コンテンツ制作における収録音のための 1 入力音声強調
Single channel Speech Enhancement for Recoreded Audio Contents

中村一文
Kazufumi Nakamura
法政大学情報科学部ディジタルメディア学科
E-mail: kazufumi.nakamura.vv@cis.hosei.ac.jp

Abstract

Speech enhancement for post-production of recorded audio contents such as radio program, TV program, and TV/radio drama needs both to keep intelligibility and reduce background noise. A lot of speech enhancement methods such as spectral subtraction(SS) and running spectrum analysis(RSA) have been developed. Both SS and RSA, however, have trade-off problem between intelligibility and noise reduction level. This paper proposes a speech enhancement using speech/noise domain discrimination in modulation spectrum. Using probabililstic distribution of the modulation spectrum in the over 2Hz domain, the proposed method decides a speech/noise dominant. In 10dB SNR, segmental SNR was improved about 1.5 dB, but the result of MOS test was not improved than RSA.

1 まえがき

ドラマ・アニメ・ラジオ等のコンテンツ制作において収録音に対する編集作業は不可欠である．収録音に雑音が重畳した場合には，編集で雑音成分を低減するか収録をやり直すのが一般的である．しかしながら，再収録にはコストがかかる．またコンテンツ制作を新規・個人で始める人にとって，収録環境を整えることは至難である．これらの問題を解決する方法として 1 入力で行える自動音声強調システムが必要であると考えられる．1 入力音声強調は最低限の収録機材を用意するだけで良いので，コンテンツ制作を新規・個人で始める人にも制作環境を提供することができる．

コンテンツ制作においてはどの音が雑音になるかは制作しているコンテンツの内容によって変わってくる．しかし，楽曲制作を除けば，ドラマ・アニメ・ラジオの収録音の主たる音は音声である．そのためコンテンツ制作における雑音除去とは音声強調と捉えることができる．そしてコンテンツに用いるためこの音声強調によって強調された音声は聴覚上．雑音を抑えつつ音声の明瞭性を保つことが必要である．しかしながら音声編集ソフトウェアに付属しているノイズ除去機能は雑音除去性能と音声保存性能がトレードオフの関係にあるものが少なくない．また機能設定を細かく調整することで正しく雑音を除去することも可能なものもあるが調整の基準がわからない，機能を使用するたびに設定を調整し直す必要がある等，利便性に欠ける．このことからも自動音声強調システムの必要性がうかがえる．

本論文では変調スペクトル上における音声と雑音の性質を利用し，音声/雑音領域を決定する手法を提案する．定常的な雑音は変調周波数 2Hz 以上ではスペクトルが一様に分布している．一方，音声は周波数が増えるにつれスペクトルが弱くなる．よって変調周波数 2Hz 以上のスペクトル列に対する回帰直線は，雑音の傾きが 0，音声の傾きは負になる．提案手法ではこの差に着目し，しきい値を用いて音声/雑音領域を決定した．強調した音声を客観評価である SegmentalSNR の改善度と主観評価である MOS テストを用いて評価した結果，提案方法は MOS テストでは従来法 [1][2] と同等のスコアであったが，SegmentalSNR の改善度 [3] では従来手法を上回るスコアを得た．また雑音を低減する割合を調節することで雑音下の原音より聴きやすい音を作成できた．これにより 1 入力音声強調として提案方法の有効性を示す．

Supervisor: Prof. Katunobu Itou

2 収録における雑音

表 1. 音の分類と指摘数

分類	音の種類	指摘数
I	車の音	55
	電車の音	32
II	電話の音	13
	救急車のサイレン	11
III	人の声	13
	人の話し声	13
IV	鳥の声	6
	動物の声	3
V	工事現場の音	8
	コンピューターの端末の音	3

実際に音声を収録する場合を考える．無音での収録が難しい場合，収録する環境による背景雑音が入る可能性がある．文献 [4] では都市の音環境を把握するために個人の 1 日の行程履歴から調べるという方法で 82 人から聞き取りをして実環境に出てくる音の頻度を調査している．[4] では音の分類を 5 種類に大別しており，
I) 交通騒音， II) 人の注意を引きつける音，III) 人的要因の音，IV) 自然の音，V) 機械からの音
となっている．表 1 に [4] において指摘の多かった音の種類を分類毎に示す．表より車など乗り物の音が特に多く，次に人の声とある．背景音における人の声はバブルノイズに近いと考えられる．また録音入力が低いと空調や PC のファンなどの音を拾い，「サー」というノイズが知覚される．この音は白色雑音に近い音である．このことから本論文では実験で用いる雑音として白色雑音，バブルノイズ，自動車の走行音，雨音を選んだ．雨音は [4] において指摘に出てこなかったが天候に関わらず収録が行われる可能性があることから加えた．

本論文で取り扱う雑音はこのようにマイクの外から入ってきた音である．コンテンツ作品に用いる雑音除去にドルビーノイズリダクション [5] があるが，これは主に機材の特性からくる雑音の低減手法でありマイクの外から入ってきた雑音を抑制するものではない．提案する手法はラジオ番組などの音声をメインとするコンテンツ作成を補助するために音声以外を雑音とみなし変調スペクトルにおける音声の特徴を用いて外からの雑音を抑制する．

このような雑音を抑制する従来の 1 入力音声強調法にスペクトルサブトラクション法 (SS 法)[6] やランニングスペクトル

フィルタ (RSF)[7] がある．SS 法は 1 入力における雑音除去法として効果的であるが雑音推定に誤りがあると音声歪が発生してしまう．一方，RSF はランニングスペクトルの時間軌跡にフィルタリングする手法で，雑音を推定することなく定常的な雑音を除去し，音声の明瞭性も保つことができる．また，RSF における FIR 型フィルタによる遅延を防ぐためにランニングスペクトルをフーリエ変換した変調スペクトル上で直接処理を施すランニングスペクトルアナリシス (RSA)[2] がある．しかしながら，RSF も RSA も各周波数帯に変調スペクトルにバンドパスもしくはハイパスをかける手法であり，パス内にある雑音成分は残り，パス外に含まれる音声成分が除去され程度は微弱だが音声の明瞭性は劣化する．さらに 1 入力で行える音声/雑音領域判別手法としては [1][8] が提案されている．これらの手法はランニングスペクトル上で判別をしているが，音声の歪や雑音の引き残しが目立つ．提案手法について，各種雑音を付加した音声信号を用いて [1] と [2] の手法を従来手法として比較評価し，また実際にコンテンツの素材を用い，提案手法と原音の比較をした．

3 変調スペクトルに基づく音声/雑音領域判別

3.1 ランニングスペクトル

ランニングスペクトルは時間軸上に並べた短時間スペクトルであり，信号の時間-周波数特性を見ることができる．本論文では短時間フーリエ変換 (STFT) でシフトしながら短時間スペクトルを求めることで図 1 のようなランニングスペクトルを求めている．図 1 よりフレーム毎, 周波数毎にスペクトルが求められていることがわかる．このようにフレーム周期ごとに求められるスペクトルの時系列データのことをランニングスペクトルと呼ぶ．

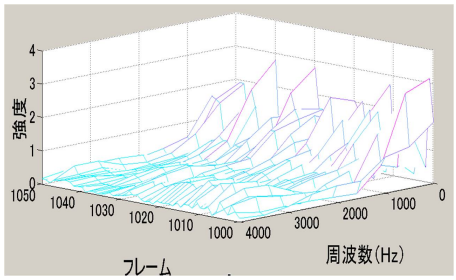

図 1. ランニングスペクトル

3.2 変調スペクトル

変調スペクトルはランニングスペクトルをフーリエ変換したものである．仮にランニングスペクトル上で 1kHz の周波数位置で時間軸に沿ってスライスする場合を考える．スライスした断面からは「1kHz におけるパワーの時間変化」を見ることができる．これをフーリエ変換すると「パワーの時間変化に対するスペクトル」を求められる．求められたスペクトルが変調スペクトルであり，変調スペクトルの横軸は変調周波数と呼ばれる．図 2 に音声と白色雑音の変調スペクトルを示す．

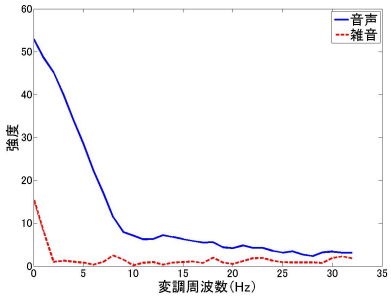

図 2. 音声と白色雑音の変調スペクトル

これらの図を比較すると白色雑音は 1Hz 以下の低域にエネルギーがほとんど集中しており，音声は高域になるほど小さくなるがエネルギーが広く分布している．白色雑音ような定常的な雑音はランニングスペクトルでの時間変化が小さいことから変調スペクトルでは低域に成分が集中している．また音声は非定常的な音であり，スペクトルの時間変化が大きいので変調スペクトルが広く分布している．特に音声認識において重要な変調周波数は 1〜16Hz であり [9]，また知覚実験により 17Hz 以下のみを用いても音声の明瞭性にはほとんど影響がない [10] とされている．よって雑音成分が集中している 1Hz 以下のエネルギーを抑えることで音声の明瞭性を保ちながら雑音を除去できるとされている．

3.3 ランニングスペクトルアナリシス (RSA)

変調スペクトル上で直接重み付けを行うのがランニングスペクトルアナリシス (RSA) である．文献 [2] では音声認識に必要な音声成分を強調するために 1〜7Hz を残すように重み付けされている．

しかし図 2 より，音声と同じく雑音も 1Hz 以上にスペクトルが分布しているのがわかる．よって雑音がある無音声区間において 1〜7Hz を残すような音声成分の強調を行うと雑音の引き残しが生じる．図 3 に変調スペクトル 1〜16Hz を強調するようにクリーン音声のスペクトログラムと RSA 処理した白色雑音を付加した音声のスペクトログラムを載せる．比較すると雑音成分が多く残っていることが見て取れる．また音声も 1〜16Hz 以外に成分があり，文献 [11] では 0Hz 付近，17Hz 以上の成分にも話者情報が含まれていることを示唆している．よって音声を歪ませず雑音を除去するためには各周波数帯における音声領域と雑音領域とを分けて適時処理する必要がある．

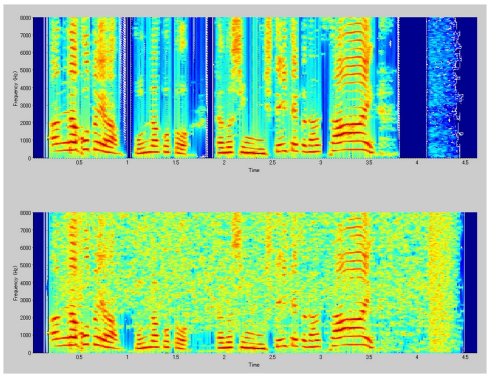

図 3. クリーン音声と RSA 処理した白色雑音付き音声

3.4 提案方法

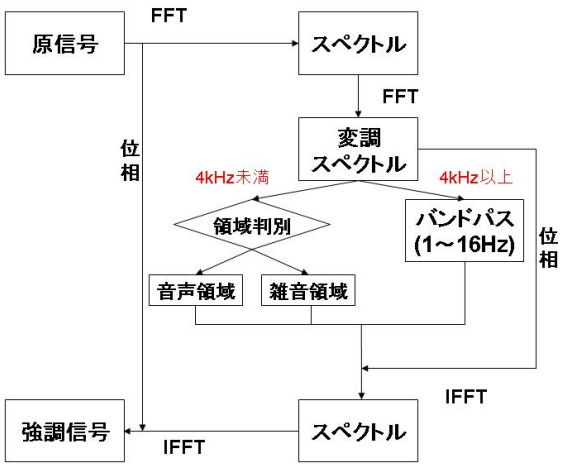

図 4. 提案方法の構成

3.4.1 音声/雑音領域判別法

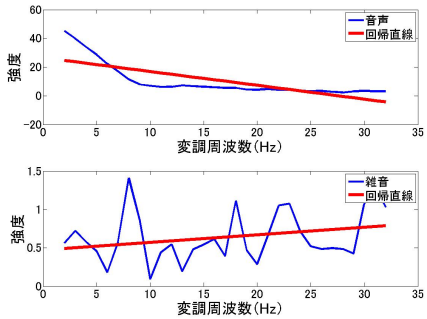

図 5. 2 H z 以上での変調スペクトル列に対する回帰直線

図 2 より，雑音成分は全体的に分布しているが 1Hz 以上ではその強さはある程度一定である．また音声成分は全体的に分布しているが高域になるにつれ緩やかに下降している．この差を利用して雑音と音声の判別を行う．具体的には 2Hz 以上におけるスペクトルの並びから回帰直線を求めて，その傾きから雑音と音声の判別を行う．先に述べた通り，雑音成分の強さはある程度一定であるので回帰直線の傾きが小さい可能性が高い．また音声成分は高域になるにつれ緩やかに下降しているので傾きが雑音成分の傾きより大きい可能性が高い．しきい値を設定して回帰直線の傾きがしきい値以下なら雑音，それより上の場合は音声と判断する．音声だと判断されれば成分をそのまま残し，雑音だと判断されれば成分を全体的に大きく減らす．

図 5 に音声の変調スペクトルから求めた回帰直線と白色雑音の変調スペクトルから求めた回帰直線をそれぞれ示す．

3.4.2 しきい値設定

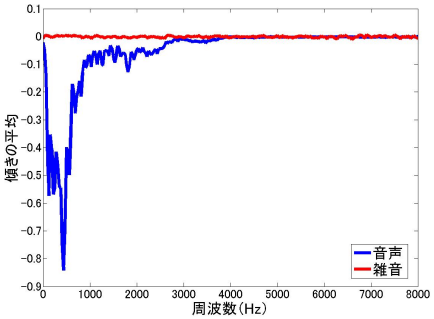

図 6. 各周波数における回帰直線の傾きの平均

変調スペクトルの強さはランニングスペクトルの周波数帯によって変わってくる．特に音声は低域に比べ高域になるとスペクトルの強さが弱くなるので雑音との回帰直線の傾きの差が小さくなる．そのため一意にしきい値を設定してしまうと高周波数において雑音成分の引き残し，または音声成分の引き過ぎが目立つ可能性がある．よって本論文ではしきい値をランニングスペクトルの周波数帯毎に設定する．本論文では以下の式でしきい値を設定し，音声/雑音領域を判別した．

$$Th(i) = \mu(i) - (\sigma(i) * d) \tag{1}$$

$$\text{領域}\ (r,i) = \begin{cases} \text{音声領域} & if \quad a(r,i) < Th(i) \\ \text{雑音領域} & if \quad a(r,i) \geq Th(i) \end{cases} \tag{2}$$

ここで i はランニングスペクトルの各周波数帯，r は変調スペクトルでのフレーム番号，a(r,i) は回帰直線の傾きである．また $\mu(i)$ は各周波数帯での傾きの平均，$\sigma(i)$ は標準偏差である．そして d は音声/雑音領域の分類を正しくおこなうためのパラメータである．　しかしながら図 6 より音声は 4kHz 以上で回帰直線の傾きがほとんど無いことがわかる．よって提案手法では 4kHz 以上の音声/雑音領域の判別が困難となる．音声を残そうとすれば雑音もほとんど残り，雑音を除去しようとすれば音声もほとんど除去されてしまう．そのため 4kHz 以上の成分については別の手法を検討することが必要となる．今回は 4kHz 未満に成分については領域判別を用い，4kHz 以上については従来法である RSA を用いることで音声の明瞭性を保持しながら雑音を除去することにする．図 4 に提案手法の構成図を示す．

4 評価実験

4.1 従来法との比較

4.1.1 実験条件

実験は 16kHz，16bit のサンプリングで録音した音声を用い，男女各 2 名づつに「あらゆる現実をすべて自分のほうへねじ曲げたのだ」という文章を読み上げてもらった．録音した音声に白色雑音・バブルノイズ・自動車の走行音・雨音をそれぞれ付加した．雑音の SNR は 10dB である．ランニングスペクトルは，短時間スペクトルを 512 点 FFT で求め，フレームシフト量を 256 点として計算した．変調スペクトルはランニングスペクトルの時間軌跡に対しフレーム幅を 4 点，フレームシフト量を 2 点として計算した．パラメータ d の値は 1 である．実験評価には Segmental SNR の改善度 [3] と MOS テストを用いて，従来手法である SS と RSA との比較をした．MOS テストの被験者は 5 名とし，音の聴きやすさを 5 段階評価した．

4.1.2 実験結果

表 2. Segmental SNR の改善度

SNR	Method	White	Babble	Rain	Car
10	Proposed	5.13	4.78	5.31	5.60
	SS	3.58	2.79	5.46	5.75
	RSA	3.57	3.06	4.42	5.69

表 3. MOS テスト

SNR	Method	White	Babble	Rain	Car
10	Proposed	2.90	3.05	3.25	2.5
	SS	2.00	2.00	2.05	1.65
	RSA	3.05	2.90	3.31	2.95

表 3，表 4 に Segmental SNR の改善度と MOS テストの結果を示す．また図 7 に白色雑音を付加した音声に対する各処理後のスペクトログラムを示す．Segmental SNR の改善度については RSA より雑音除去性能が高い，SS と比べても雑音によって変わるが常に同等かそれ以上の改善結果が出ている．これは提案手法により 4kHz 以下の雑音が従来方法より除去できているからであると考えられる．また MOS テストより車の走行音以外，提案手法は RSA と同等のスコアを得られた．

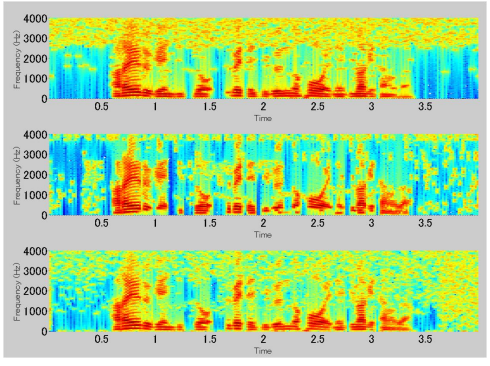

図 7. 各処理後結果 (上:提案，中：SS，下:RSA)

4.1.3 考察

車の走行音にについて，Segmental SNR の改善度ではどの手法においても大きな違いは見られなかった．これは車が目の前を通り過ぎる音が SNR10dB より高く，音声がマスキングされた結果，判別による雑音除去が行えなかったことが原因であると考えられる．しかし通り過ぎる前の音や通り過ぎた後の音は従来法より除去できている．そして従来手法では全体的にある程度消えるので，全体として SNR の改善度の結果は提案手法と従来手法では大きく変わらなかったと考えられる．しかしながら MOS テストの結果において提案手法は従来法 RSA に及ばなかった．これは音声に重畳している車の走行音が目の前を通り過ぎるときの音なので，領域判別で除去できていないことが原因である．SS より結果が良いのは，SS は音声に歪が顕著に発生しているため音声が聴き取り辛くなったためであると考えられる．

他の雑音については提案手法の低域における領域判別ができており Segmental SNR の改善度は RSA より平均 1.4dB 改善され，SS に対しても白色雑音，バブルノイズにおいてそれぞれ 1.5dB，2dB 改善されており，定常的な雑音に対する提案手法の有効性を示せた．一方，MOS テストの結果が RSA と同等なのは 4kHz 以上の雑音の残りかたが同じためであると考えられる．低域の雑音除去性能の差よりも高域での雑音が気になるため主観評価は変わらない結果となった．SS よりも提案手法と RSA の MOS テスト結果が良いのは音声の歪が SS と比べてほとんど無いためである．

4.2 原音との比較

4.2.1 実験条件

雑音が除去されても原音より聴きにくいのではコンテンツに用いることはできない．そこで原音と提案手法によって強調された音声をそれぞれ聴いてもらい，どちらの音が聴きやすいか主観で評価してもらう．被験者は 5 名である．評価に用いたデータは大学の教室で収録した音声を用いた．空調の音がノイズとして主に入っており SNR は約 15dB である．また提案手法による強調音声は雑音領域での低減する割合を 100% から 10% ずつ変えた強調音声を用意した．

4.2.2 実験結果

表 4. 原音と提案手法との比較結果

除去割合 (%)	提案手法が良い (人)	原音が良い (人)
100	0	5
90	0	5
80	2	3
70	4	1
60	5	0

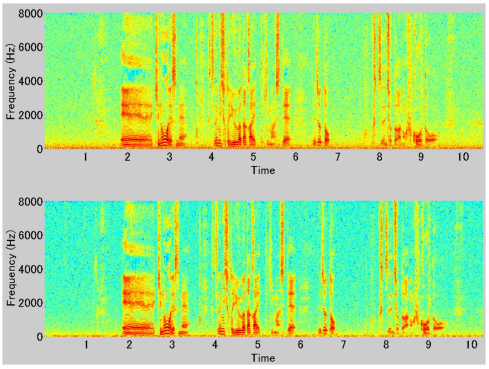

図 8. 音声のスペクトログラム (上:原音，下:処理後)

評価結果を表 5 に示す．表 5 より雑音領域をすべて除去した音声が一番聴きにくい結果となった．これは領域の誤判別によって発生した音声の歪や雑音の引き残しが音として目立ち，自然な音に聴こえなかったためだと考えられる．しかし，雑音を低減する割合を 60% にした場合，被験者全員が原音より聴きやすいと回答した．これは雑音領域の成分を残すことにより領域誤判別で生じる音声の歪が補完され自然な音声に聴こえ，引き残しによる雑音も気にならなくなったことと全体の雑音は原音より 60% 低減されたためと考えられる．図 8 に原音と雑音低減割合を 60% にしたときの強調音声のスペクトログラムを示す．図 8 より音声成分を残しながら帯域全体の雑音が薄くなっていることがわかる．この結果より提案手法を用いることで音声の明瞭性を保ちつつ雑音を抑制できることが示せた．

5 あとがき

コンテンツ制作のための音声強調として変調スペクトルを利用した音声/雑音領域判別法を提案した．提案手法では音声と雑音の判別として変調スペクトル上でのスペクトル列に対する回帰直線の傾きを利用した．提案手法は 4kHz 以上の領域判別が困難であることから 4kHz 以上については RSA を用いた．提案手法について SegmentalSNR の改善度と MOS テストによる性能評価を行った．その結果，SegmentalSNR の改善度は従来手法よりスコアが改善されたが，MOS テストのスコアは従来法と同等だった．しかし雑音領域での雑音低減の割合を調節することで音声の明瞭性を保ちつつ雑音を抑制することができ，原音より聴きやすい音を作成できた．このことよりコンテンツ素材のための音声強調としての提案手法の有効性を示すことができた．今後は低域における判別能力を向上するために適応的にパラメータ d を設定する方法と高域における処理を新たに考える必要がある．

参考文献

[1] 野村行弘他，“ 雑音量に依存しない音声/雑音領域判別法を利用した音声強調の改良 ” 日本音響学会誌 62 巻 1 号 (2006)，pp. 12-22

[2] 石田訓子他，“ ランニングスペクトルアナリシスを用いた雑音にロバストな音声認識 ” 電子情報通信技術研究報告 SIS2004-5 (2004-06)，pp. 23-28

[3] J.R. Deller Jr.，J.Hansen and J.G Proakis， Discrete-Time Processing of Speech Signals (IEEE Press, New York, 2000)，Chap.9，pp.568-597

[4] 木村英司他，“ 現代都市の具体的音環境把握のための研究 (その 1) ” 日本建築学会大会学術講演集 (1990)，pp. 895-896

[5] 山本研二，“ ドルビー・B タイプ・ノイズ・リダクション ” 日本音響学会誌 29 巻 6 号 (1973)，pp. 379-387

[6] Steven F. Boll, ”Suppression of acoustic noise in speech using spectral subtraction.” IEEE Trans.Acoust., Speech and Signal Process.， vol. ASSP-27， no. 2，(1979)，pp113-120

[7] 藤岡一馬他，“ ランニングスペクトルフィルタリングを用いた音声の雑音低減法 ” 電子情報通信学会論文誌 Vol. J88-D-Ⅱ No4 (2005)，pp. 695-703

[8] S.Yoon and C.D Yoo, et. al.“ Speech enhancement based on speech/noise-dominant decision. ” IEICE Trans. Inf. Syst.，E85-D，(2002)，pp. 744-750

[9] 金寺登他，“ 変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識 ” 電子情報通信学会論文誌 Vol.J84-D-Ⅱ No7 (2001)，pp. 1261-1269

[10] 早坂昇他，“ ランニングスペクトルフィルタを用いた雑音にロバストな音声認識 ” 電子情報通信技術研究報告 CAS2003-6，VLD2003-16，DSP2003-36 (2003-06)，pp. 31-36

[11] 金寺登他，“ 音声の変調スペクトル中に含まれる情報の調査-音声認識情報と話者識別情報との比較- ” 電子情報通信技術研究報告 SP2000-34 (2000), pp. 15-22