話者正規化スペクトルサブバンドパラメータを用いた雑音下での音声認識

柘植 覚†‡　深田 俊明†　シンガー ハラルド†

†ATR 音声翻訳通信研究所

‡徳島大学

†〒 619-0288　京都府相楽郡精華町光台 2-2

Tel.: 0774-95-1397　e-mail: {stsuge, fukada, singer}@itl.atr.co.jp

あらまし　本稿では、雑音下での音声認識における補助的特徴量として、話者正規化 SSC (spectral subband centroids) を提案する。SSC は、サブバンド内に含まれる音声パワースペクトルのセントロイド周波数として定義される。この特徴量は、雑音環境下においても比較的変動の少ない、スペクトルのピーク (フォルマント) が示す周波数をおおまかにとらえるため、雑音に対してロバストな特徴量であると考えられる。SSC はスペクトルのピークが示す周波数に依存する特徴量のため、スペクトル形状の異なる複数話者から求めた SSC の分布は広がり、異なる音素の分布間に大きな重なりが生じると考えられる。そこで、この分布の重なりを低減するため、話者正規化手法を SSC の計算に取り入れた話者正規化 SSC を提案する。自由発話音声を用いた連続音声認識実験により、話者正規化 SSC を補助的特徴量として用いた場合、20.3% (SNR=15dB) の誤り改善率を得ることができた。また、話者正規化手法を用いない SSC との比較においても、14.3% (SNR=15dB) の誤り改善率を得ることができた。

キーワード　スペクトルサブバンドセントロイド, 雑音環境, 話者正規化, 音声認識

Speaker normalized spectral subband parameters for noise robust speech recognition

Satoru Tsuge†‡　Toshiaki Fukada†　Harald Singer†

†ATR Interpreting Telecommunications Research Laboratories

‡Tokushima University

†2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288

Tel.: 0774-95-1397　e-mail: {stsuge, fukada, singer}@itl.atr.co.jp

Abstract　This paper proposes speaker normalized spectral subband centroids (SSCs) as supplementary features in noise environment speech recognition. SSCs are computed as frequency centroids for each subband from the power spectrum of the speech signal. This feature can be obtained reliably even under noisy conditions because SSC are mainly computed from spectral peaks such as formants whose positions are almost unchanged in a noisy environment. Since the conventional SSCs depend on formant frequencies of a speaker, the distributions of SSCs computed from large amounts of speakers will be highly overlapped between different phones. Therefore, we introduce a speaker normalization technique into SSC computation to reduce the speaker variability. Experimental results on spontaneous speech recognition show that the speaker normalized SSCs are more useful as supplementary features for improving the recognition performance than the conventional SSCs. We observed a significant improvement in error rate by 20.3% and 14.3% at SNR=15dB by adding speaker normalized SSCs to the conventional features and by incorporating a speaker normalized technique into the conventional SSCs, respectively.

key words　Spectral subband centroids, Noise environment, Speaker normalization, Speech recognition

1 まえがき

音声認識システムは、音声波形から後の認識に有益な特徴パラメータを抽出する信号処理部(フロントエンド)を含んでいる。特徴パラメータは、直接音声認識システムの認識性能に影響を与えるため、特徴パラメータを適切に選択することは認識システム構成において重要な問題である[1][2]。

現在、音声認識システムに有益な特徴パラメータとして、線形予測分析、フィルタバンクから計算されるケプストラム係数が広く用いられている[3]。これらの特徴パラメータは、音響モデル作成のための学習データとそのモデルを用い認識を行う評価データの発声環境が同一の条件の場合、認識に非常に有効な情報を与えることがわかっている。しかし、実環境において、音声認識システムを使用する場合、背景雑音、マイクからの位置のずれ等の影響により、学習データの発声環境と評価データの発声環境の間にミスマッチが生じてしまう。このような環境のミスマッチにより、入力音声のケプストラムと学習データから構築された音響モデルとの間にはずれが生じ、認識性能の劣化を引き起す。

これらの環境のミスマッチに対して、頑健な音声認識手法の研究が多数行われている。例えば、加算性雑音に対する雑音対策手法として、スペクトルサブトラクション法(SS)[4]やParallel Model Combination (PMC)法[5]などが提案されている。また乗算性雑音に対する手法として、ケプストラム平均を減算するケプストラム平均正規化手法等[6]が挙げられる。SSは入力音声のスペクトルから推定雑音のスペクトルを除去する方法であり、PMC法はモデル側において、雑音モデルを重畳することにより雑音処理を行っている。

一方、これらの手法以外に環境のミスマッチに対応するため、環境のミスマッチに変動の少ない頑健な特徴量を音声波形から抽出し、特徴パラメータとして用いるRelative Spectra (RASTA)法ある[7]。しかし、これらの手法は、雑音比や、ノイズの推定スペクトラム等の知識が必要となる。このような雑音に対する知識を必要とせず、環境のミスマッチに頑健な音声認識に有効な特徴パラメータとしてスペクトル・サブバンド・セントロイド(Spectral Subband Centroid : SSC)が近年、提案されている[8][9]。

SSCは、周波数帯を複数のサブバンドに分割し、各サブバンド内に含まれる音声波形のパワースペクトルを用い、サブバンド内のセントロイド周波数として計算される。このセントロイド周波数は、各サブバンド内のパワースペクトルのピークが示す周波数をおおまかにとらえる。これらのスペクトルのピークが示す周波数は、雑音下においても変動が少ないと考えられる。つまり、環境のミスマッチが生じても、学習時の発音環境に近い特徴量が抽出できる。このため、SSCを補助的パラメータとして用いることにより、環境のミスマッチによる認識性能の劣化が低減できると考えられる。

実際に、SSCの有効性は、文献[9]に示されている。しかし、この文献で行われた実験は、特定話者に対するアルファベット音声認識と非常に小さいタスクで行われている。さらに、SSCが雑音環境下においてもクリーン環境下に近い特徴量を表す分析例は示されているが、音声認識実験によるSSCの雑音環境下における有効性は示されていない。このため、我々は、自然発話に対し、連続単語認識実験を行いSSCが雑音による認識性能の劣化を低減できることを報告した[10]。

ところが、SSCはスペクトルのピークが示す周波数を表現する特徴量であるため、スペクトル形状が異なる複数話者から求めたSSCの分布は大きく広がり、異なる音素の分布間に大きな重なりが生じてしまうことが予想される。そのため、不特定話者に対する音声認識においては、SSCの有効性が減少してしまうと考えられる。そこで、我々は、話者により異なるスペクトル形状の正規化を行いSSCを計算する、話者正規化SSCを提案する。

本稿では、SSC、話者正規化SSCの雑音環境下における有効性を示すため、言語モデルを用いた単語認識率による評価を行った。以下、**2**ではSSC、**3**では話者正規化SSC、**4**では認識実験について、**5**では本稿のまとめを述べる。

2 スペクトル・サブバンド・セントロイドを用いた雑音下での音声認識

2.1 スペクトル・サブバンド・セントロイド

SSC[8][9]は、周波数帯$[0, F_s/2]$ (F_s: サンプリング周波数)をM個のサブバンドに分割を行い、各サブバンドに含まれる音声信号のパワースペクトルを用い、各サブバンドのセントロイド周波数として次式により定義される。

$$C_m = \frac{\int_{l_m}^{h_m} f \cdot P^{\gamma}(f) df}{\int_{l_m}^{h_m} P^{\gamma}(f) df} \quad (1)$$

ここで、fは周波数、$P(f)$は周波数fにおけるFFTパワースペクトルを示す。γはパワースペクトル・ダイナミックレンジ・コントロール変数であり、$0 < \gamma$の定

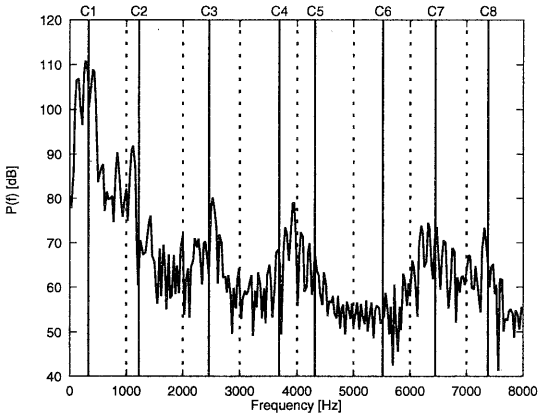


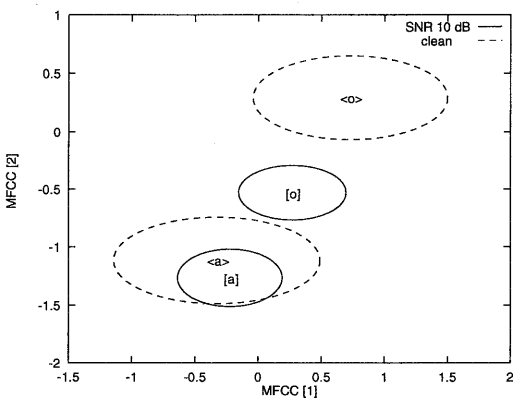
図 1: SSC の分析例 (日本語母音 /u/、 $M = 8$、 $\gamma = 0.5$)

数である。例えば、$\gamma = 0.5$ とすると、C_m は振幅スペクトルから計算される各サブバンドのセントロイド周波数となる。また、l_m、h_m は m 番目のサブバンドの開始および、終了周波数を表す。各サブバンドの開始、終了周波数は、重なりなく周波数帯を等分割した場合、$l_1 = 0$、$h_M = F_s/2$、$l_{m+1} = h_m = m * F_s/(2 * M)(m = 1, 2, \ldots, M-1)$ となる。本稿では、周波数帯を等分割したサブバンドから SSC を計算したが、メルスケール、バークスケールを用い分割したサブバンドからの計算も可能である。また、FFT パワースペクトル以外にスペクトル包絡から SSC を計算することも可能である。

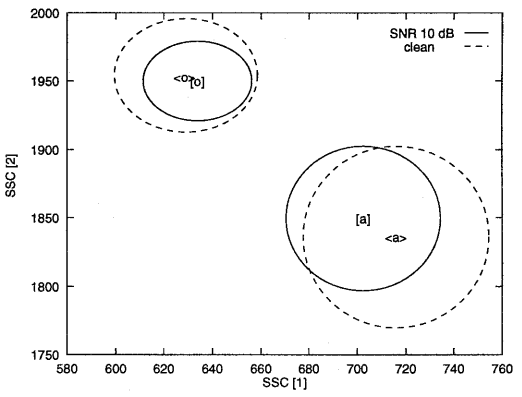
2.2 SSC の分析例

図 1に、成人男性がクリーンな環境下で発声を行った日本語母音 /u/ をサンプリング周波数 16kHz で標本化し、その標本化された音声に対するパワースペクトルを示す。この図において、周波数帯 [0, 8kHz] を $M = 8$ に重なりなく等分割を行ったサブバンド ($l_1 = 0$Hz, $h_1 = l_2 = 1000$Hz, $h_2 = l_3 = 2000$kHz, ...) から、$\gamma = 0.5$ とし式 (1) を用い計算を行った SSC が示した周波数を縦の実線で、分割したサブバンドを縦の点線で示した。この図より、いくつかの SSC (C_1, C_2, C_3 等) は、フォルマント周波数をおおまかにとらえていることがわかる。また、この分析例以外のデータにおいても、おおまかはあるが、SSC はフォルマントが示す周波数を表していた。

また、雑音による特徴量の変動を調べるため、図 2に特徴量が雑音により変動する例を示す。図 2の上図 (a) は、成人男性が発声した自然発話音声中に出現した日本語母音 /a/, /o/ に対する従来用いられている特徴量である



(a) MFCCs



(b) SSC

図 2: 男性 1 話者に対する日本語母音 /a/ と /o/ の分布

MFCC (Mel Frequency Cepstral Coefficients) の低次 (1 次、2 次) の分布を示している。同様に下図 (b) に $M = 6$、$\gamma = 0.5$ とし、式 (1) で計算を行った SSC の低次 (1 次、2 次) の分布を示す。図中の実線は、クリーンな環境下で発声した音声の分布、点線はその発声に SNR=10dB になるように計算機雑音を重畳した音声から求めたそれらの分布である。図中に示した分布は、全て $\mu \pm 0.5\sigma$ の範囲を楕円で表した (μ: 平均、σ: 標準偏差)。また、図中の <>、[ ] はそれぞれ、クリーンな発声に対する分布の平均、雑音を重畳した発声に対する分布の平均を示す。

図 2から、MFCC は、クリーンな環境で発声した音声とその音声に雑音を重畳した音声から求めた分布間にずれが生じていることがわかる。雑音等の影響により発声環境が変動した場合、クリーンな環境に近い特徴を示すことができず、音響モデルが示す特徴との間にずれが生

じ、認識性能が劣化すると考えられる。しかし、SSC は、雑音の重畳した音声に対しても、分布の変動が少ない。そのため、雑音下においてもクリーン環境と同様の特徴を保て、環境の変動に対する認識性能の劣化を抑えることができると予想される。

3 話者正規化 SSC

前節において、SSC はスペクトルのピークが示す周波数をおおまかにとらえていることを示した。このスペクトルピークが示す周波数は、環境の変動に対し変化が少ないため、この周波数をおおまかに示す SSC は、環境のミスマッチに対して頑健な特徴量であると考えられる。しかし、スペクトルのピークは話者に依存する値であるため、複数話者から得られる SSC は、話者により大きく異なる値を示すことが予想される。そのため、各音素の分布が広がり、異なる音素間における分布に非常に大きい重複がおこることが予想される。不特定話者音声認識の場合では、SSC の有効性がこのような重複によって減少していると考えられる。そこで、我々は、これらの話者性を減少させ、分布の重なりを低減するために、SSC の計算式に話者正規化手法 [11][12] を取り入れた話者正規化 SSC を提案する。本稿では、話者正規化手法に周波数軸を伸縮する周波数ワーピングに基づく声道長正規化を用いた。

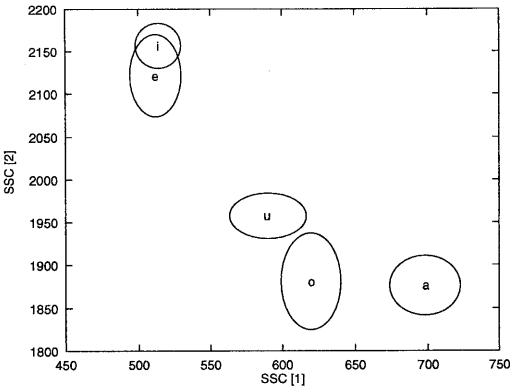
周波数軸の伸縮を決定する周波数ワーピング係数 α を式 (2) により計算を行う。

$$\alpha = \frac{\sum_{p \in \mathbf{P}} F_{c,p}}{\sum_{p \in \mathbf{P}} F_{s,p}} \tag{2}$$

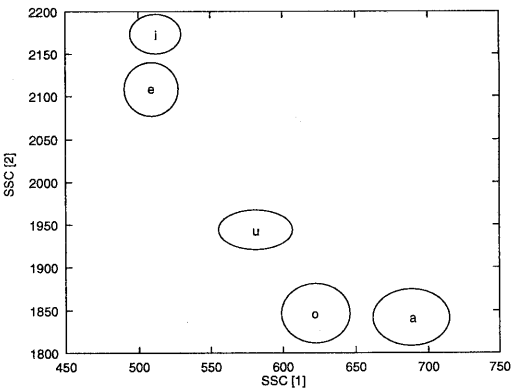
ここで、$\mathbf{P}$ は日本語母音 {/a/, /i/, /u/, /e/, /o/} を示す。また、$F_{c,p}$ は、各母音 p の学習に用いた全話者 c の平均第 2 フォルマント周波数を示し、$F_{s,p}$ は、各母音 p に対する正規化を行う各話者 s の平均第 2 フォルマント周波数を示す。この式より、計算した周波数ワーピング係数 α を用い、話者正規化 SSC ($\bar{C}_m$) を式 (3) と定義する。

$$\bar{C}_m = \frac{\int_{l_m}^{h_m} f \cdot P^{\gamma}(\alpha^{-1} f) df}{\int_{l_m}^{h_m} P^{\gamma}(\alpha^{-1} f) df} \tag{3}$$

話者正規化の有効性を示すため、図 3 (a) に SSC (話者正規化なし) の分布、(b) に話者正規化 SSC の分布を示す。これらの楕円で示した分布は、音響モデル学習に用いる 230 話者から計算を行った日本語 5 母音に対する、



(a) SSC (話者正規化なし)



(b) 話者正規化 SSC

図 3: 230 名の日本語 5 母音の分布 ($\mu \pm 0.5\sigma$)。(上図: 話者正規化を行わない SSC に対する分布、下図: 話者正規化 SSC に対する分布)

SSC、話者正規化 SSC の 1 次、2 次の値を $\mu \pm 0.5\sigma$ の範囲で示した分布である。

この図より、話者正規化手法を SSC の計算に取り入れることにより、/i/、/e/ 間に見られる分布の重複を低減することが可能であることがわかる。このように、分布の重複が低減可能であるため、話者正規化 SSC は、従来の SSC より認識性能を向上できる特徴量であると期待される。

4 音声認識実験

SSC、話者正規化 SSC の雑音下における有効性を調べるため、Travel Arrangement をタスクとする自由発話音声データベース [13] を用いた連続音声認識実験を行った。

表 1: 音響分析条件

サンプリング周波数	16kHz
プリエンファシス	0.98
フレーム周期	10 msec
フレーム長	20 msec (ハミング窓)
ケプストラム次数	12
フィルタバンク次数	16
SSC 次数	6

4.1 実験条件

音響モデル作成のための学習データとして、230 名 (男性 100 名、女性 130 名) が発声した自然発話音声を用いた。評価データは、学習に用いていない話者 42 名 (男性 17 名、女性 25 名) が発声した自然発話音声、約 40 分を用いた。実験に用いた特徴ベクトルは、表 1に示す音響分析条件で分析を行った以下の 4 種類を用いた。

SSC の雑音下における有効性を調べるために、MFCC 12 次元と対数パワーと各々の一次、二次回帰係数を加えた合計 39 次元 (**MFCC**)、この **MFCC** に SSC 6 次元とその一次回帰係数を併用した合計 51 次元 (**MFCC + SSC**) の特徴ベクトルを用いた。ここで、SSC はナイキスト周波数 (8000Hz) を 6 個のサブバンドに重なりなく等分割 ($M = 6$) を行い、式 (1) で $\gamma = 0.5$ として計算を行った。また、3で述べた話者正規化手法を **MFCC** と **MFCC + SSC** に使用した **SN-MFCC**、**SN-MFCC + SN-SSC** を特徴ベクトルとして用いた。各話者の周波数ワーピング係数 α はクリーンな発声に対し、Waves+ で求めたフォルマント周波数を用い、式 (2) で計算を行った。話者正規化 SSC は、SSC と同条件で計算を行った。

これらの 4 種類の特徴ベクトル (**MFCC**、**MFCC + SSC**、**SN-MFCC**、**SN-MFCC + SN-SSC**) を用い、ML-SSS[14] により分割を行った、総状態数 800、各 5 混合の HMnet (音素環境依存 HMM) を音響モデルとして用いた。また、3 状態 10 混合の HMM を無音モデルとして用いた。

音声認識システムを実環境下で使用する場合、学習に用いた音声の収録環境と認識を行う環境とのミスマッチが生じることが多い。そのため、本稿では、音響モデル学習時には雑音を付加しないクリーンな音声を用い、評価データには様々な雑音比 (SNR = 10, 15, 20, 30dB) となるように発話単位で雑音比を計算し、雑音を重畳した音声を用いた。この評価データに対し、品詞および可変長単語列の複合 N-gram [15] を言語モデルとして用い、単語グラフによるビームサーチ手法 [16] の 1 位の候補で評価を行った。認識単語語い数は約 7000 単語である。また、評価データに重畳した雑音は、電子協雑音データベース [17] から計算機室雑音 (ワークステーション) を用いた。

表 2: 認識結果 (単語認識率 (%))

feature vector	SNR [dB]			
	10	15	20	30
MFCC	19.7	44.4	63.1	72.5
MFCC+SSC	30.1	51.2	65.0	71.0
誤り改善率 (%)	13.0	12.2	5.2	-5.5

表 3: 話者正規化手法を用いた認識結果 (単語認識率 (%))

feature vector	SNR [dB]			
	10	15	20	30
SN-MFCC	20.0	47.6	66.5	73.7
SN-MFCC+SN-SSC	32.9	58.2	68.7	74.1
誤り改善率 (%)	16.1	20.3	6.5	1.5

4.2 認識結果

表 2に、**MFCC** と **MFCC + SSC** の認識実験結果を示す。この結果より、雑音による音声の劣化が著しい発声 (SNR = 10, 15, 20) に対して、SSC を従来の特徴量 **MFCC** の補助的パラメータして用いた場合、従来の特徴量の認識性能を向上することがわかる。特に、SNR = 10dB においては、従来の特徴量の誤りを 13.0% 改善することができた。しかし、雑音による劣化が少ない発声 (SNR = 30dB) に対しては、**SSC** を補助的パラメータとして用いた場合、従来の特徴量の認識性能を劣化させる結果となった。

SSC の問題と考えられる異なる音素間の重なりを低減するため、話者正規化を行った特徴ベクトル **SN-MFCC**、**SN-MFCC+SN-SSC** を用いた認識実験結果を表 3に示す。表 2に示した通り、話者正規化を行わない SSC を用いた場合、雑音による劣化が少ない (高い SNR) 発声に対して、認識性能を劣化させていた。しかし、話者正規化 SSC を補助的パラメータとして用いた場合には、そのような劣化は生じず、全ての雑音比において **SN-MFCC** の認識性能を改善していることがわかる。特に SNR = 15dB において、20.3% の誤り改善率を得ることができた。

さらに、表 2と表 3の比較において、話者正規化 SSC の誤り改善率は、SSC の誤り改善率より全雑音比に対して高い値を示した。このことより、話者正規化 SSC が、SSC より補助的パラメータとしての有効性が高いことわかる。特に、SNR = 15dB において、**MFCC + SSC** から

SN-MFCC + SN-MFCC への変更は、14.3% (51.2% から 58.2%) の誤り改善率を得ることができる。よって、話者正規化手法を SSC の計算手法に組み込むことは、雑音環境下における認識性能の向上に有効であることがわかった。

5 むすび

本稿では、雑音環境下における音声認識に有効な補助的パラメータである話者正規化スペクトル・サブバンド・セントロイド (SSC) を提案した。

自然発話を用いた音声認識実験の結果、話者正規化 SSC は、雑音環境下において、補助的パラメータとして用いた場合、雑音による認識性能の劣化を低減することが可能であることをを示した。また、話者正規化 SSC は、全雑音比において、従来の SSC に比べ高い誤り改善率を得ることができた。特に、提案した話者正規化 SSC を補助的パラメータとして用いた場合、雑音比の低い SNR = 10dB では、16.1%、SNR = 15dB では、20.3% の誤り改善率を得ることができた。

本稿の実験では、雑音下における音声を計算機により人工的に作成した。しかし、広く知られている通り雑音下においては、雑音比によってフォルマント周波数が変化する (ロンバード効果) ため、今後は、実環境下において収録した雑音音声を用いた音声認識実験を行う予定である。また、他の雑音除去手法、特にスペクトルサブトラクション法との併用によりさらに認識性能の劣化の低減が可能であると考えられる。

謝辞

研究の機会を与えて頂いた ATR 音声翻訳通信研究所 山本 誠一社長、第一研究室室長 匂坂 芳典室長に深く感謝をします。話者正規化のためにフォルマント周波数の計算を行って頂いた、第一研究室 内藤 正樹 氏ならびに日頃熱心にご討論頂く ATR 音声翻訳通信研究所の皆様に深く感謝します。また、ATR 音声翻訳通信研究所にて、研究を行う機会を与えて頂いた徳島大学工学部 北研二 助教授に感謝をします。

参考文献

[1] J. Picone. Signal modeling techniques in speech recognition. In *Proceedings of the IEEE*, Vol. 81, pp. 1215–1247, 1993.

[2] Ch. Jankowski Jr, H. Vo, and R. Lippmann. A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 4, pp. 286–293, July 1995.

[3] L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[4] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. In *IEEE Trans. Acoustics, Speech and Signal Processing*, pp. 113–120, April 1979.

[5] M. J. F. Gales and S. Young. An improved approach to the Hidden Markov model decomposition of speech and noise. In *Proc. ICASSP*, pp. 233–236, 1992.

[6] F. H. Liu, R. M. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *Proc. DARPA Workshop*, pp. 69–74, March 1993.

[7] H. Hermansky and N. Morgan. Rasta processing of speech. In *IEEE Trans. Speech and Audio Processing*, Vol. 2, pp. 578–589, October 1994.

[8] K. Paliwal, 柘植覚, Harald Singer, 深田 俊明, 北研二. 自然発話音声認識における音響分析の比較. 音講論, pp. 5–6, September 1997.

[9] K. Paliwal. Spectral subband centroids features for speech recognition. In *Proc. ICASSP*, pp. 617–620, 1998.

[10] 柘植 覚, 深田俊明, Harald Singer, K. Paliwal. スペクトルサブバンドセントロイドを用いた雑音下での 連続音声認識. 音講論, pp. 89–90, March 1998.

[11] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. ICASSP*, pp. 346–348, 1996.

[12] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proc. ICASSP*, pp. 353–356, 1996.

[13] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka. Japanese speech databases for robust speech recognition. In *Proc. ICSLP*, pp. 2199–2202, Philadelphia, 1996.

[14] M. Ostendorf and H. Singer. HMM topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, Vol. 11, No. 1, pp. 17–41, 1997.

[15] 政瀧浩和, 松永昭一, 匂坂芳典. 品詞および可変長単語列の複合 N-gram の自動生成. 信学論 (D-II), Vol. J81-D-II, No. 9, pp. 1929–1936, September 1998.

[16] 清水徹, 山本博史, 政瀧浩和, 松永昭一, 匂坂芳典. 大語い連続音声認識のための単語仮説数削減. 信学論 (D-II), Vol. J79-D-II, No. 12, pp. 2117–2124, December 1996.

[17] 板橋秀一. 騒音データベースと日本語共通音声データ DAT 版. 日本音響学会誌, Vol. 47, No. 2, pp. 951–953, 1991.