Internal Use Only（非公開）

TR－SLT－0027

文クラスタリングによる複数言語モデルを用いた誤認識文の推定

Mis-recognized Utterance Detection Using Multiple Language Models Generated by Clustered Sentence

藤永 勝久
Katsuhisa FUJINAGA
小窪 浩明
Hiroaki KOKUBO
山本 博史
Hirofumi YAMAMOTO
菊井 玄一郎
Genichiro KIKUI

2002年10月31日

本稿では，音声認識結果の発話単位の正解判定法について提案する．本手法は，複数のシステムが同じ単語を出力している部分は正解である可能性が高いというROVER法の考え方に基づき，複数の認識システムの信任投票により正解を判定するものである． ROVER法には（1） 認識システムを複数用意することが困難，（2） 計算コストがシステム数に応じて増加，という問題点がある．本稿では最初の問題に対しては，コーパスの自動クラスタリングにより任意の数の言語モデルを生成し，2番目の問題に対しては，リスコアリングを用いる．本手法に対し，大語彙連続認識結果の正解判定による評価を行った．その結果，正解判定を行わない場合と比較して，認識結果に含まれる正解文を10%捨てることで18ポイント，20%捨てることで24ポイント高い適合率が得られた．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

目次

1 はじめに **1**

2 従来法による認識結果の正解判定 **2**

2.1 従来法の文正解判定への応用 2

2.2 従来法における問題点 2

3 提案手法 **3**

3.1 コーパスのクラスタリングによる言語モデルの生成 3

3.2 複数の言語モデルを用いた発話単位の正解判定 3

4 提案法の評価: 実験条件 **5**

4.1 実験条件 5

4.2 評価尺度 6

5 提案法の評価: 実験結果 **7**

5.1 提案法の正解判定性能 7

5.2 従来法との比較 7

5.3 クラスタ数の違いによる比較 9

5.4 線形補完係数の違いによる比較 9

5.5 クラスタリング手法の違いによる比較 11

5.6 正解判定を行う認識結果の選択基準の比較 11

5.7 提案法により得られる単語正解精度 11

6 実装方法の改良による高速化 **15**

7 まとめ **17**

参考文献 **18**

付録 A 提案法と従来の信頼度尺度の併用 **19**

付録 B 提案法による単語単位の正解判定 **20**

B.1 単語単位での信任投票 20

B.2 評価実験 20

1 はじめに

近年，音声認識に対して統計的モデルの研究が進んできており、条件によっては認識結果に対してかなり高い精度が期待できるようになってきている．しかしながら，統計的モデルの性質ゆえに原理的にも，また実際上も誤認識が生ずることは避けがたい．そのため，音声認識を用いたシステムにおいては，あらかじめ誤認識が生ずることを前提とした対策を用意しておくことが必要となってくる．特に，音声翻訳の入力部分として音声認識を用いる場合では，たった一語の認識誤りが翻訳に対して大きな影響を与え，結果として全く異なる文に翻訳されてしまう可能性があり，誤認識への対策はきわめて重要な問題となってくる．

誤認識への対策のためには，まず誤認識個所の推定が必要となってくる．上記の音声翻訳等への応用を考えた場合，どの単語が誤っているかではなく，認識結果に誤りが含まれるかどうかが重要と考えられる．そこで本稿では誤認識単語の推定ではなく，得られた認識結果が正解文であるかどうかの推定を行うことを目的とする．

2 従来法による認識結果の正解判定

2.1 従来法の文正解判定への応用

誤認識単語の推定の方法として複数の認識システムの出力結果の共通部分を用いる方法が提案されている [1]. この方法は，複数のシステムが同じ単語を出力している部分は正解である可能性が高いというROVER法 [2] の考え方に基づくものであり，その単語を認識結果のうち信頼性の高い部分として出力するものである．本稿では，これを認識結果全体の信頼性を得るための方法に応用することとする．すなわち，複数のシステムが全く同じ認識結果を出力しているならば正しい認識結果である可能性が高いとするものである．

2.2 従来法における問題点

ROVER法における最大の問題点は複数の認識システムを用意しておき、それらを並列に実行させなければならない点にある。用いる認識システムの数は多いほど効果が見込まれるが，性格の近いシステム同士の組み合わせや，他に比べて性能の低いシステムとの組み合わせでは効果が薄いことから，適切な認識システムの組み合わせをそろえることはかなり難しくなり，認識システムの構築にかかるコストも増大する．また，複数の認識システムを並列に実行させなければならないため，実行に必要な計算コストもまたシステム数に応じて増加する点も大きな問題となってくる．

3 提案手法

我々は，2.1節で挙げたROVER法の考えに基づき，言語モデルの違う複数の認識システムの認識結果の共通部分を用いた発話単位での正解判定法を提案する．提案法では，基準となる認識システム及び正解判定に用いる複数の認識システムにより音声認識を行い，正解判定用の認識システムの認識結果の多くが基準となる認識システムの認識結果と同じであれば，基準の認識結果を正解と判定する．複数のシステムを用いる場合には2.2節で挙げたような問題が生じるが，認識システムの構築の問題に対しては，コーパスの自動クラスタリングにより任意の数の言語モデルを生成し，計算コストの問題に対しては，リスコアリングを用いることで解決する．

3.1 コーパスのクラスタリングによる言語モデルの生成

提案法では，与えられたコーパスの全てのデータを用いて学習したベースライン言語モデル及び，コーパスの自動クラスタリングを行い各クラスタ毎に学習したクラスタ言語モデルを用いる．本稿では，クラスタリング手法としてエントロピーの総和を最小とするクラスタリングを用いた[3]．

ベースライン言語モデルは学習データが最も多いため，これを用いた認識システムは，各クラスタ言語モデルを用いたものより高い音声認識性能が期待される．このため，基準となる認識システムにはベースライン言語モデルを用い，正解判定に用いる認識システムには各クラスタ言語モデルを用いる．

クラスタ数の増加に従い個々のクラスタの学習データ量が減少し，クラスタ言語モデルの性能が低下する可能性がある．そのため，ベースライン言語モデルとクラスタ言語モデルを以下のように線形補完して用いる．

$$\hat{P}(W_i|\cdot)_{C_n} = (1-\lambda)P(W_i|\cdot)_{base} + \lambda P(W_i|\cdot)_{C_n} \quad (1)$$

ここで$P(W_i|\cdot)_{base}$はベースライン言語モデルのN-gram確率，$P(W_i|\cdot)_{C_n}$はクラスタ言語モデルのN-gram確率，λは線形補完係数である．

提案法では用いるコーパスは1つであるため，コーパスを収集するコストが低い．また，自動クラスタリングにより，容易に任意の数の言語モデルを生成することが可能である．

3.2 複数の言語モデルを用いた発話単位の正解判定

提案法を用いた正解判定システムの概要を図1に示す．正解判定システムは2パスデコーダと投票による正解判定器によって構成される．はじめに，1パスの処理において，ベースラインの2-gram言語モデルを用いて入力音声の認識を行い，単語ラティスを得る．次に，2パスの処理として，ベースラインと各クラスタそれぞれにおいて，3-gram言語モデルを用

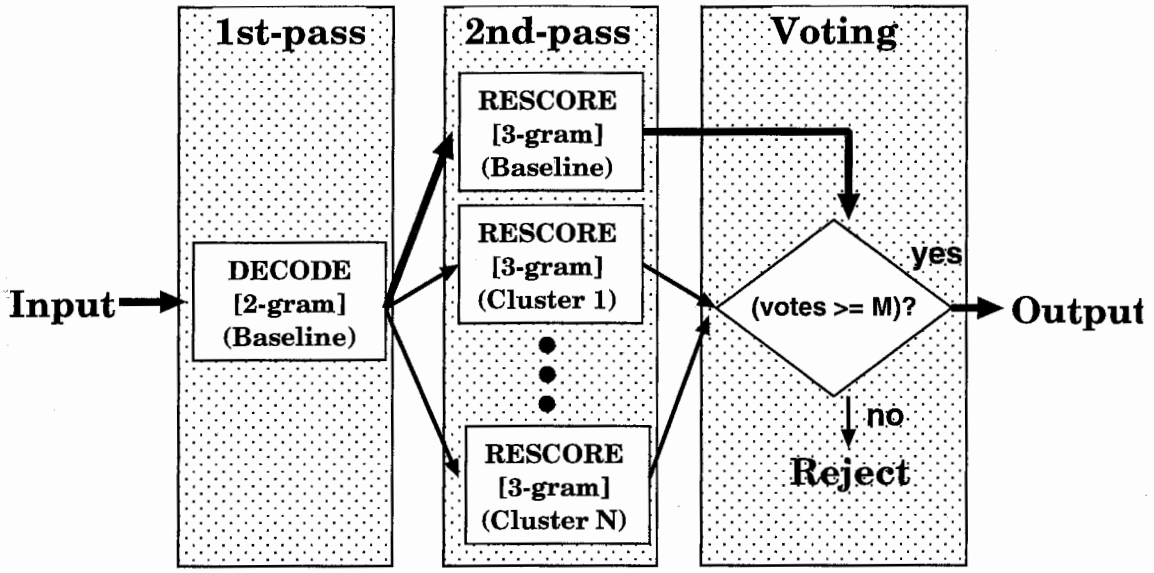

図 1: 認識システムの概要

いて単語ラティスのリスコアリングを行い認識結果の 1-best を得る．最後に，ベースラインの認識結果に対し，各クラスタの認識結果を用いて信任投票を行う．本報告では，各クラスタの認識結果がベースラインと同じ場合に 1 票投じた．得票数が閾値（M）より多い場合は正解と判定して受理し，少ない場合は誤りとして棄却する．

ROVER 法では各システムが独立して認識を行うため，システム数に比例して計算コストが増加する．これに対し，提案法では計算コストの最もかかる 1 パスの計算を 1 度しか行わないため，計算コストを大幅に削減できる．

文数	171,894
平均文長	6.9
総単語数	1,183,175
語彙数	20,561

表 1: コーパスの統計値

標本化周波数		16KHz
分析フレーム長		25ms
分析フレーム周期		10ms
特徴量		12 次 MFCC+12 次 ΔMFCC+Δpower
音響モデル		性別依存，1400 状態，5 混合 HMnet
言語モデル	1 パス	単語 2-gram
	2 パス	単語 3-gram
デコーダ		ATRSPREC[4]
単語辞書サイズ		36,810

表 2: 実験条件

4 提案法の評価: 実験条件

大語彙連続音声認識結果に対する正解判定により提案法の評価を行った.

4.1 実験条件

我々は，旅行会話の自動翻訳のための音声認識部の開発を進めている．したがって，我々の音声認識のタスクは旅行会話である．そのため，評価に用いるコーパスとして，旅行者向けのフレーズブックに現れるような旅行会話の基本表現を大量に集めたものを用いた．表1に使用したコーパスの統計値をまとめる．各発話には，それがどのような状況で用いられるかにより,「ビジネス」「留学」など 10 個のトピックの 1 つが振られている．評価実験においては，161,744 文を用いてモデルの学習を行い，510 文を用いて評価を行った．その他の実験条件は表 2 に示す.

学習データを全て用いて学習したベースライン言語モデルを使った場合に得られた認識率は，発話正解精度（utterance accuracy）が 68.63%，単語正解精度（word accuracy）が 88.89%であった.

4.2 評価尺度

一般に，正解判定では，入力の正解，不正解をどれだけの精度で認識できるかを評価する．しかし，音声翻訳を目的とする正解判定の場合，正解発話をどれだけの精度で受理したかが重要になるため，評価尺度として次式で計算した適合率と再現率を用いて評価を行う．

$$\text{適合率 (Precision)} = 100 \times \frac{\text{受理された正解発話数}}{\text{受理された発話数}}$$

$$\text{再現率 (Recall)} = 100 \times \frac{\text{受理された正解発話数}}{\text{ベースラインの正解発話数}}$$

正解判定を行わず全ての認識結果を受理した場合，適合率はベースラインの発話正解精度と等しく，再現率は100%である．

音声翻訳を目的とした場合，適合率が高いことが最も要求されるが，再現率が低いと受理されない発話が増え音声翻訳システムが成立しない．そのため，本稿では，一定以上の再現率においてどれだけの適合率を得るかを重視して評価を行う．

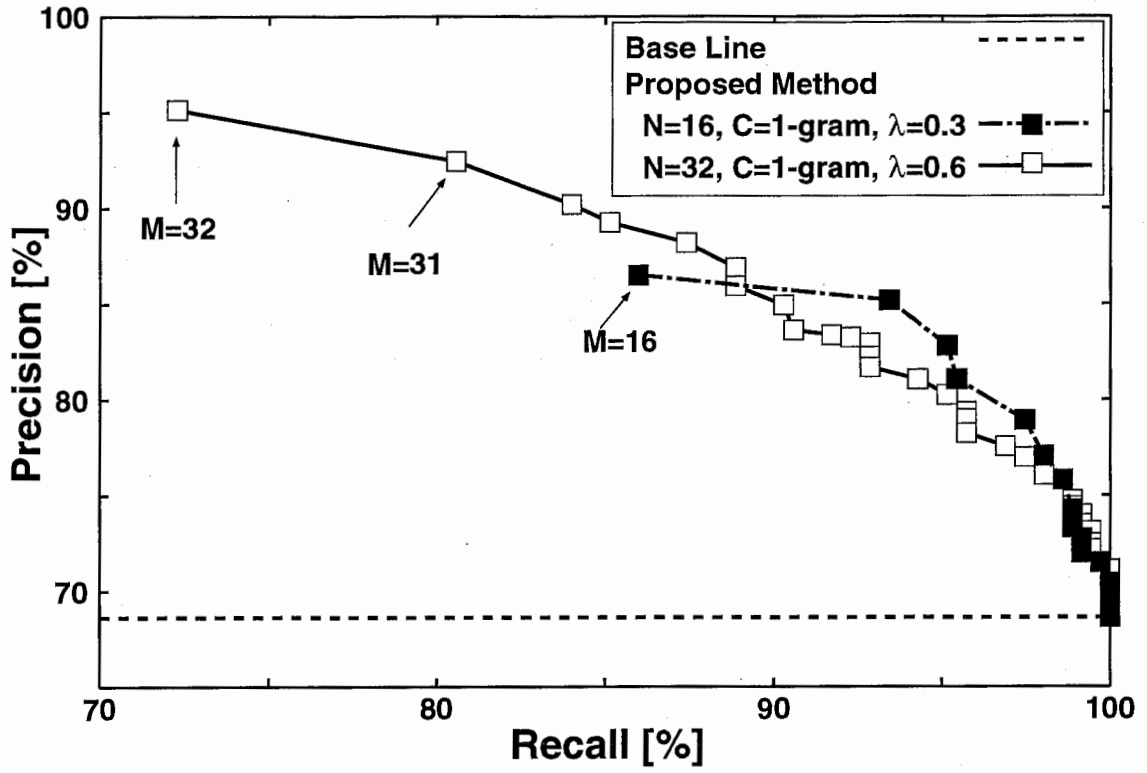

図 2: 提案法の正解判定性能

5 提案法の評価: 実験結果

5.1 提案法の正解判定性能

本節では提案法の正解判定性能の評価を行う．再現率が 80%，90%での適合率が最も高く得られたクラスタ数（N）=32，エントロピー法でクラスタリングを行う際に用いるエントロピーの計算に用いる N-gram（C）=1-gram，言語モデルの線形補完係数（λ）=0.6 の場合と，N=16，C=1-gram，λ=0.3 の場合の実験結果を図 2 に示す．図中の各点は信任投票の閾値（M）毎の実験結果を表す．

提案法は閾値の増加に従い，再現率が低下し適合率が増加した．ベースラインに対し，90%程度の再現率で 18 ポイント程度，80%程度の再現率で適合率が 24 ポイント程度向上した．

提案法は，λ などの条件の違いにより，適合率の最大値や同じ適合率に対する再現率などが大きく異なる傾向が確認できた．このため，必要とする適合率，再現率に応じて，条件を決定する必要があることがわかる．

5.2 従来法との比較

本節では提案法を従来法と比較する．従来法として，正規化尤度を用いた方法，正規化尤度，事後確率，認識結果の第一候補と第二候補の正規化尤度の比を SVM で統合した方法，

	音響モデル	言語モデル		認識率 [%]	
		1パス	2パス	発話正解精度	単語正解精度
SYSTEM 1	1400状態, 5混合, 性別依存 HMnet	単語 2-gram	単語 3-gram	68.63	88.89
SYSTEM 2	1400状態, 5混合, 性別依存 HMnet	多重クラス 複合 2-gram[5]	–	65.69	88.55
SYSTEM 3	2000状態, 16混合, 性別非依存 HMM	単語 2-gram	単語 3-gram	67.25	88.89

表 3: ROVER法で用いた認識システム

ROVER法の3つを用いた．正規化尤度を用いた方法では，各発話のフレーム単位の正規化尤度が閾値以上の場合を正解と判定して受理した．SVMで統合した方法では，学習データや評価データに含まれない1010発話を用いて学習したSVMを用いて正解判定を行った．ROVER法では3つの認識システムを用いて正解判定を行った．本来，ROVER法は音声認識精度を向上する手法であり，正解判定法ではないが，本稿では発話単位で多数決を行い，選ばれた認識候補の得票数が閾値以上の時を正解として受理した．ROVER法に用いた認識システムの概要と個々の認識結果を表3に示す．SYSTEM 1は提案法のベースライン言語モデルを用いた認識システムと等しい．また，SYSTEM 3の音響モデルは情報処理振興事業協会（IPA）の「日本語ディクテーション基本ソフトウェア」[6]収録されているものを用いた．他の条件に関しては4.1節と共通である．

提案法の5.1節と同じ条件での結果を他の手法と共に図3に示す．ROVER法において閾値が1の時に再現率が100%を越えた．これは，多数決の結果，ベースラインの認識結果より多くの正解発話が選択されたためである．

はじめに，提案法及びROVER法を他の手法と比較する．提案法及びROVER法は，正規化尤度を用いた方法より高い性能を示た．また，複数の信頼度尺度をSVMで統合したものと比較すると，同じ再現率において同等以上の適合率を得た．加えて，提案法及びROVER法では閾値を変更することにより再現率を落として適合率を向上させるなどの調整が可能だが，SVMを用いた方法では困難である．これらより，複数の認識システムの出力結果の共通部分を用いる方法は他の手法と比べて発話単位の正解判定に有効であると考えられる．

次に，提案法をROVER法と比較する．実験の結果，95%以下の再現率においては提案法が高い適合率を得た．また，提案法はROVER法より多くのシステムを用いることが可能なため，閾値による再現率，適合率の調整が容易である．更に，提案法は計算コストが低い．これらより，正解文を5%程捨てることを許容するならば，ROVER法と比較して提案法が有効であると考えられる．

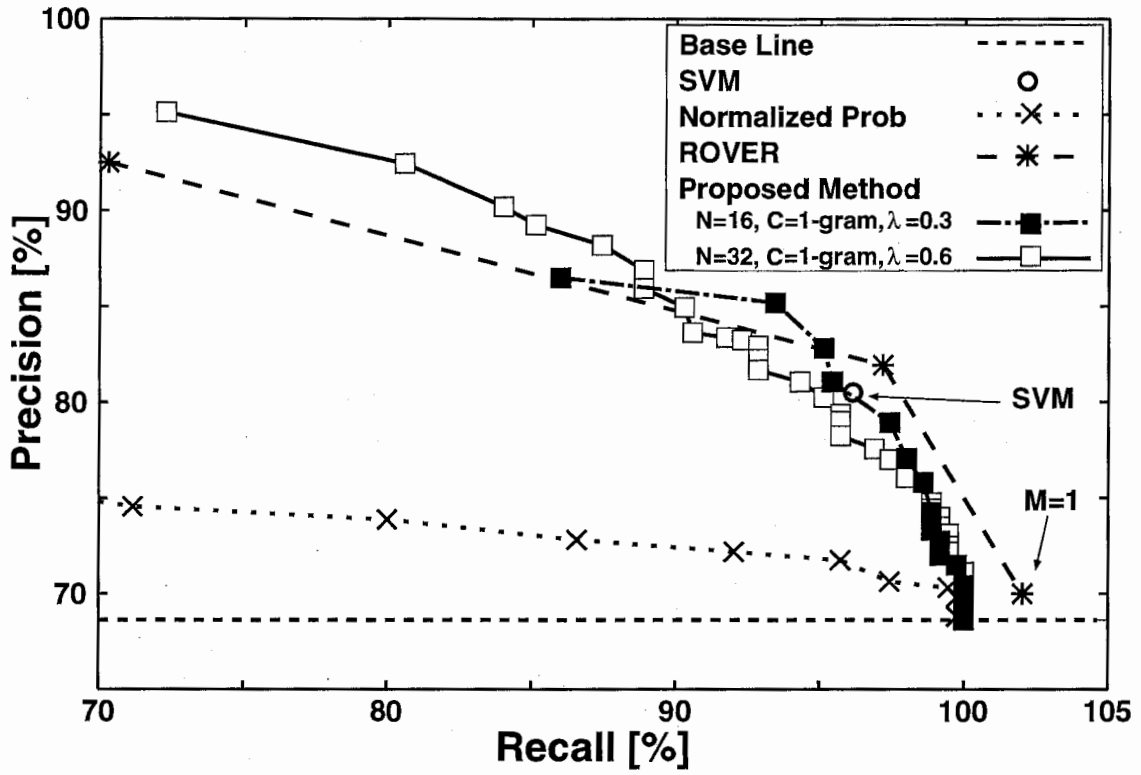

図 3: 提案法と従来法の比較

5.3 クラスタ数の違いによる比較

本節ではクラスタ数の違いによる比較を行う．C=1-gram，λ=0.6 の場合の結果を図 4 に示す．

提案法では，クラスタ数の増加に応じて適合率の最大値が向上した．しかし，90%以上の再現率における適合率は，あまり変わらない傾向がみられた．また，クラスタ数の増加に応じて閾値による適合率，再現率の変化が小さくなり，閾値による調整が容易になる傾向がみられた．

5.4 線形補完係数の違いによる比較

本節では提案法の線形補完係数の違いによる比較を行う．N=10，C=1-gram の場合の結果を図 5 に示す．

提案法では線形補完係数の増加に応じて適合率の最大値が向上し，同じ適合率を得る再現率は低下する傾向がみられた．その結果，再現率ごとに最大の適合率を得る線形補完係数が異なる傾向がみられた．また，線形補完を行わない場合（$\lambda = 1.0$）は大きく性能が低下した．これらより，各クラスタの言語モデルをベースラインのモデルと線形補完することは有効であり，線形補完係数は必要とする再現率に応じて決定することで高い適合率が得られることがわかる．

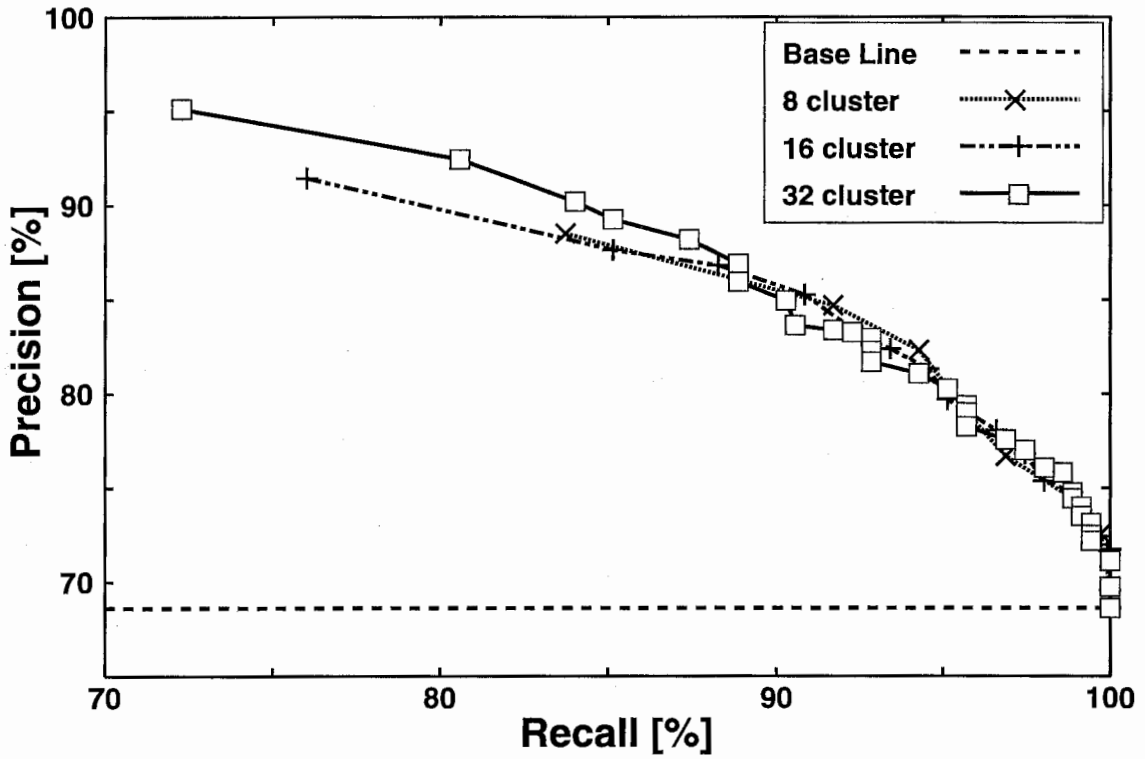

図 4: クラスタ数による比較: $C = 1$-gram, $\lambda = 0.6$

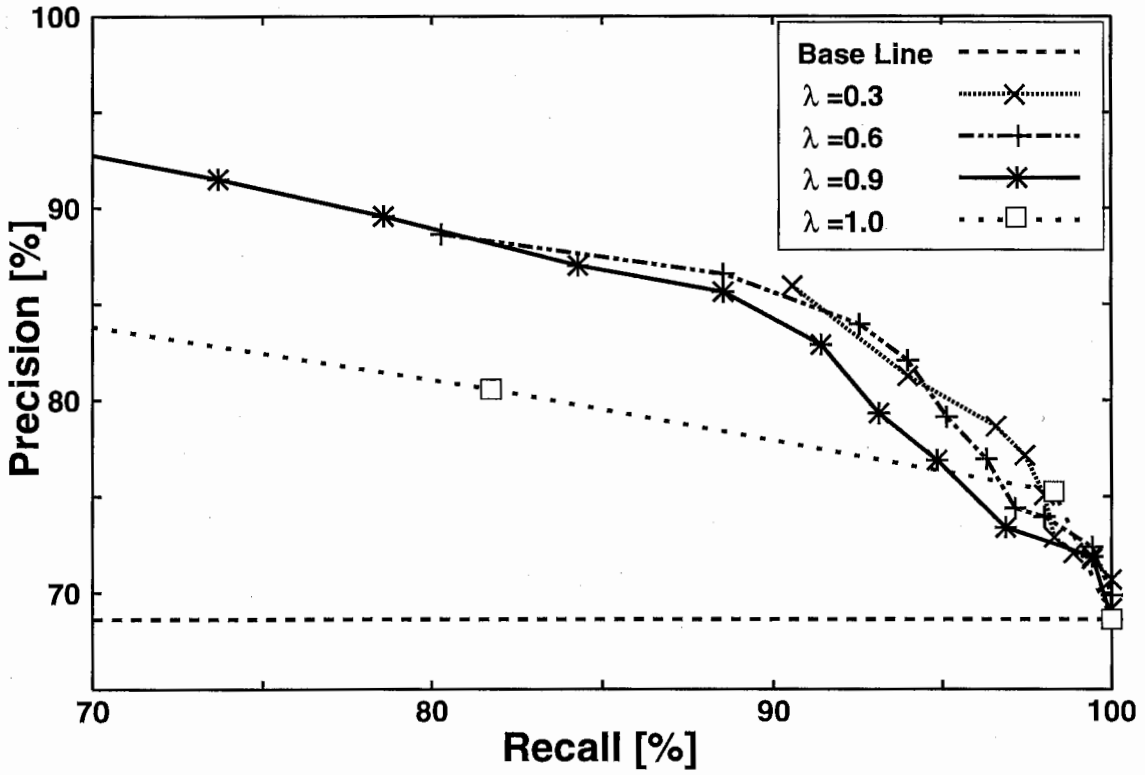

図 5: 線形補完係数による比較: $N = 10$, $C = 1$-gram

5.5 クラスタリング手法の違いによる比較

本節では提案法のクラスタリング手法の違いによる比較を行う．比較したクラスタリング手法は，エントロピー法（1,2,3-gram）とランダム，コーパスに記述されたトピックによるクラスタリングである．N=10，λ=0.9の場合での実験結果を図6，7に示す．

エントロピー法の計算に使うN-gramを比較すると，1-gramが最も性能が高く，3-gramが最も性能が低い傾向がみられた．エントロピー法（1-gram）と他のクラスタリングを比較すると，エントロピー法が最も性能が高く，ランダムが最も低い傾向がみられた．しかし，ランダムの場合でも90%程度の再現率で適合率が12ポイント程度向上する結果が得られた．これより，クラスタリング手法に関わらず，複数の言語モデルを用いることにより高い精度で正解判定を行うことができることがわかる．

5.6 正解判定を行う認識結果の選択基準の比較

提案法では，ベースラインの認識結果に対し各クラスタの認識結果による信任投票を行い正解判定を行う．これに対し，ROVER法と同様にベースラインと各クラスタの認識結果の多数決を行い，選択された認識結果の得票数が閾値以上の時を正解とする方法も考えられる．すなわち，提案法では正解判定を行う対象はベースラインの認識結果であるのに対し，多数決により正解判定を行う対象を決定する方法である．この方法は，ベースラインで誤認識をした発話を各クラスタで正確に認識することにより，提案法より多くの正解発話を得ることができる可能性がある．ベースラインと各クラスタの認識結果が全て同じもののみを受理する場合では，提案法と多数決法の受理する認識結果は同じとなる．

本節では，提案法と多数決法を比較した．N=10，C=1-gramの場合のλ=0.2，0.9の結果を図8，9に示す．λが小さい場合は提案法と多数決法に有意な差は見られないが，λの増加に従い，多数決法の性能が低下する傾向がみられた．これは，各クラスタの線形補完前の言語モデルの性能がベースラインのモデルより低いためクラスタ言語モデルの重みが増加するにつれて誤認識が増え，多数決の結果，誤りを選択してしまったためだと考えられる．この結果より，ベースラインの認識結果に対する信任投票が有効であることがわかる．

5.7 提案法により得られる単語正解精度

以上の実験では受理された発話の発話正解精度を表す適合率で評価を行った．これに対し，本節では正解判定を行い受理された発話の単語正解精度を調査した．再現率80%，90%で高い適合率を得た5.1節と同じ条件においての提案法の実験結果を図10に示す．再現率は他の実験と同じく発話単位で求めたものである．

提案法では，90%前後の再現率で7ポイント程度，80%前後の再現率で9ポイント程度単語正解精度が向上した．

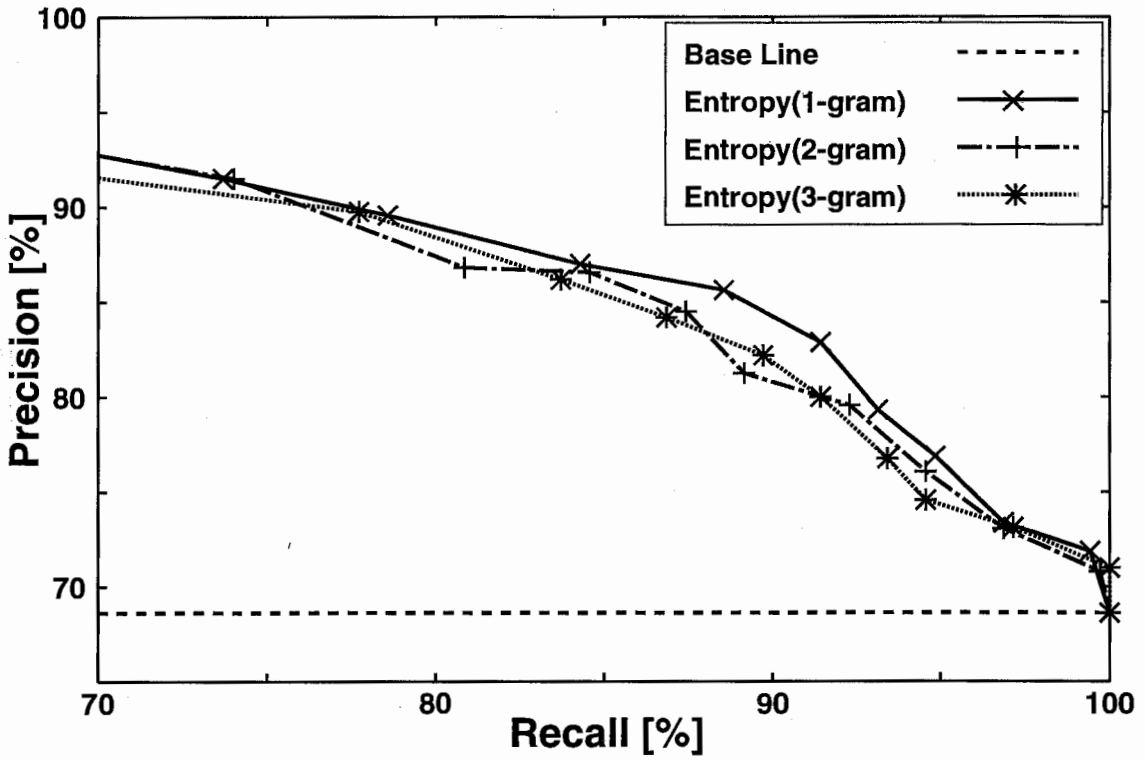

図 6: クラスタリング手法の比較 (1) : $N = 10, \lambda = 0.9$

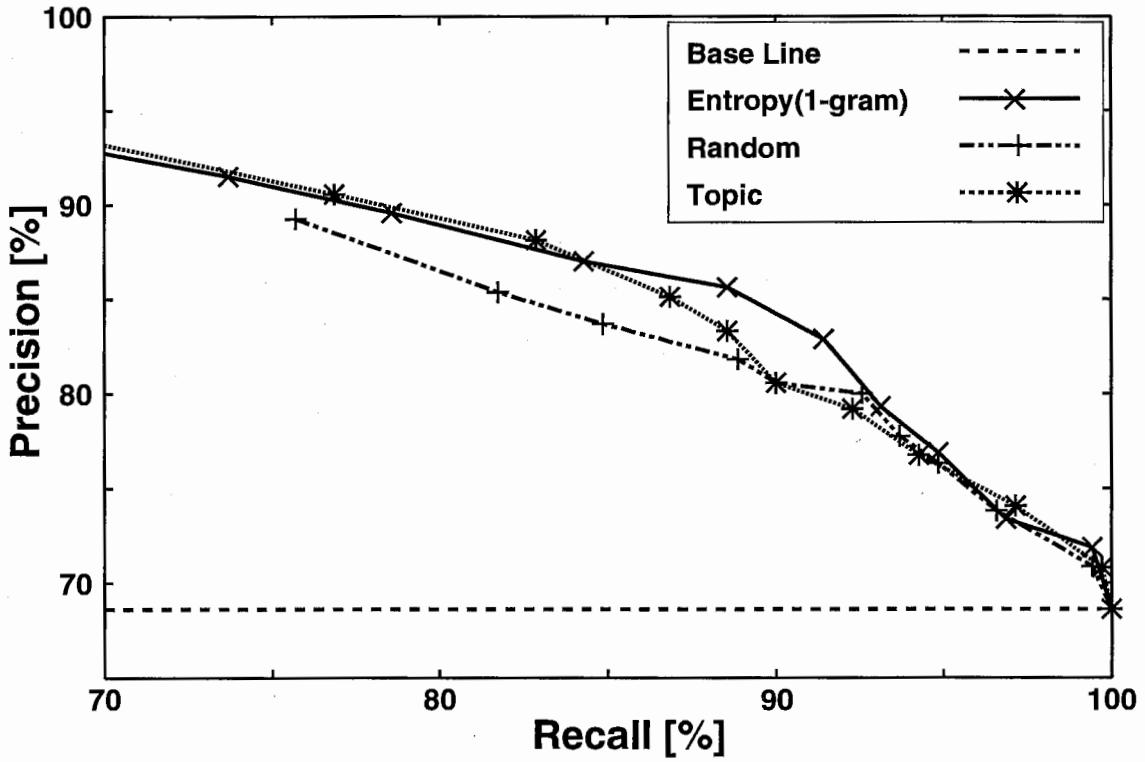

図 7: クラスタリング手法の比較 (2) : $N = 10, \lambda = 0.9$

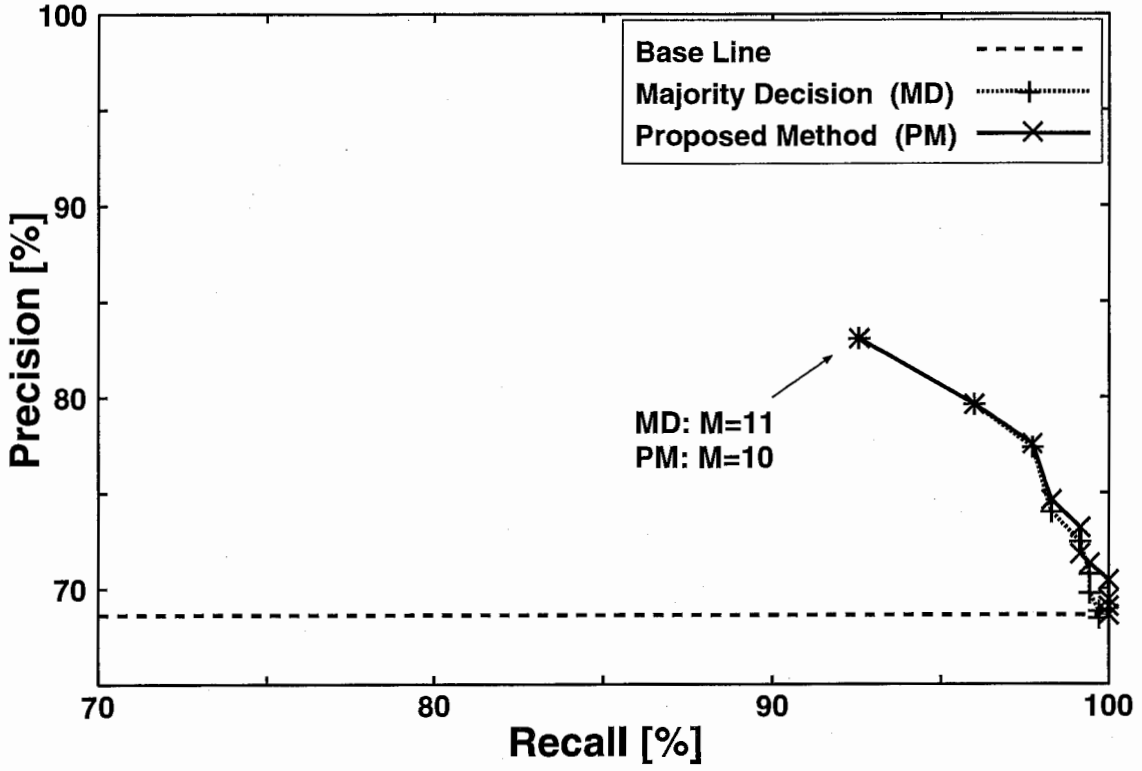

図 8: 認識結果の選択基準の比較 (1) :$N = 10$, $C = 1$-gram, $\lambda = 0.2$

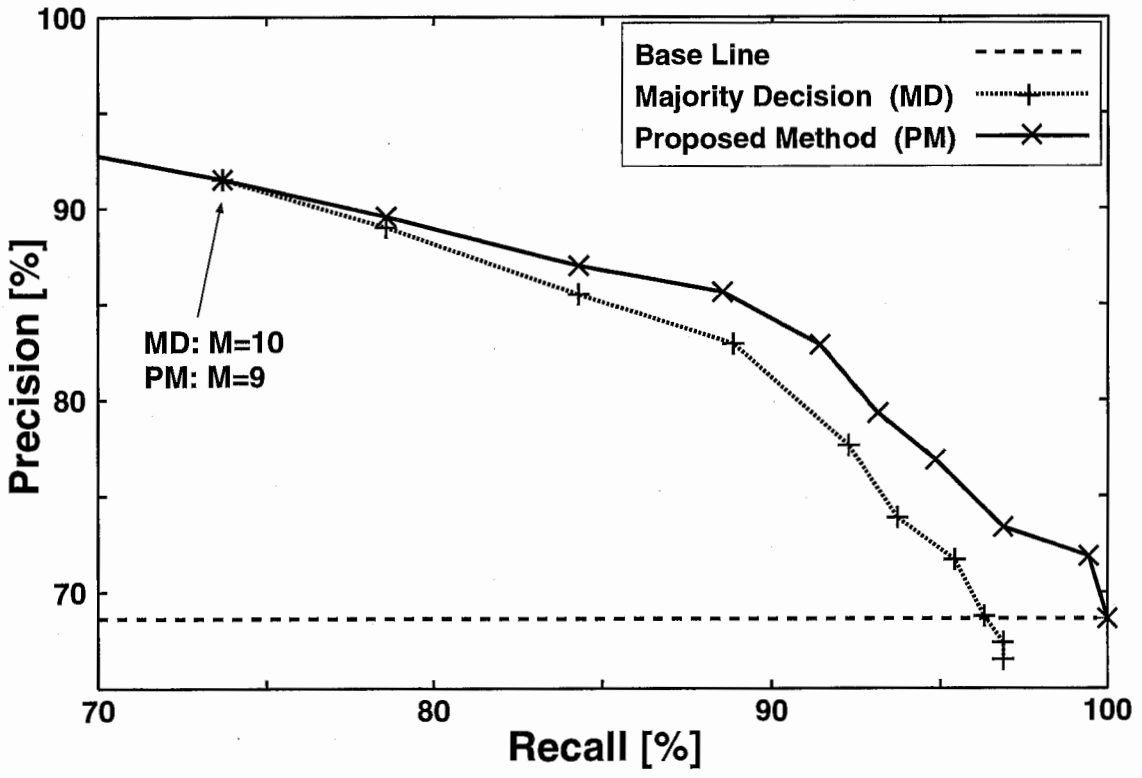

図 9: 認識結果の選択基準の比較 (1) : $N = 10$, $C = 1$-gram, $\lambda = 0.9$

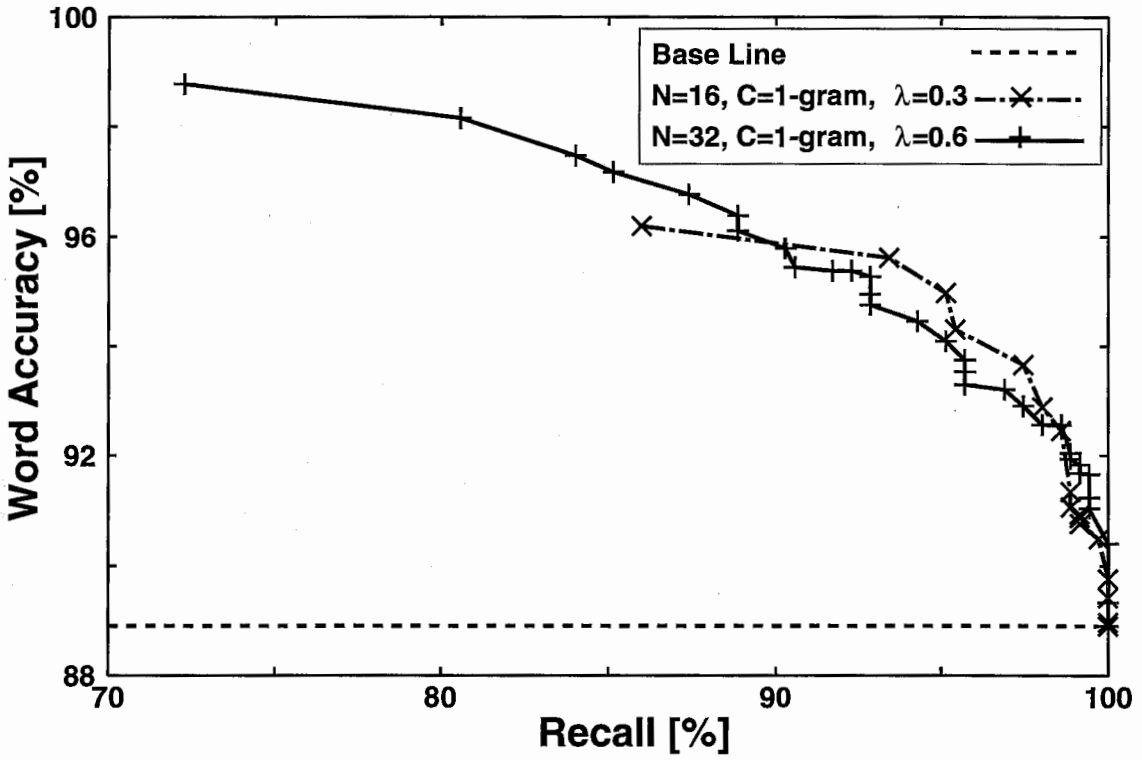

図 10: 提案法で得られる単語正解精度

6 実装方法の改良による高速化

本章では，提案法の更なる高速化について説明する．提案法はROVER法と比べ計算コストが低いが，実装方法を改良することにより更に高速な処理が可能となる．改良が可能な点は以下の2つである．

1. 言語モデル読み込みの高速化
   ベースライン言語モデルの学習データをクラスタリングして各クラスタ言語モデルを生成するため，各クラスタ言語モデルのエントリはベースライン言語モデルを越えることはない．そのため，ベースライン言語モデルと各クラスタ言語モデルの言語スコアを並べて記述することにより，言語モデルの読み込みは一度に行うことができる．

2. 1-best の同時探索
   単語ラティスの各ノードに各クラスタがベースラインと同じパスを選択したか否かを記録するキャッシュを用意することにより，ベースラインの 1-best を求めると同時に同じ 1-best を出力するクラスタの数も求めることができる．1-best 探索の例を図 11 に示す．1-best 探索は以下の手順で行われる．

   (a) 発声始端のキャッシュを全て 1 とする．

   (b) 発声始端以外では，ベースライン言語モデルを用いて，発声始端からあるノードに至るパスのうち，最大の尤度を持つパスを求める．また，ベースラインの最尤パスが各クラスタ言語モデルでも最尤であるかを求め，最尤なら 1，異なる場合は 0 でベースラインの最尤パスの 1 つ前のノードのキャッシュの値と AND をとり，そのノードのキャッシュに入れる．

   (c) 全てのノードの計算が終了するまで （ b ）に戻る．

   (d) 終端ノードからバックトレースを行い，ベースラインの 1-best を求める．

   (e) 終端ノードのキャッシュが 1 になるものは，ベースラインと同じ 1-best をもつ．

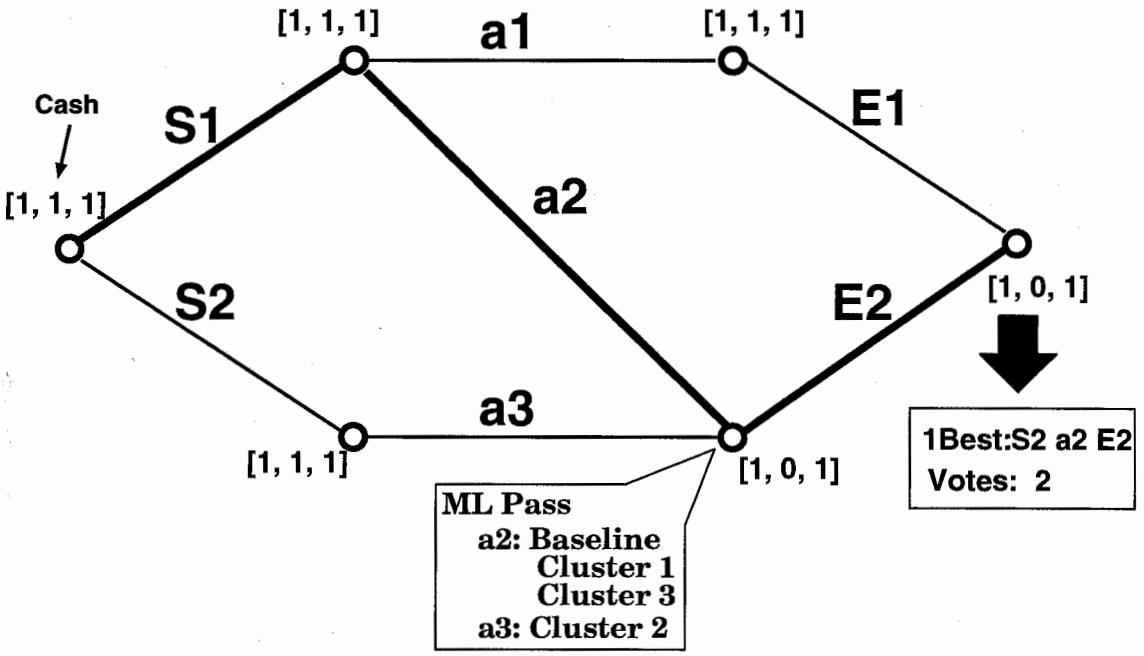

図 11: 1-best の同時探索

7 まとめ

本稿では，音声認識結果の発話単位の正解判定法として，言語モデルが違う複数のシステムの共通部分を用いた発話単位の正解判定法を提案した．本手法はROVER法に対し次のような利点がある．

- 与えられたコーパスに対し自動クラスタリングを行い，任意の数の言語モデルを自動生成することにより，容易に複数の音声認識システムを構築できる．
- 音声認識で最も計算コストの高い1パスの計算はベースライン言語モデルを用いた1度だけであるため，計算コストが非常に低い．

提案法を評価するために，大語彙連続音声認識結果に対する正解判定を行った．その結果，提案法はベースラインに対し90%程度の再現率で18ポイント程度，80%程度の再現率で24ポイント程度適合率が向上し，他の手法と比較して同等以上の性能を得た．また，信任投票の閾値により，容易に再現率，適合率の調整が行えることが確認した．次にクラスタ数や線形補完係数など条件の違いによる評価を行い，次のような結果を得た．

- クラスタ数の増加に応じて適合率の最大値が向上した．
- 線形補完係数の増加に応じて適合率の最大値が向上し，同じ適合率を得る再現率の値は低下した．
- エントロピー計算に1-gramを用いたエントロピー法によりクラスタリングを行った場合において，最も高い性能を得た．
- ベースラインの認識結果への信任投票が高い性能を得た．

最後に，実装方法の改良による高速化について論じた．

参考文献

[1] Y. Kodama, T. Utsuro, H. Nishizaki, S. Nakagawa. "Experimental Evaluation on Confidence of Agreement among Multiple Japanese LVCSR Models". Proc. EUROSPEECH 2001, pp. 2549-2552, 2001.

[2] J. G. Fiscus. "A Post-processing System to Yield Reduced Error Word Rates: Recognizer Output Voting Error Reduction (ROVER)". Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347-354, 1997.

[3] 清水 徹, 大野 晃生, 樋口 宜男. "文セットのクラスタリングに基づく統計的言語モデル". 音響学会講演論文集, 1-6-14, pp. 31-32, 1998-3.

[4] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, and Y. Sagisaka. "Spontaneous Dialogue Speech Recognition Using Cross-word context Constrained Word Graphs". Proc. ICASSP 1996, pp. 145-148, 1996.

[5] 山本 博史, 匂坂 芳典. "接続の方向性を考慮した多重クラス複合N-gram言語モデル". 電子情報通信学会論文誌, Vol.J83-D-II, No.11, pp. 2146-2151.

[6] 河原 達也, 李 晃伸, 小林 哲則, 武田 一哉, 峯松 信明, 嵯峨山 茂樹, 伊藤 克亘, 伊藤 彰則, 山本 幹雄, 山田 篤, 宇津呂 武仁, 鹿野 清宏. "日本語ディクテーション基本ソフトウェア (99年度版)". 日本音響学会誌, Vol.57, No.3, pp. 210-214, 2001.

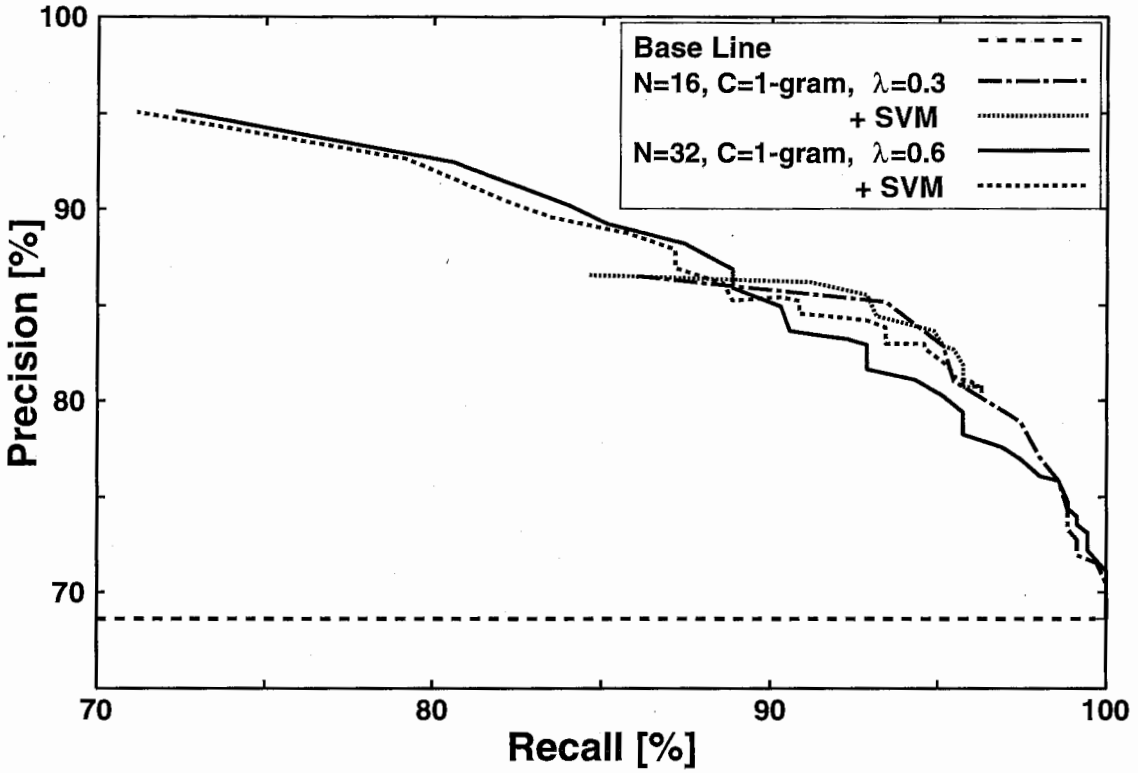

図 12: 他の手法との併用

付録 A　提案法と従来の信頼度尺度の併用

提案法と従来法を併用した実験を行った．提案法で正解と判断された認識結果に対し，従来法を用いて正解判定を行い，共に正解と判断された結果を受理した．提案法と併用した従来法は正規化尤度，事後確率，認識結果の第一候補と第二候補の正規化尤度の比を SVM で統合した手法である．これは 5.2 節で用いたものと等しい．

実験結果を図 12 に示す．従来法と併用した場合，90%以上の再現率において適合率が向上する傾向がみられた．しかし，提案法において高い再現率で高い適合率を得るように条件を設定した場合 (N=16, C=1-gram, λ=0.3)，従来法と併用しても性能は向上しなかった．

付録 B　提案法による単語単位の正解判定

本章では，提案法を単語単位での正解判定に適用する．

B.1　単語単位での信任投票

提案法を単語単位に適応するのは容易である．コーパスをクラスタリングすることにより任意の数のクラスタ言語モデルを生成し，リスコアリングによりベースライン及び各クラスタの認識結果を得るまでは，発話単位の場合と同様の処理を行う．正解判定は，ベースラインの認識結果に含まれる各単語に対し，各クラスタの認識結果に含まれる各単語による信任投票を行うことにより行う．すなわち，各クラスタの認識結果の各単語がベースラインの認識結果に含まれる場合，ベースラインの単語に対し1票を投じる．ベースラインと各クラスタの認識結果の各単語の対応付けはDPを用いて行う．また，提案法では全てのシステムにおいて同じ構造のラティスを用いるため，各クラスタの1-bestがベースラインの1-bestと同じパスを通るか判定することでも，対応付けが可能である．

B.2　評価実験

大語彙連続音声認識結果に対する正解判定により提案法の評価を行った．実験条件は4.1と等しい．また，比較のために単語事後確率を用いた正解判定とROVER法による正解判定を行った．単語事後確率を用いた正解判定では，ベースラインのシステムから得られる単語事後確率が閾値以上の時に正解とした．ROVER法では，3つの認識システムの認識結果をDPにより単語単位でアライメントをとり，多数決を行った．多数決により選択された単語の得票数が閾値以上の場合を正解とした．ROVER法に用いた認識システムの概要と個々の認識結果を表4に示す．

評価尺度は以下の式で求めた適合率，再現率を用いた．

$$\text{適合率 (Precision)} = 100 \times \frac{\text{受理された正解単語数}}{\text{受理された単語数}}$$

$$\text{再現率 (Recall)} = 100 \times \frac{\text{受理された正解単語数}}{\text{ベースラインの正解単語数}}$$

ベースライン言語モデルを用いた認識システムが出力する認識結果の単語数は3435であり，正解単語数は3133であった．このため，正解判定を行わず，全ての認識結果を受理した場合，適合率は91.21%，再現率は100%となる．

提案法と比較手法の実験結果を図13に示す．提案法では，95%の再現率で5ポイント，90%の再現率で6.5ポイント適合率が向上した．

提案法を他の手法と比較すると，正規化尤度を用いた手法より高い性能を得られ，ROVER法よりわずかに性能が劣った．しかし，提案法はROVER法より計算コストが低く，適合率，再現率の調整が容易である．

	音響モデル	言語モデル		認識率 [%]	
		1パス	2パス	発話正解精度	単語正解精度
SYSTEM 1	1400 状態, 5 混合, 性別依存 HMnet	単語 2-gram	単語 3-gram	68.63	88.89
SYSTEM 2	1400 状態, 5 混合, 性別依存 HMnet	多重クラス 複合 2-gram[5]	–	65.69	88.55
SYSTEM 3	2000 状態, 16 混合, 性別非依存 HMM	多重クラス 複合 2-gram[5]	–	60.78	86.30

表 4: ROVER 法で用いた認識システム

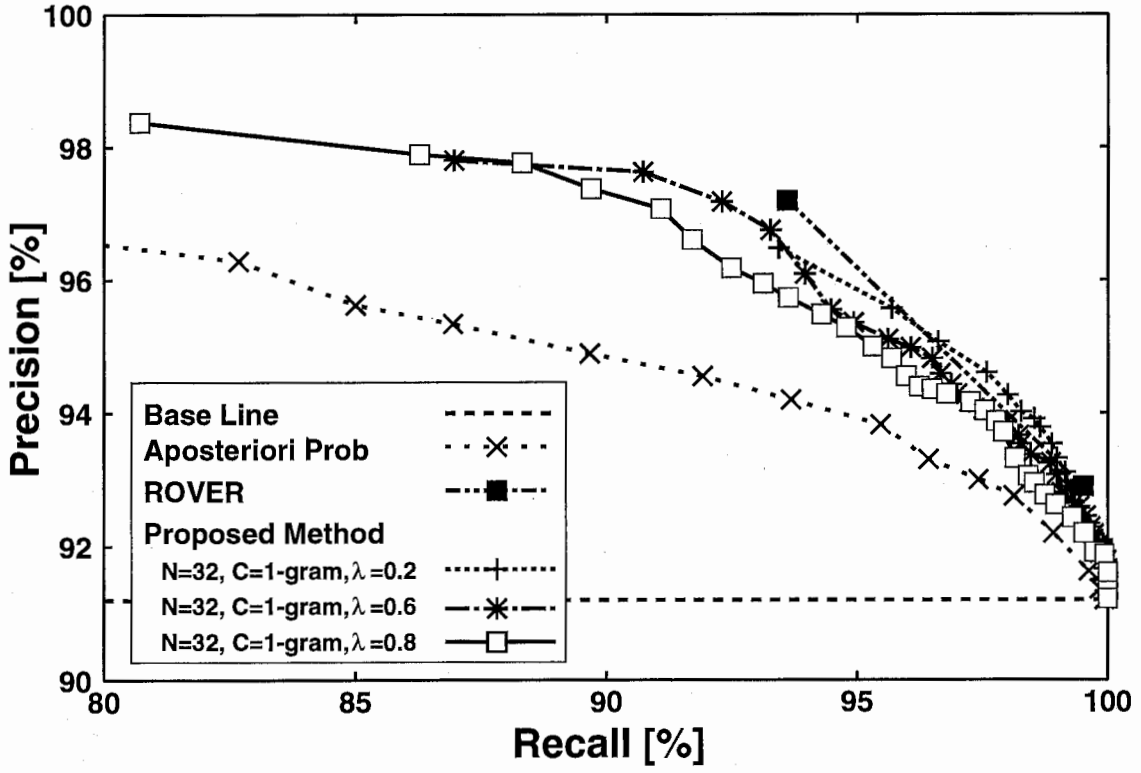

図 13: 提案法の正解単語判定性能