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

2005－ＳＬＰ－59（19）
2005／12／22

統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換

秋田 祐哉† 河原 達也†

† 京都大学 学術情報メディアセンター
〒606-8501 京都市左京区吉田二本松町

あらまし 講演や会議のような話し言葉の音声認識では，言語モデルの学習に際してタスクにマッチしたデータ，すなわち忠実な書き起こしテキストの量が限られていることが問題となっている．本稿では，大規模な文書データベースに基づく言語モデルの統計量から，話し言葉言語モデルの統計量を推定する変換手法を提案する．提案する統計的変換モデルでは，話し言葉に特徴的な言語表現がモデル化され，それらの変換確率も推定される．変換の発生する文脈パターンと確率は，音声の忠実な書き起こしとそれを文書スタイルに整形したテキストからなる小規模パラレルコーパスを用いて学習される．変換の適用範囲を広げ，信頼性を高めるために，単語を文脈としたモデルから品詞を文脈としたモデルにバックオフする枠組みを導入する．提案法を国会音声の認識タスクに適用したところ，テストセットパープレキシティを大きく改善することができた．
キーワード 話し言葉，音声認識，言語モデル，統計的機械翻訳，発話スタイル

Speaking Style Transformation of Language Model Based on Statistical Machine Translation Framework

Yuya AKITA† and Tatsuya KAWAHARA†

† Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract One of the most significant problems in language modeling of spontaneous speech such as meetings and lectures is that only limited amount of matched training data, i.e. faithful transcript for the relevant task domain, is available. In this paper, we propose a novel transformation approach to estimate language model statistics of spontaneous speech from a document-style text database, which is often available with a large scale. The proposed statistical transformation model is designed for modeling characteristic linguistic phenomena in spontaneous speech and estimating their occurrence probabilities. These contextual patterns and probabilities are derived from a small amount of parallel aligned corpus of the faithful transcripts and their document-style texts. To realize wide coverage and reliable estimation, a model based on part-of-speech (POS) is also prepared to provide a back-off scheme from a word-based model. The approach has been successfully applied to estimation of the language model for National Congress meetings from their minute archives, and significant reduction of test-set perplexity is achieved.
Key words Spontaneous speech, Speech recognition, Language model, Statistical machine translation, Speaking style

1. はじめに

近年の音声認識の対象は，講義・講演や会議・討論，会話といった自然な自発音声（話し言葉音声）に移ってきている．話し言葉音声では音響的・言語的な多様性がいちだんと大きくなることから，これらを適切にモデル化することは重要な課題である．特に言語的な面では，対象音声のドメイン（話題）に関連する表現に加えて，フィラーや口語表現などの話し言葉に特徴的な表現もカバーする言語モデルの構築が求められている．

このような言語モデルは，これら2つの言語的特徴を十分に含んだコーパスを利用して学習することが理想的であるが，現実にそのようなコーパスを得ることは難しい．ドメインをカバーするコーパスとしては，規模が大きく収集も比較的容易な，講演録や議事録などの文書データベースが考えられる．しかし，講演録や議事録のテキストでは文書化のための整形が行われており，話し言葉表現はほとんど含まれない．一方，音声を忠実に書き起こした話し言葉テキストのコーパスは，作成に多大なコストが必要となるため，ドメインごとに大量に収集することは困難である．いくつかの大規模話し言葉コーパスも存在するが，特定のドメインに限定されている．

したがって，話し言葉音声認識のための言語モデル構築手法としては，ドメインをカバーする文書データベースと，何らかの話し言葉テキストコーパスを混合・補間する手法が一般的である．例えば講義や講演の音声認識[1],[2]では，講義テキストや講演予稿などの書き言葉テキストと会話音声の書き起こし（Switchboard・Fisherコーパス）を用いて言語モデルを構築している．また会議音声の認識[3],[4]でも，会議コーパスやWebテキストコーパスとSwitchboardコーパスが組み合わされている．このような混合・補間手法は有効に機能するが，ドメインと無関係な表現まで言語モデルに含まれたり，あるいは単語の接続関係を十分にカバーできないなどの問題もあり，これらによる音声認識性能の低下が懸念される．このほか，書き言葉テキストにフィラーをランダムに挿入して擬似的な話し言葉テキストを生成し，これを用いて言語モデルを学習する手法も提案されている[5]．しかしこの手法ではフィラーしか扱われておらず，また挿入の確率も適切に推定されているとはいえない．

そこで本稿では，話し言葉表現に関する言語モデルの統計量（N-gram単語列とその出現頻度）を，大規模文書コーパスから変換モデルを用いて推定する手法を提案する．変換モデルは統計的機械翻訳で広く用いられている枠組みに基づいており，話し言葉表現への変換パターンを確率的にモデル化できるよう設計されている．この変換パターンと確率は，文書スタイルのテキストと話し言葉スタイルのテキストが対応づけられたパラレルコーパスを用いて学習される．広範囲に適用可能なモデルを少量のデータからでも構築できるように，単語の一致に基づく変換パターンに加えて，品詞の一致に基づく変換パターンもあわせて学習する．本稿では国会音声の認識タスクにおいて評価を行い，提案法がパープレキシティを大きく削減することを示す．

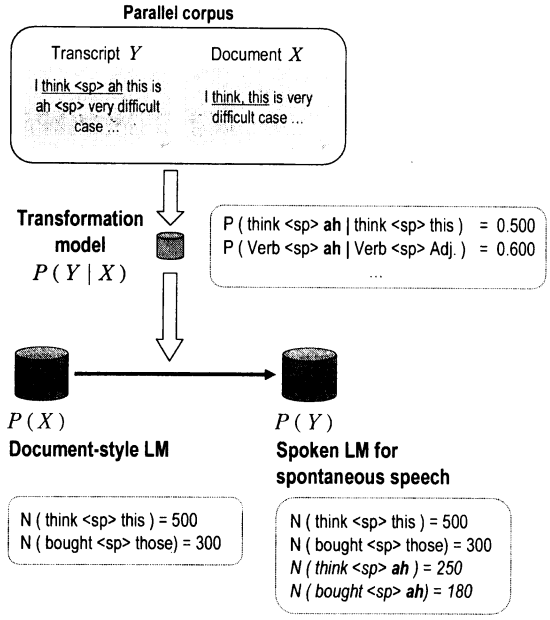

図1　統計的機械翻訳に基づく言語モデル変換の概念図
Fig. 1　Conceptual image of SMT-based transformation

2. 統計的機械翻訳の枠組みに基づく変換

統計的機械翻訳（Statistical Machine Translation, SMT）[6]では，翻訳元の言語における文 X と翻訳先の言語の文 Y について，事後確率 $P(Y|X)$ が最大となる Y を翻訳文として出力する．Bayes則に基づき $P(Y|X)$ は(1)式で与えられる．

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

ここで $P(X)$ と $P(Y)$ はそれぞれ翻訳元・翻訳先の言語の言語モデルである．$P(X|Y)$ は翻訳モデルと呼ばれ，両言語の対応関係を規定している．実際のSMTでは，$P(X)$ は Y の選択に寄与しないことから無視される．

表 1　文書スタイルと話し言葉スタイルのおもな違い
Table 1　Major differences between spontaneous speech and document-style text

種類	$P(Y\|X)$	$P(X\|Y)$	文書スタイル X	→	話し言葉スタイル Y
フィラーの挿入	要 推定	1	私 は 思い ます	→	私 は **あー** 思い ます
助詞の脱落	要 推定	要 推定	私 **は** 思い ます	→	私 思い ます
口語表現の置換	要 推定	1	さ **せ** て いただき	→	さ **し** て いただき

本研究では，文書スタイル（X）と話し言葉スタイル（Y）をそれぞれ別の言語としてとらえ，文書スタイルの言語モデル $P(X)$ から話し言葉スタイルの言語モデル $P(Y)$ を生成することを考える．図 1 にこの変換の概念図を示す．(1) 式より，$P(Y)$ は (2) 式で表される．

$$P(Y) = P(X)\frac{P(Y|X)}{P(X|Y)} \tag{2}$$

言語モデルとして N-gram モデルを用いるとすると，N-gram 単語列の出現確率は学習コーパスにおける出現頻度により定められる．したがって，文書スタイルの n-gram 単語列 x_1^n と話し言葉スタイルの単語列 y_1^n の出現頻度 $N(x_1^n)$・$N(y_1^n)$ には，(2) 式より (3) 式の関係が導かれる．

$$N(y_1^n) = N(x_1^n)\frac{P(y_1^n|x_1^n)}{P(x_1^n|y_1^n)} \tag{3}$$

(3) 式は，$N(x_1^n)$ から $N(y_1^n)$ が推定可能であることを示している．条件付き確率 $P(y_1^n|x_1^n)$・$P(x_1^n|y_1^n)$ は，音声の書き起こしと，それを文書スタイルに編集したテキストによるパラレルコーパスを用いて推定することが可能である．

3.　話し言葉表現の統計量の推定

3.1　対象とする話し言葉表現

本研究では，表 1 に示す 3 種類の話し言葉表現を主な対象とする．以下では，それぞれのケースについて話し言葉変換の確率 $P(Y|X)$・$P(X|Y)$ を検討する．

フィラーの挿入は話し言葉特有の現象であり，発話の冒頭・末尾あるいはポーズにおいて頻繁に観測される．しかし，これらの点で常にフィラーが挿入されるわけではなく，またフィラー単語の種類は前後の文脈に依存する．例えば表 1 の例では，直前の音節が「は」(/w a/) であるため，同一の母音からなるフィラー単語「あー」が挿入される場合が他と比較して多いと考えられる．したがって $P(Y|X)$ は学習データをもとに推定する必要がある．一方，文書スタイルではフィラーは必ず削除されることから，$P(X|Y)$ は常に 1 と考えられる．

助詞の脱落も話し言葉音声によくみられる現象であるが，フィラーの場合と同様に，必ずしも全ての助詞が脱落するわけではない．例えば係助詞「は」・「が」や格助詞「を」には脱落が比較的多くみられるが，格助詞「の」の脱落は少ない．逆に話し言葉から文書スタイルへの編集においても，助詞の挿入可能な点の全てで挿入が行われるわけではない．したがってこのケースでは $P(Y|X)$・$P(X|Y)$ のいずれも学習データを用いて推定する必要がある．

話し言葉音声では，書き言葉の表現が発声の怠けなどによって変化した口語表現が見られる．もとの表現や話者によって口語表現の種類や変化の割合は異なるが，実際にみられた口語表現は文書として整形される際に丁寧な書き言葉的表現に改められる．そのため，口語表現の場合もフィラーと同様に $P(Y|X)$ は学習データから推定し，$P(X|Y)$ は 1 とする．

3.2　変換確率の推定

条件付き確率 $P(Y|X)$・$P(X|Y)$ は，文書スタイルと話し言葉スタイルのテキストのパラレルコーパスを用いて学習される．両スタイル間の表現の差異（すなわち編集操作）はテキストにあらかじめタグ付けされているとする．本研究では，変換に際して前後の単語文脈を条件とし，これを含めた変換パターンと確率を学習する．これに加えて前後の単語の品詞を条件とした変換パターンと確率も求める．

まず，2 つのテキストで異なっている（タグ付けされた）表現について，文書スタイルのテキストにおける当該単語列 x_1^n の頻度 $N(x_1^n)$ と，編集により x_1^n となる，もとの話し言葉スタイルの単語列 y_1^n の頻度 $N(y_1^n|x_1^n)$ を求める．x_1^n・y_1^n には文脈となる前後の単語も含め，文脈ごとの頻度を求める．これらの頻度を用いて，単語文脈の場合（単語ベース）の条件付き確率 $P_{\mathrm{word}}(y_1^n|x_1^n)$ は (4) 式により定められる．

$$P_{\mathrm{word}}(y_1^n|x_1^n) = \frac{N(y_1^n|x_1^n)}{N(x_1^n)} \tag{4}$$

次に，文脈の単語を品詞や活用形ごとにまとめて，品詞文脈ごとの頻度を同様に求める．本研究では品詞は形態素解析器を用いて自動的に付与している．利用した解析器は茶筌 Ver.2.2.3+IPADIC-2.4.4 である [7]．文脈を

表 2　言語モデルの仕様
Table 2　Specifications of language models

モデル	ベースライン ("Baseline")	CSJ	混合モデル		変換によるモデル	
			CSJ ("+CSJ")	書き起こし ("+Transcript")	ベースラインから ("Proposed-BL")	+Transcript から ("Proposed-Trn")
学習コーパス	衆議院 会議録	日本語 話し言葉 コーパス	会議録 + CSJ	会議録 +書き起こし	会議録 +パラレル	会議録 +書き起こし +パラレル
スタイル 総単語数	文書 71M	話し言葉 2.9M	話し言葉 71M+2.9M	話し言葉 71M+0.7M	話し言葉 71M+0.7M	話し言葉 71M+0.7M+0.7M
語彙サイズ Trigram エントリ数	30,386 3.63M	5,896 0.25M	31,019 3.77M	30,431 3.78M	30,431 4.14M	30,431 4.39M

品詞にバックオフすることにより，パラレルコーパスに出現しない単語に対しても変換が可能になるとともに，確率の推定がより頑健になることが期待される．求められた頻度を用いて，品詞文脈の場合（品詞ベース）の条件付き確率 $P_{\mathrm{POS}}(y_1^n|x_1^n)$ が (4) 式と同様に定められる．

なお，条件付き確率 $P_{\mathrm{word}}(x_1^n|y_1^n)\cdot P_{\mathrm{word}}(x_1^n|y_1^n)$ についても，これらの手順において x_1^n と y_1^n を入れ替えることで求めることができる．

3.3　変換モデルの適用

学習された変換モデルを，(3) 式に基づき文書スタイルの言語モデルに適用する．まず，言語モデルのそれぞれの N-gram 単語列について，単語ベースのパターンの適用を試みる．パターンに適合した場合は話し言葉の N-gram 単語列を生成するとともに，変換確率 $P_{\mathrm{word}}(y_1^n|x_1^n)/P_{\mathrm{word}}(x_1^n|y_1^n)$ をもとの N-gram 単語列の頻度に乗じることで，生成された N-gram の頻度を推定する．単語ベースのパターンが適合しない場合は品詞ベースのパターンの適用を試み，適合した場合は同様に変換確率 $P_{\mathrm{POS}}(y_1^n|x_1^n)/P_{\mathrm{POS}}(x_1^n|y_1^n)$ を用いて頻度を推定する．

4.　評価実験

4.1　評価タスク

本研究では国会音声の認識タスクで提案法の評価を行った．評価で扱う言語モデルの仕様を表 2 に示す．

ベースライン言語モデルは，1999 年から 2002 年までの 4 年分の衆議院会議録を用いて学習した．会議録ではフィラーや口語表現・文末表現などの典型的な話し言葉表現が削除・修正されているため，ベースラインモデルは文書に近いスタイルとなっている．テキストの総単語数は 71M，出現する trigram の異なり総数は 11M であり，このうち出現回数が 1 回の trigram エントリは言語モデルの構築時に除外している．

本研究では 2003・2004 年における衆議院の会議の一部について書き起こしを作成し，これらの会議の会議録とあわせてパラレルコーパスを構築した．コーパスの例を図 2 に示す．図 2 では，実際に発声されたにもかかわらず会議録では削除されている部分を中括弧 "{...}" で，発声されていないにもかかわらず挿入されている部分を括弧 "(...)" で示している．表現の置換が行われた個所は "{ 置換前/置換後 }" で示す．用意した書き起こしの総単語数は 666K である．

評価用のテキストは，同時期における別の会議の書き起こしである．これらの会議では経済・外交や安全保障などさまざまな話題が扱われている．話者の総数は 23 名であり，種々の話し言葉表現が観測される．評価テキストの総単語数は 63K である．

比較のため，我々がこれまで用いてきた，会議録と「日本語話し言葉コーパス」(CSJ) の混合モデル ("+CSJ") もあわせて評価を行った．CSJ は主として学術講演と模擬講演（自由なスピーチ）から構成されており，これらの音声についてフィラーや口語表現などが忠実に書き起こされている．したがって，これらの表現を含まない会議録と組み合わせることで，話し言葉音声認識のための言語モデルが構築できる．本研究では模擬講演のみを用いて言語モデルを構築し，ベースラインモデルと混合した．CSJ の学習テキストの総単語数は 2.9M である．なお，混合には相補的バックオフに基づく手法 [8] を用いている．

さらに，変換モデル学習用の書き起こしと会議録テキストを混合して学習したモデル ("+Transcript") も用意した．書き起こしは会議録と比べて総単語数が約 100 分の 1 と小さいため，テキストを 10 倍にコピーすることで重み付けを行った．この混合モデルの語彙は，後述する変換によるモデルと同一に設定している．

例 1:
{えー}それでは少し、今{その一}最初に大臣からも、{その一}貯蓄から投資へという流れの中に
{ま}資するんじゃないだろうかとかいうような話もありましたけれども、{だけど/だけれども}、
{まあ}あなたが言うと本当にうそらしくなる{んで/ので}{ですね、えー}もう少し{ですね、あの一}
これは{あー}財務大臣に{えー}お尋ねをしたいんです{が}。
例 2:
{ま}その{あの}見通しはどうかということでありますけれども、これについては、{あの一}委員
御承知の{その}「改革と展望」の中で{ですね}、我々の今{あの一}予測可能な範囲で{えー}見通せる
ものについてはかなりはっきりと書かせていただいて (い) るつもりでございます。
例 3:
{あの}このような問題をこの第一次提案{ていう/という}のは含んでいる。{で}お米はもちろん
だけれども、ほか{な/の}農産物にも多大な影響を及ぼし、さらにはこの後発途上国の問題なども
問題は含んでいるということを大臣 (は) お認めになったと思うので、これは毅然として{ですね}
はねのけていく{と}、これは確認できると思うんですけれども、もう一度お願いします。

図 2　国会音声の書き起こしと編集タグの例
Fig. 2　An example of tagged transcript of congressional meetings

4.2　実験結果と考察

まず，会議のパラレルコーパスを用いて変換モデルの学習を行った．低頻度の話し言葉表現による不適切なエントリの湧き出しを防ぎ，かつ変換確率の信頼性を確保するため，1 回しか観測されなかったパターンは除外した．この結果学習されたパターンの総数は 5,719 であり，うち 2,187 (38%) が品詞ベースのパターンである．

次に，この変換モデルをベースライン言語モデルに適用し，話し言葉スタイルの言語モデル (“Proposed-BL”) を生成した．生成されたモデルの語彙サイズは 30,431 となったが，これはベースライン言語モデルの語彙サイズとほぼ同等であり，CSJ 混合モデルのものより小さい．これは，変換モデルにより追加される単語は話し言葉表現に特有のもののみであるのに対して，CSJ 混合モデルでは無関係な話題語なども追加されるためである．一方，trigram エントリの総数は 4.14M となり，話し言葉表現に関する trigram を CSJ 混合モデル (3.77M) や書き起こし混合モデル (3.78M) よりも多く生成することができた．

本実験では，変換モデルを書き起こし混合モデルに対しても適用した．これにより生成されたモデル (“Proposed-Trn”) の trigram エントリ数は，書き起こし混合モデルよりも多い 4.39M となった．これは，書き起こしに存在しない trigram も変換モデルにより生成できたことを示している．

これらのモデルについて，評価テキストにおけるパープレキシティと trigram ヒット率により評価を行った．パープレキシティの計算に際しては未知語も算入されている．なお，未知語率はベースラインモデルでは 3.64%，CSJ 混合モデルでは 0.47%，書き起こし混合モデルと変換による 2 つのモデルでも 0.47%である．混合モデルおよび変換によるモデルでは，話し言葉特有の単語が追加されたことにより，ベースラインモデルと比較して未知語率が大きく改善されている．

図 3 に各モデルによるパープレキシティと trigram ヒット率を示す．ベースライン言語モデルのパープレキシティは 68.9 であるのに対して，CSJ 混合モデル (“+CSJ”) と書き起こし混合モデル (“+Transcript”) によるパープレキシティはそれぞれ 66.5，49.2 となった．提案法による，ベースラインモデルから変換されたモデル (“Proposed-BL”) と，書き起こし混合モデルから変換されたモデル (“Proposed-Trn”) のパープレキシティはそれぞれ 51.9，42.8 となった．CSJ 混合モデルでは trigram エントリが増加しているにもかかわらずパープレキシティの改善はほとんどみられないが，書き起こし混合モデルとベースラインから変換したモデルでは，パープレキシティをほぼ同様に削減している．これら 2 つのモデルでは trigram ヒット率も CSJ 混合モデルより高くなっており，同一ドメインの話し言葉テキストにより有効な trigram エントリを追加または生成できたためパープレキシティが改善できたと考えられる．さらに書き起こし混合モデルから変換したモデルにおいても，trigram ヒット率とパープレキシティのさらなる改善が見られる．これにより，同

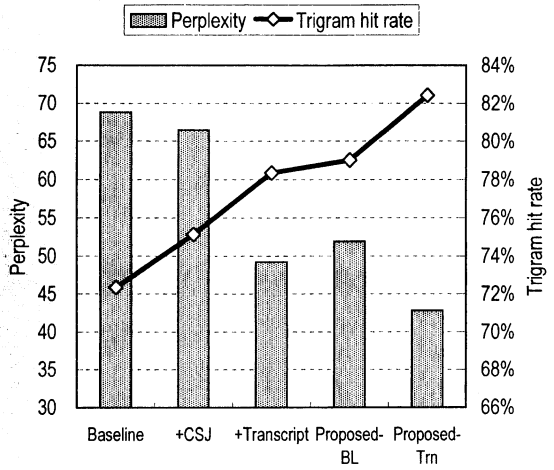

図 3 各言語モデルによるパープレキシティと trigram ヒット率

Fig. 3 Perplexity and trigram hit rates by the language models

ードメインのテキストを用いたコーパス混合手法でも補えない trigram についても．提案法により適切な変換パターンと確率が推定されて生成が可能になったといえる．

5. おわりに

本稿では，話し言葉の言語モデルを構築するための，統計的変換モデルに基づくスタイル変換の枠組みを提案した．変換モデルは，文書スタイルから話し言葉スタイルへの，文脈に応じた変換パターンと確率から構成される．これらのパターンと確率は，忠実な書き起こしとその整形テキストからなるパラレルコーパスにおける出現頻度をもとに定められる．提案法では変換パターンの文脈を品詞にバックオフすることより，規模の小さな学習コーパスでも頑健な変換モデルの構築が可能である．この変換モデルを文書スタイルの言語モデルの N-gram 頻度に適用し，話し言葉スタイルの N-gram とその推定頻度を求める．評価実験により，提案法が話し言葉言語モデルを効率的かつ効果的に生成でき，パープレキシティを削減できることが実証された．今後は音声認識における評価を行う予定である．

文　献

[1] A. Park, T. Hazen, and J. Glass. Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling. In *Proc. ICASSP*, Vol. 1, pp. 497–500, 2005.

[2] L. Lamel, G. Adda, E. Bilinski, and J.-L. Gauvain. Transcribing Lectures and Seminars. In *Proc. Eurospeech*, pp. 1657–1660, 2005.

[3] F. Metze, C. Fügen, Y. Pan, and A. Waibel. Automatically Transcribing Meetings using Distant Microphones. In *Proc. ICASSP*, Vol. 1, pp. 989–992, 2005.

[4] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of Conference Room Meetings: An Investigation. In *Proc. Eurospeech*, pp. 1661–1664, 2005.

[5] H. Schramm, X.L. Aubert, C. Meyer, and J. Peters. Filled-Pause Modeling for Medical Transcriptions. In *Proc. SSPR*, pp. 143–146, 2003.

[6] P. Brown, S. Pietra, V. Pietra, and R. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[7] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』version 2.2.3 使用説明書. Feb. 2001.

[8] 長友健太郎, 西村竜一, 小松久美子, 黒田由香, 李晃伸, 猿渡洋, 鹿野清宏. 相補的バックオフを用いた言語モデル融合ツールの構築. 情処学論, Vol. 43, No. 9, pp. 2884–2893, 2002.