインと唯一異なるのは，特徴量の計算の際に CMN を適用していることである．よって，評価カテゴリ [3] は 0（バックエンドの変更なし）である．本実験では，学習データに対しても認識時と同様の雑音抑圧処理を行っている．

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
高域強調	$1 - 0.97z^{-1}$
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）

表 4　$N = 20$ としたときの Relative performance
Table 4　Relative performance when $N = 20$.

Relative performance				
	A	B	C	Overall
Clean Training	61.70%	61.94%	54.33%	60.43%
Multicondition training	29.03%	50.45%	40.88%	43.36%
Average	45.36%	56.20%	47.60%	51.90%

3.2 実験結果

まず，提案法において N を変化させたときの単語正解精度を図 5 と図 6 に示す．図 5 は Clean training の場合，図 6 は Multicondition training の場合である．各図には，4 つの雑音抑圧手法単体の単語正解精度も示している．ここで，単語正解精度は全体の平均である．

図から，Clean training の場合，Multicondition training の場合共に，N が増加するにつれ単語正解精度が高くなっていることが分かる．また，手法単体で最も単語正解精度が良かった手法 (G) と比較すると，Clean training の場合では $N = 20$ 付近においてわずかながら改善していることが分かる．また，Multicondition training の場合，$N = 2$ のときにすでに手法単体で最も単語正解精度が良かった手法 (S) から改善が見られる．その一方で，$N = 10$ を越えたあたりからは単語正解精度がほぼ横ばいとなっている．これは，Multicondtion training の場合は，正解が上位の候補に含まれていることが多いからであると考えられる．表 3 と表 4 に，$N = 20$ とした時の単語正解精度，及び AURORA-2J ベースライン性能からの Relative performance を示しておく．

最後に，複数の雑音抑圧手法を統合することにより得られる最

表 3　$N = 20$ としたときの提案法の単語正解精度

Table 3　%Acc of the proposed algorithm when $N = 20$.

Clean Training (%Acc)	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.97	99.97	100.00	99.94	99.97	99.97	99.97	100.00	99.94	99.97	99.97	99.97	99.97	99.97
20 dB	99.02	99.67	99.67	99.48	99.46	99.05	99.03	99.05	98.55	98.92	99.05	99.09	99.07	99.17
15 dB	96.62	98.52	99.16	98.15	98.11	96.87	97.49	97.85	97.07	97.32	96.68	97.79	97.24	97.62
10 dB	89.68	93.17	92.69	94.66	92.55	89.16	91.26	92.31	89.66	90.60	89.38	91.63	90.51	91.36
5 dB	70.53	72.19	66.90	80.28	72.48	70.46	71.95	76.41	67.76	71.65	66.07	69.41	67.74	71.20
0 dB	36.94	34.55	25.14	43.23	34.97	33.62	35.16	40.29	30.61	34.92	30.21	31.86	31.04	34.16
-5 dB	12.40	9.31	8.29	15.21	11.30	6.20	11.06	15.15	6.48	9.72	10.99	12.61	11.80	10.77
Average	78.56	79.62	76.71	[illegible]	79.51	[illegible]	[illegible]	[illegible]	[illegible]	78.68	76.28	77.96	77.12	78.70

Multicondition Training (%Acc)	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	100.00	99.82	99.91	99.91	99.91	100.00	99.82	99.91	99.91	99.91	99.88	99.79	99.84	99.90
20 dB	99.75	99.79	99.82	99.75	99.78	99.54	99.73	99.64	99.63	99.64	99.79	99.52	99.66	99.70
15 dB	99.54	99.55	99.73	99.38	99.55	99.14	99.33	99.02	98.89	99.10	99.42	99.33	99.38	99.33
10 dB	99.20	98.82	98.57	98.86	98.86	96.71	97.91	96.60	95.43	96.66	98.34	97.55	97.95	97.80
5 dB	96.35	94.11	94.42	93.89	94.69	88.21	91.60	89.92	88.58	89.58	93.40	90.90	92.15	92.14
0 dB	82.71	72.07	72.59	80.84	77.05	59.23	72.52	67.34	66.65	66.44	73.04	64.93	68.99	71.19
-5 dB	42.55	27.18	25.02	39.99	33.69	13.85	31.32	30.69	24.53	25.10	30.27	26.30	28.29	29.17
Average	95.51	92.87	93.03	94.54	93.99	88.57	92.22	90.50	89.84	90.28	92.80	90.45	91.62	92.03

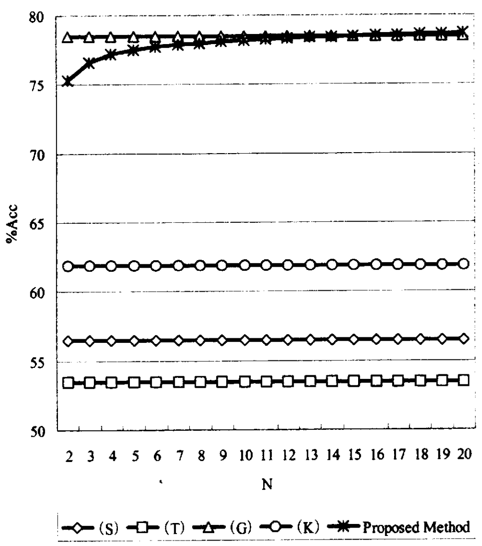

図 5　提案法の単語正解精度と N の関係 (Clean training)

Fig. 5　N-%Acc relationship for the proposed algorithm (Clean training).

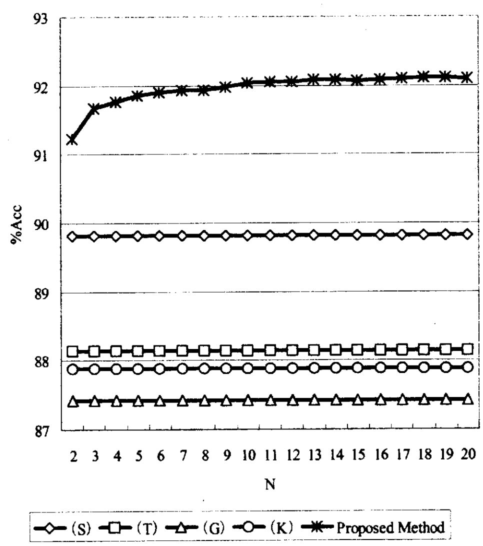

図 6　提案法の単語正解精度と N の関係 (Multicondition training)

Fig. 6　N-%Acc relationship for the proposed algorithm (Multicondition training).

高性能を調べる．N と候補となる認識結果から最も正解にマッチするものを選択した場合の単語正解精度の関係を図 7 と図 8 に示す．図 7 は Clean training の場合，8 は Multicondition training の場合である．図中の "all" とは，4 つの手法のそれぞれ上位 N 個の認識結果 ($4 \times N$ 個) を用いたときである．また，(S)，(T)，(G)，(K) はそれぞれ手法単体で上位 N 個の認識結果を用いたときである．

この結果より，手法単体では，N を大きくすることにより単語正解精度を最大で 15%程度改善できることが分かる．また，all の単語正解精度は，Clean training の場合は $N = 20$ 程度，Multicondition training の場合は $N = 10$ 程度で横ばいになっていることが分かる．そのとき，Clean training の場合で 95%弱，Multicondition training の場合で 99%の単語正解精度弱が得られることが見て取れる．

さらに，この結果を SNR 毎に分析する．all を用いるときの，N と候補となる認識結果から最も正解にマッチするものを選択した場合の単語正解精度の SNR 毎の関係を図 7 と図 8 に示す．ここで，図 9 は Clean training の場合，図 10 は Multicondition training の場合である．それぞれの値は，SNR 毎の平均の単語正解精度である．

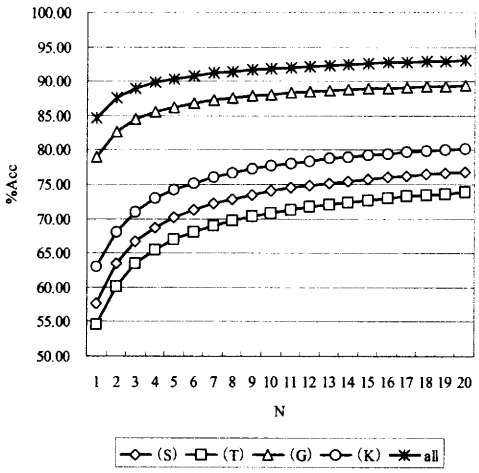

図 7　最も正解にマッチする認識結果を選択した場合の N と単語正解精度の関係（Clean training）

Fig. 7　N-%Acc relationship when the best result is selected (Clean training).

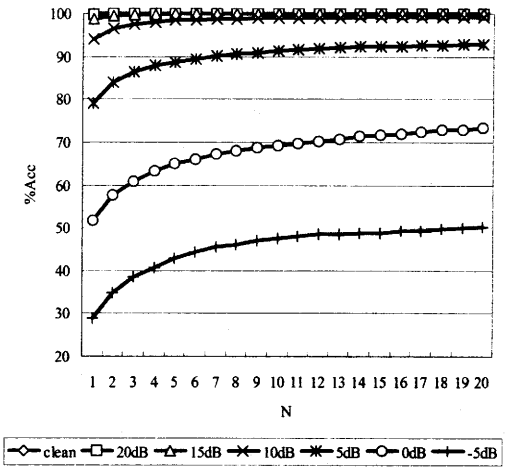

図 9　最も正解にマッチする認識結果を選択した場合の N と SNR 毎の単語正解精度の関係（Clean training）

Fig. 9　N-%$AccforeachSNRvalue$ relationship when the best result is selected (Clean training).

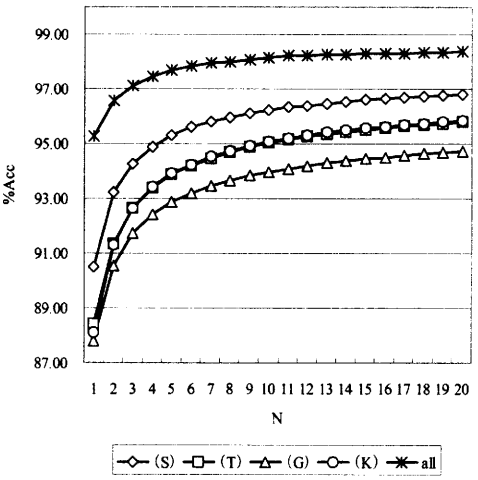

図 8　最も正解にマッチする認識結果を選択した場合の N と単語正解精度の関係（Multicondition training）

Fig. 8　N-%Acc relationship when the best result is selected (Multicondition training).

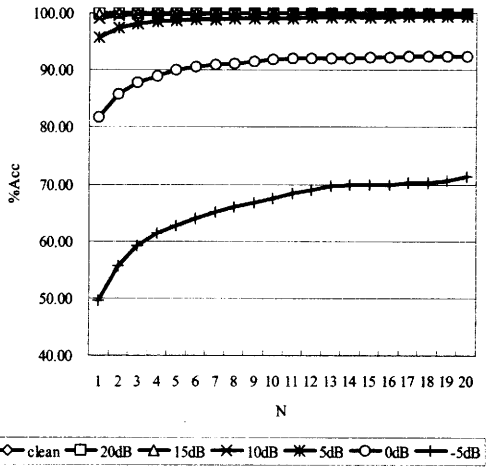

図 10　最も正解にマッチする認識結果を選択した場合の N と SNR 毎の単語正解精度の関係（Multicondition training）

Fig. 10　N-%$AccforeachSNRvalue$ relationship when the best result is selected (Multicondition training).

これらを見ると，Clean training の場合，Multicondition training の場合共に高 SNR では N を大きくしても効果が少ないことが分かる．一方，低 SNR では単語正解精度は N が増加すると共に高くなっている．これらは，高 SNR においては，正解に最もマッチする認識結果は上位に集まっており，低 SNR においては，広い順位に散らばっていることを示している．

4. むすび

本稿では，さまざまな雑音条件下でのロバストな音声認識を実現するために．4 つの雑音抑圧手法の複数の認識結果から最適な認識結果を選択する手法について検討し，その有効性を AURORA-2J を用いて調べた．その結果，Clean training では手法単体に比べ平均的な単語正解精度がわずかながらに改善し，Multicondition training では大幅に改善することが分かった．一方，最も正解にマッチする認識結果が選択可能な場合と比較すると，提案法の性能はまだ不十分であることも分かった．今後，信頼度尺度に関する更なる検討が必要である．

謝辞

音声データや各手法のプログラムをご提供頂いた，武田一哉氏，北岡教英氏，藤本雅清氏に感謝する．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] T. Yamada, J. Okada, K. Takeda, N. Kitaoka, M. Fujimoto, S. Kuroiwa, K. Yamamoto, T. Nishiura, M. Mizumachi, S. Nakamura, "Integration of Noise Reduction Algorithms for Aurora2 Task," Proc. Eurospeech2003, pp.1769–1772, 2003.

[2] 山田武志，岡田治郎，武田一哉，北岡教英，藤本雅清，黒岩眞吾，山本一公，西浦敬信，水町光徳，中村哲，"雑音下音声認識のための複数の前処理手法の統合とその AURORA-2J による評価，" 情報処理学会研究報告，SLP-47-18，2003.

[3] 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.

[4] 北岡教英，赤堀一郎，中川聖一，"スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識，" 電子情報通信学会論文誌，Vol.J83-D-II，pp.500–508，2000.

[5] 藤本雅清，有木康雄，"GMM に基づく音声信号推定法を用いた雑音下音声認識，" 信学技報 SP2002，pp.25–30，2002.

[6] S.-J. Park, M. Ikeda, K. Takeda, F. Itakura, "Improvement of the ASR Robustness using Combinations of Spectral Subtraction and KLT based Adaptive Comb-filtering," IPSJ SIGNotes, SLP-44-3, pp.13–18, 2002.