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

2006－SLP－64 (43)
2006／12／22

音声認識システムSOLONの日本語話し言葉コーパスによる評価（2006年版）

中村 篤　大庭隆伸　渡部晋治　石塚健太郎　藤本雅清
堀 貴明　エリック・マクダーモット　南 泰浩

日本電信電話株式会社
NTT コミュニケーション科学基礎研究所 メディア情報研究部
信号処理研究グループ
〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: {ats, oba, watanabe, ishizuka, masakiyo, hori, mcd, minami}@cslab.kecl.ntt.co.jp

あらまし NTT コミュニケーション科学基礎研究所では実環境での自然な話し言葉を対象とした音声認識の研究を進めている．本報告では，そのテストベッドとして開発中の音声認識ソフトウェア'SOLON'の，日本語話し言葉コーパス(CSJ: Corpus of Spontaneous Japanese)を用いたベンチマーク評価結果を報告する．音声区間の事前検出，発話速度依存音声分析，言語モデルの誤り訂正学習，全共分散型モデルの識別学習，教師なし話者適応，及びそれらの組み合わせによる効果を実験により示す．

キーワード 重み付き有限状態トランスデューサ，発話区間検出，発話速度依存分析，誤り訂正学習，識別学習，教師なし話者適応

Evaluation of the SOLON Speech Recognition System: 2006 Benchmark using the Corpus of Spontaneous Japanese

Atsushi Nakamura, Takanobu Oba, Shinji Watanabe, Kentaro Ishizuka, Masakiyo Fujimoto
Takaaki Hori, Erik McDermott and Yasuhiro Minami

NTT Communication Science Laboratories
NTT Corporation, 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237

Abstract This article describes results from the latest benchmark tests of our speech recognition system 'SOLON' using the Corpus of Spontaneous Japanese (CSJ). The improvement in recognition accuracy using several techniques, including prior voice-activity detection, speaking-rate dependent analysis, corrective language modeling, discriminative training of full-covariance parameters, unsupervised model adaptation, and their combinations, are reported.

Keyword WFST, Voice Activity Detection, Speaking-rate Dependent Analysis, Corrective Training, Discriminative Training, Unsupervised Adaptation

1. はじめに

NTT コミュニケーション科学基礎研究所では音声認識研究のためのテストベッドとして音声認識システム'SOLON'の開発を進めている．SOLONには，重み付き有限状態トランスデューサ(WFST: Weighted Finite-state Transducer) の高速on-the-fly合成による効率的認識アルゴリズム[1]を中心とした，種々の独自技法，及び主要な標準技法が実装されており，現在も機能追加が続けられている．

大規模な自然発話音声データベースである日本語話し言葉コーパス(CSJ: the Corpus of Spontaneous Japanese)[2-4]は，現状のSOLONを評価し、より良いものにしていく上で欠かせない存在となっている．実際我々はこれまでもCSJを利用した大規模なベンチマークを行い，その結果を報告してきた[5,6]．本稿では2006年に新たに実施した種々のベンチマーク評価の結果をまとめて報告する．

2005年のベンチマーク[6]では，最終的に，不特定話者音響モデルで80%を超える単語正解精度を得た．今回はこれを新たなベースラインとして，諸技法とそれらの組み合わせにより，さらなる性能向上を目指す．具体的には，まず不特定話者での認識を前提として，発話区間の事前検出，発話速度に依存した音声分析，言語モデルの誤り訂正学習，全共分散型音響モデルの識別学習等を，まずは単独で適用する．効果の検証は，高速 on-the-fly 合成を包含する形で新たに拡張した認識アルゴリズムを用いて行う．続いて，複数技法の組み合わせにより精度向上の積み上げを図る．最後に会議録作成等，音声認識のオフライン的用途を念頭におき，教師なしの話者適応を試みる．

さらに，認識誤りの傾向を概観し，音響モデル改善による単語正解精度の上限についても考察する．

2. 2006年版ベンチマーク評価の基本設定

本報告では，我々が以前に行った評価(2005年版評価[6])で用いたものを包含する形で拡張された新しい

認識アルゴリズムを使用している．評価の流れとしては，2005年版評価における最終的な結果を出発点とし，新たに種々の技法を適用して性能向上を図るという形をとる．そこでまず，新しい認識アルゴリズムについて説明した上で，2005年版評価における最終的な結果(実験条件と性能)をベースラインとして再掲する．

2.1. 認識アルゴリズム -拡張高速 on-the-fly 合成法-

これまで我々は，WFSTに基づく音声認識システムを構築し，その性能改善や音声言語処理への応用を検討してきた．そして，WFSTを用い，高速かつ省メモリで動作する音声認識アルゴリズム「高速 on-the-fly 合成法」を提案し，語彙サイズ約185万の one-pass 実時間音声認識等を実現した[1]．高速 on-the-fly 合成法は，探索中に二つのWFSTの必要な部分だけを合成する on-the-fly 合成法の発展形であり，合成時の計算共通化により大幅な高速化を実現している．このアルゴリズムはさらに拡張され、現在では任意個のWFSTを高速合成することも可能となっている[7]．以下，任意個の WFST を効率的に on-the-fly 合成可能な「拡張高速 on-the-fly 合成法」について説明する．今 M 個の WFST $T_1, T_2, \ldots, T_M$ が存在し，これらを用いて記号列変換を行うことを考える．入力記号列 X が与えられたとき，これらのWFSTによる X から重み最小記号列への変換は

$$\Omega(X \to \hat{Y}_M \mid T_1,\cdots,T_M) = \min_{Y_1,\cdots,Y_M} \{\Omega(X \to Y_1 \mid T_1) + \Omega(X \to Y_2 \mid T_2) + \cdots + \Omega(Y_{M-1} \to Y_M \mid T_M)\} \quad (1)$$

を満たす $\hat{Y}_M$ を求める問題となる．ここで $\Omega(X \to Y \mid T)$ は，WFST T を用いて記号列 X を Y に変換するときの重みである．拡張高速 on-the-fly 合成法では，これを次のように書き換える．

$$\Omega(X \to \hat{Y}_M \mid T_1,\cdots,T_M) = \min_{Y_1}\{\Omega(X \to Y_1 \mid T_1) + \min_{Y_2}\{\Omega(X \to Y_2 \mid T_2) + \min_{Y_3}\{\cdots + \min_{Y_M}\{\Omega(Y_{M-1} \to Y_M \mid T_M)\}\cdots\}\}\} \quad (2)$$

この式で，T_2 から T_M までの累積重みに当る部分を $\Omega(X \to Y_1 \mid T_1)$ の補正項と捉えると，基本的には T_1 のみを用いて $\hat{Y}_1 = \arg\min_{Y_1} \Omega(X \to Y_1 \mid T_1)$ を求める手順が $\Omega(X \to \hat{Y}_M \mid T_M)$ の計算に適用できる．これはWFSTが二つの場合のみを想定した従来の高速 on-the-fly 合成法の自然な拡張であり，M=2 のとき，その手順は従来と同一になる．

拡張高速 on-the-fly 合成法による記号列変換では，WFST T_1 に基づいて one-pass Viterbi decoding を行い，その過程で生成される部分記号列仮説の累積重みを式(2)のように計算する．この際，各段階のWFSTから見て上位のWFSTを用いた部分記号列仮説の補正を，当該WFSTからの出力記号列が変化したときのみ上位に向かって順に繰り返し行う．特に，出力記号に ε (何も出力しないことを表す記号) が多い場合は，各段階で出力される記号列が変化する頻度が少ないため，補正計算の回数は少なく抑えられる．X の入力に沿って最後まで処理を進め，$\hat{Y}_1$ が求まった時点で瞬時に $\hat{Y}_2$ から $\hat{Y}_M$ までを求めて行くことができる．

本アルゴリズムを音声認識に適用した例としては，二つの WFST だけではメモリサイズの制約等により困難であったクラス N-gram 確率モデルや単語・クラス混合 N-gram 確率モデルの導入等において，既にその効果を示してきた[7]．本報告では，全体的な評価の枠組みとして，この拡張高速 on-the-fly 合成法を用いている．ベースラインを含め，従来通り単語 N-gram 確率モデルのみを言語モデルとして利用する際は M=2 とし，さらに上位に，学習データから得られる誤り傾向を考慮した補正WFSTを配置する場合(3.3節)は M=3 としている．

2.2. 実験条件とベースライン認識性能

表 1～4 に実験条件を示す．CSJ 公開版で規定されている「男性話者による学会 10 講演('test-set 1')」[3]を評価データとする．ベースラインにおいては，事前に発話単位への分割等を行うことなく，講演 1 回分全体をそのまま認識対象としている．性能は CSJ 公開版の短単位[3]による単語正解精度で評価する．学習データとしては，音響モデル用に評価データを除く男・女性話者による学会講演を，言語モデル用に評価データを除く学会・模擬の全講演を用いた．2005年版評価においては，音響モデルとして対角共分散型モデルに大語彙識別学習を施したもの(Diag-MCE)，全共分散型モデルを最大尤度学習したもの(Full-ML)がほぼ同等の性能となっていることから，この二種類をベースラインとして用意した．大語彙識別学習の具体的な方法については[17]を参照されたい．言語モデルは，学習データに出現する全単語(約 10 万語)を語彙としてカバーしており，backoff による N-gram 確率の平滑化手段として，Kneser-Ney(K-N)法[8]を用いている．

以上の条件の下で，拡張高速 on-the-fly 合成法(M=2)によって得られる性能をベースラインとして表 5 に示す．ビーム幅は精度が十分飽和する値を用い，言語モデルによる対数化先験確率へのスケーリング係数は 13 に設定している．これらが，2006年版ベンチマークの出発点となる．

表 1. 音声分析

標本化周波数/ 量子化精度	16kHz/16bit
特徴量ベクトル (39 次元)	(12 次 MFCC+ 対数パワー) + Δ + Δ²
分析フレーム長/周期	25ms/10ms (Hamming) Δwindow: ±2 frames
特徴量正規化	ケプストラム平均正規化

表 2. 音響モデル

音響モデル種別	3 状態 left-to-right HMM	
パラメータ学習	対角共分散型 (Diag-MCE)	最尤学習+ 識別学習
	全共分散型 (Full-ML)	最尤学習
音素カテゴリ数	43	
音素決定木質問数	144	

表 3. 言語モデル

言語モデル種別	単語 trigram 確率モデル
語彙数	100,808
N-gram 確率平滑化	Backoff 平滑化 (Kneser-Ney 法)
N-gram カットオフ	bigram:1, trigram: 1

表 4. 学習／評価データ（全て CSJ 公開版より）

音響モデル 学習データ	学会 961 講演 234 時間
言語モデル 学習データ	学会及び模擬 2,672 講演 異なり単語数 100,807 延べ単語数 6,830,031
評価データ	男性話者学会 10 講演 延べ単語数 26,329， パープレキシティ 75.3 OOV 率 1.36%

表 5. ベースライン認識性能（単語正解精度(%)）

Diag-MCE (5000 状態×32 混合)	**80.2**
Full-ML (2000 状態×16 混合)	**80.5**

3. 諸技法の実装と効果

種々の技法による性能の向上を試み，不特定話者音声認識における，各々の技法単独での効果を検証する．ここではベースライン音響モデルとして，Full-ML を用いている．

3.1. 音声区間の事前検出

CSJ は，主として講演音声を収録した音声データで構成されており，データ中には，発表スライドの切り替え等に伴う長時間の無音区間が含まれている．このような音声をそのまま認識すると，長時間の無音区間において発生する突発的な雑音等の影響により，挿入誤りが増加し，性能が劣化する恐れがある．そこで事前に音声区間検出(VAD: Voice Activity Detection)を適用して不要な無音区間を削除し，上記挿入誤りを抑制することを考える．同時にこれによって認識器への入力音声長が短縮され，認識に要する全体的な計算量を削減することもできる．

本報告における VAD は，非音声／音声状態遷移モデルに基づく手法[9]により行う．音声のオンセット，オフセット時における語頭，語尾の途切れを回避するために，VAD が検出した音声区間の両端には 100msec のマージンを付与する．全ての区間を再連結し，各講演単位でのケプストラム平均正規化を行った上で，認識器に入力する．

以上を Full-ML 音響モデルの下で適用した際の単語正解精度は **80.8%**であった(図 1 "Pre-VAD")．特に挿入誤りが 2 割近く減少しており，予想通り，無音区間での誤りが抑制されていることがわかる．また，VAD を用いることにより入力音声の総時間長が約 1 割短縮された．これにより，音声認識に要する処理時間も同様に約 1 割短縮されている．

3.2. 発話速度依存音声分析

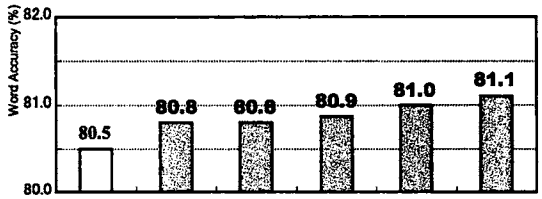
講演のような自然発話音声においては，読み上げ音声に比べ発話速度(以下「話速」)の変動が大きくなる．話速が平均的な値から外れると音声認識精度が劣化することが知られており[10]，話速に応じて HMM の状態遷移確率を変化させるなど，音響モデルにより対処する手法が提案されている[10-14]．ここでは，話速が速くなることで生じる調音のなまけや音節の短縮化を，音声分析の時間分解能を上げてより正確に分析することを考える．具体的には，フレーム周期 10 ms (Δwin.: ±2)で構成した音響モデルとフレーム周期 8 ms (Δwin.: ±3)で構成した音響モデルの二種類を用意した上で，評価データの各講演の平均話速を推定し，推定値に応じてフレーム周期 8 ms と 10 ms を使い分ける．

話速の推定は，音声信号の振幅変動の周波数スペクトルの 1 次モーメントである *Enrate* [15]により行った．この方法では，まず有限区間の音声波形を半波整流した後，遮断周波数 16 Hz の低域通過フィルタによって振幅変動を抽出する．さらにその周波数スペクトル $\Psi(k)$をもとに，当該音声区間内の推定音節数に対応する値を以下により求める．

$$Enrate = \frac{\sum_{k=s}^{K} k|\Psi(k)|^2}{\sum_{k=s}^{K}|\Psi(k)|^2} \qquad (3)$$

ここで K は振幅変動の周波数スペクトルのうち 16 Hz に相当する周波数ビン番号，s は直流成分の影響を受けない最小の周波数ビン番号である．今回の評価では，1 講演の平均話速に対応する値として，1 秒を単位区間とし 500 ms ずつずらしながら各区間の *Enrate* を算出した上で，全区間の値を平均したものを用いた．

以上を適用した結果得られた単語正解精度は **80.8%** であった(図 1 "Enrate-Dep. Analysis")．話速が高いと推定され，フレーム周期 8 ms で分析を行った話者 4 名の内 3 名について向上が見られ．特にベースラインにおいて最も正解精度の低い話者(A03M0156)での向上の幅は 2.5%に達した．全体的には，主として脱落誤りが減少している．これはまさに音声分析の時間分解能を上げたことの効果と考えられる．なお，[20]では，CSJ



	Baseline (Full-ML)	Pre-VAD	Enrate-Dep. Analysis	Sentence-Div LM	Corrective LM	Full-MCE
WER	19.5	19.2	19.2	19.1	19.0	18.9
Sub	12.4	12.3	12.5	12.3	**12.0**	12.4
Del	4.8	5.0	**4.4**	4.8	4.6	**4.3**
Ins	2.3	**1.9**	2.3	**2.0**	2.0	2.2

図 1. 各技法単独による精度の向上

を用いた評価においてフレーム周期を一律8msとしている．今回このように一律に分解能を上げた分析を試みた場合でも，同等の効果が得られた．しかし，話者による効果のばらつきは大きく，10名中4名で精度が向上，5名で劣化，1名は変化なしという結果となった．

3.3. 言語モデルの学習データ分割と誤り訂正学習

・言語モデルの学習データ分割

話し言葉においては書き言葉の場合に比べ，文に相当する言語的単位が曖昧であると言われている．このことから，ベースライン言語モデルの学習に用いたCSJのデータでは基本的に各講演を，文境界を意識しない一連なりの単語列と見なしていた．一方で，同データには，「文と認定できる」とされる単位の切れ目(絶対境界)，発話の大きな切れ目(強境界)等，文境界相当の情報も，自動処理(一部は人手処理)で付与されている[3]．さらには，長い無音区間で挟まれた単語列の各々を便宜的に文と定義することも考えられる．

文境界を明示的に意識し，データを分割して言語モデルを学習することで，途切れてしかるべき単語並びの存在が，誤った統計量として言語モデルに反映されることを抑止できる等の利点が考えられる．同時に，本来曖昧な境界を確定的に決めてしまうことで言語モデルの頑健性が損なわれる危険もある．これらが実際の音声認識性能にどのような影響を与えるのか，興味深いところである．ここでは，CSJで付与された絶対境界，同じく絶対境界+強境界，0.5秒以上の無音区間の三種類を各々文境界と考えて学習した言語モデルを用いる場合の認識性能を検証した．

まずCSJで付与された境界を用いた場合，単語正解精度は絶対境界で80.5%，絶対境界+強境界で80.4%となり，いずれも精度の向上は見られなかった．一方無音区間を文境界と考えた場合は**80.9%**となった(図1 "Sentence-Div. LM")．主として挿入誤りが削減されており，音声－無音境界付近での不要な単語仮説の生成が抑えられたものと考えられる．

・言語モデルの誤り訂正学習

N-gram確率モデルには，相対的に短い系列や，曖昧性の小さい経路に対応する系列の先験確率が不正に高くなることがあるという問題が存在する．また通常，音響モデルと言語モデルは独立に学習されており，これを相補的な関係とすることにより，全体的な性能の向上を見込むことが出来る．これらを踏まえ，音響モデルの学習データから得られる誤りの傾向をもとに言語モデルを補正することを試みる．

まず学習データを入力とする音声認識の結果からn-best単語列リストを得る．リストの中で正解に最も近い単語列(oracle単語列)に注目し，これに含まれる単語N-gramが優先され，それ以外が棄却されるようにパーセプトロン・アルゴリズムを用いて誤り訂正学習を進める[16]．具体的には，まず各単語列について，N-gram頻度を素性値とする素性ベクトルを考える．単語列リストを素性ベクトルとパラメータベクトルの内積値に基づく降順に並べ替え，最上位がoracle単語列でない場合にパラメータベクトルから最上位単語列の素性ベクトルを差し引き，oracle単語列の素性ベクトルを加える，という手順を繰り返す．これにより，最終的にパラメータベクトルの各要素には個々のN-gramに対応する補正値が得られることになる．

ここではさらに，拡張高速on-the-fly合成法(2.1節)の枠組みを用い，one-pass Viterbi decodingの中で言語モデルの補正を実現する．上記補正値に基づく重みをもつ補正WFSTを構成し，通常の二つのWFSTのさらに上位に配置する．これは式(2)において$M=3$として，累積重み最小の単語列を探索することに相当する．

言語モデル補正後の単語正解精度は**81.0%**であった(図1 "Corrective LM")．誤り訂正学習の効果として，置換誤りが0.4%削減されている．

3.4. 全共分散型モデルの識別学習

2005年版評価では対角共分散型の音響モデルのみを対象として大語彙識別学習の効果を実証した[17]．ここでは全共分散型モデルについても同様に大語彙識別学習を実装し，その効果を検証する．

ここで用いる大語彙識別学習は2005年版評価と同様に，string-level最小識別誤り(MCE)学習の考え方から派生した，線形の損失関数を用いる定式化に基づいている．大語彙連続音声認識を対象として効果的に学習を進めるために，対立単語列生成用の言語モデル，先験確率付き正解単語列の双方をWFST形式で表現する．これらのWFSTを用いた音声認識の結果をもとに算出される，学習データ全体についての損失値(総損失)を最適化手法によって最小化する[17]．

全共分散型のモデルに対して最適化手法を適用するために，共分散パラメータについての偏導関数を導入する．今，多変量$\mathbf{x}$のガウス密度関数$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$を考え，逆共分散行列$\boldsymbol{\Sigma}^{-1}$のCholesky分解を$\boldsymbol{\Sigma}^{-1}=\mathbf{L}\mathbf{L}'$としたとき，$\mathbf{L}$についての$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$の偏導関数は，以下により与えられる[18]．

$$\frac{\partial \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \mathbf{L}} = \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\left\{(\mathbf{L}^{-1})' - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\mathbf{L}\right\} \qquad (4)$$

これを利用して総損失の導関数も容易に求めることが出来る．

最適化手法としてRprop法を用いた大語彙識別学習後の音響モデルによる単語正解精度は**81.1%**であった(図1 "Full-MCE")．ベースラインからの精度向上幅(0.6%)は単独の技法によるものとしては最も大きなものとなった．系列を全体として正解に近づけていくstring-levelの識別学習により，ここでは特に脱落誤りの削減が顕著となっている．

4. 諸技法の組合せと教師なし話者適応

諸技法の組み合わせによって，不特定話者音声認識性能のさらなる向上を図る．教師なし話者適応の効果も併せて考察する．

4.1. 諸技法の組合せによる効果

複数の技法間で効能の重複があれば，それらを単純に組み合わせても個々の精度向上幅を積み上げたような結果が得られないことは容易に想像できる．図1で各技法による誤り削減の傾向を見ると，大まかな組み合わせの相性が見えてくる．例えば発話区間事前検出(Pre-VAD)と文分割(Sentence-Div. LM)とでは，いずれも主に音声区間の両端付近や長い無音区間中に湧き出す挿入誤りが削減しているものと考えられ，これらの組み合わせは重複的ということになる．実際に様々な

組み合わせを試みた結果を図 2 に示す．予想通り，Pre-VAD 後に Sentence-Div. LM を適用してもほとんど上積みは得られない．また，ベースラインを Diag-MCE とした場合，Pre-VAD，Sentence-Div. LM 共に単独適用の時点で効果が見られなかった．これは，事前の識別学習が無音モデルの識別能力を高めており，両技法による誤り削減の余地を減らしたためと考えられる．言語モデルの誤り訂正学習(Corrective LM)と全共分散型モデルの識別学習(Full-MCE)は，図 1 の誤り削減傾向を見る限り効能の重複が少ない．実際，両者の組み合わせでは，ほぼ加法的に精度向上が積み上げられている．最終的に，組み合せによる不特定話者認識としての単語正解精度最良値は **81.9%** (ベースラインからの誤り削減幅: 1.4%，相対削減率: 7.2%)となった．

続いて言語単位を，仮名漢字，仮名，音素として認識結果単語列からの再集計により正解精度を算出し，ベースラインと比較した(表 6)．どの言語単位でも相対誤り削減率は 7%以上となっている．さらに同一条件での評価を別のデータ(CSJ「男女 5 名ずつによる学会講演('test-set 2')」[3])を用いて行った結果(表 7)から，諸技法の効果が特定の評価データに依存したものではないことがわかる．なお，'test-set 2'からは，言語モデルの学習データにも含まれている話者の講演を除外し，男性 4 名，女性 5 名の評価データとして用いている．

4.2. 教師なし話者適応による効果

講演音声の認識は会議録作成等，オフラインでの用途も想定され，その場合特に，音響モデルの教師なし話者適応が精度の向上に有効と考えられる，そこで，今回得た不特定話者モデルを初期モデルとする教師なし話者適応について考察を行った．

標準的なモデル適応法として，ガウス分布の平均ベクトルを対象とした最尤線形回帰(MLLR: Maximum Likelihood Linear Regression)法を用いる[19][20]．スパースなデータからの回帰行列推定において過学習を回避するため，事前に初期モデルのガウス分布を平均ベクトルのユークリッド距離をもとにクラスタリングして木構造を作り，各ノードに割り当てられたガウス分布同士で回帰行列を共有する．データの各フレームとガウス分布は，初期モデルを用いた認識の結果に基づいて対応付ける．これは WFST を用いた大語彙識別学習[17]と類似の枠組みで実装されている．適応後のモデルを初期モデルと入れ替え，同じ手順を必要に応じ繰り返す．なお今回，全共分散型モデルに対しては，回帰行列推定時に対角成分のみが寄与する近似的解析解を用いている．教師なし適応において仮にデータの音声区間が予め分かっていれば，境界付近の音素コンテキストをより正確に与えることができるという意味で有効であろう．そこで，3.1 節の検出法[9]により，データから音声区間を事前に切り出すことを考える．対角共分散型モデルの適応(1 回)では，切り出しなしで 83.4%，切り出しデータで 83.6%となり，音声区間情報が適応において有効であることがわかる(図 2)．これを受け，全共分散型モデルに対し，切り出しデータを用いた適応を 2 回繰り返した結果，**83.8%**の単語正解精度を得た(図 2)．教師なし言語モデル適応との組み合わせ[20]等により，さらなる精度の上積みが期待できる．

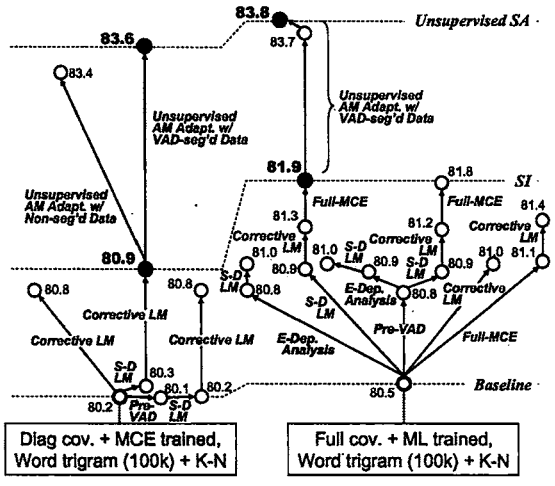


図 2. 諸技法の組合せと教師なし話者適応

表 6. 種々の言語単位に基づく正解精度 (test-set 1)

単語 (CSJ短単位)	仮名漢字	仮名	音素
80.5→**81.9**	83.6→**84.9**	88.2→**89.1**	90.5→**91.2**

(ベースライン→**2006 年版評価**)

表 7. 種々の言語単位に基づく正解精度 (test-set 2)

単語 (CSJ短単位)	仮名漢字	仮名	音素
81.7→**83.3**	84.3→**85.7**	89.3→**90.4**	91.4→**92.4**

(A01M0056 を除く；ベースライン→**2006 年版評価**)

5. 考察

5.1. 誤り傾向の概観

不特定話者認識として最良の単語正解精度(81.9%)をもたらした実験の認識結果テキストをから，残された誤りの内容を大まかに調べた．助詞を中心とする一拍語の置換または脱落誤りや，間投詞同士の置換誤りが目立つこと等，誤り傾向の大勢に 2005 年版評価[6]との差異は認められない．間投詞の扱いを考慮した再評価(詳細は[6]参照)の結果は，全ての間投詞を単一化した場合 83.6%，間投詞を除外した場合 84.0%となり，間投詞に閉じた誤りの全体に占める割合も 2005 年版評価と同等(1 割程度)であった．総じて，今回試みた諸技法の組合せによる効果は，特定種別の誤り削減に偏らず，あまねくもたらされているといえる．

5.2. 理想音響モデルによる単語列同定

現在の標準的な音声認識技術において音響モデルの果たす役割の比重は非常に大きく，音響モデルの改善により，多くの認識誤りが回避可能であると考えられている．しかしながら，特に日本語に多いとされる同音異義語等，音響モデルでは対処が不可能な場合も存在する．ここでは「理想的な音響モデルを手にした」との仮定の下で，WFST の枠組みを用い，単語列の同定がどの程度可能か検証する．

まず評価データ('test-set 1')の書き起こしテキスト中の仮名表記を音素表記に変換し，対応するHMM状態の列を得る．これを仮想的な音声入力として，WFSTデコーダにより，基本的に認識と同様のプロセスで重み最小単語列への変換を行う．音響モデルから与えられる重み付けの代わりに，WFSTデコーダが仮説として生成するノードの音素と入力されるHMM状態の音素が一致しているとき0，それ以外は1を重みの値とする．これにより，書き起こしテキストの仮名表記と単語発音モデルの間に不整合がない限りにおいて，単語列同定の精度は言語モデルに委ねられることとなる．

ベースライン言語モデルを用い，実際に単語列同定を行った結果として得られた単語正解精度は92.9%であった．残りの誤り(7.1%)は現在の枠組みでは解消が難しい言語的な曖昧性がもたらすものということになる．今回は音素列の同定が完全に誤りなく行われることを前提としたが，これは必ずしも現実的でないと考えられるため，ベースライン言語モデルとの組み合わせにおいては，音響モデル改善による単語正解精度の上限は90%弱付近と見るのが妥当と推測される．

6. まとめ

本稿では音声認識システムSOLONの，日本語話し言葉コーパス(CSJ)を用いたベンチマーク評価の結果を報告した．最終的にCSJ 'test-set 1'による最良の単語正解精度は不特定話者音声認識で**81.9%**，音響モデルのみの教師なし話者適応後で**83.8%**となった．これらを始めとして，今回実験結果として得た正解精度の各々は，我々独自の技法，既存の技法，及びそれらの組み合わせによるものである．今後，独自技法のさらなる研究と併せて，有力な既存技法の評価も引き続き検討していく．近年の動向からは，効果が大きいとされる技法として，識別的特徴量，話者正規化，教師なし言語モデル適応，複数の音声認識システムからの出力統合等が挙げられる[21].

謝辞「日本語話し言葉コーパス」の構築に尽力されたすべての皆様に敬意と謝意を表します．

文　献

[1] T. Hori, C. Hori and Y. Minami, "Fast on-the-fly composition for weighted finite-state transducers in 1.8 million-word vocabulary continuous-speech recognition," in Proc. Interspeech'04, pp. 289–292, 2004.

[2] 国立国語研究所, "The Corpus of Spontaneous Japanese," http://www2.kokken.go.jp/~csj/public/index_j.html.

[3] 国立国語研究所, "CSJ Release Information," http://www2.kokken.go.jp/csj/public/releaseinfo/index.htm.

[4] 古井貞熙, "一里塚としての『日本語話し言葉コーパス』," 音講論集(2006春), 3-1-4, pp. 1191-1194, 2006.

[5] 堀貴明, 渡部晋治, マクダーモットエリック, 南泰浩, 中村篤, "音声認識システムSOLONの日本語話し言葉コーパスによる評価," 第3回 話し言葉の科学と工学ワークショップ, pp. 85-91, 2004.

[6] 中村篤, 大庭隆伸, 渡部晋治, 石塚健太郎, 堀貴明, マイク・シュスター, エリック・マクダーモット, 南泰浩, "音声認識システムSOLONの日本語話し言葉コーパス(公開版 Ver.1.0)による評価," 信学技法, SP2005-106, pp. 7-12, 2005.

[7] T. Hori and A. Nakamura, "Generalized fast on-the-fly composition algorithm for WFST-based speech recognition," in Proc. *Interspeech'05*, pp. 557-560, 2006.

[8] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in Proc. ICASSP'95, vol. 1, pp. 181-184, 1995.

[9] 藤本雅清, 石塚健太郎, 加藤比呂子, "音声と雑音両方の状態遷移過程を有する雑音下音声区間検出," 情処研報, 2006-SLP-64-3, 2006.

[10] M. A. Siegler and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in Proc. ICASSP'95, pp. 612-615, 1995.

[11] N. Mirghafori, E. Fosler and N. Morgan, "Towards robustness to fast speech in ASR," in Proc. ICASSP'96, pp. 335-338, 1996.

[12] H. Nanjo, K. Kato and T. Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," in Proc. Interspeech'01, pp. 2531-2534, 2001.

[13] J. Zheng, H. Franco and A. Stolcke, "Effective acoustic modeling for rate-of-speech variation in large vocabulary conversational speech recognition," in Proc. Interspeech'04, 2004.

[14] T. Shinozaki and S. Furui, "Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation," in Proc. ASRU'03, pp. 417-422, 2003.

[15] N. Morgan, E. Fosler and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in Proc. Eurospeech'97, pp. 2079-2082, 1997.

[16] B. Roark, M. Saraclar and M. Collins, "Corrective language modeling for large vocabulary ASR with the perceptron algorithm," in Proc. ICASSP'04, pp. 749-752, 2004.

[17] E. McDermott, T.J. Hazen, J. Le Roux, A. Nakamura and S. Katagiri, "Discriminative training for large vocabulary speech recognition using minimum classification error," To appear, Jan 2007, *IEEE Trans. ASLP*.

[18] V. Valtchev, "Discriminative methods in HMM-based speech recognition," Ph.D. thesis, University of Cambridge, 1995.

[19] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[20] 阿部拓也, 草間隆, 武田千春, 加藤正治, 小坂哲夫, 好田正紀, "日本語話し言葉コーパスを用いた教師なし適応による講演音声認識の性能改善,"音講論集(2006春), 3-1-8, pp. 1205-1206, 2006.

[21] M. J. F. Gales and P. C. Woodland, "Recent progress in large vocabulary continuous speech recognition an HTK perspective," in Tutorial Notes ICASSP'06, Tutorial 10, 2006.