音声言語情報処理　44－20
(2002. 12. 19)

講演音声認識のための音響・言語モデルの検討

堤　　怜介† 　加藤　正治† 　小坂　哲夫† 　好田　正紀†

† 山形大学工学部
〒992-8510 米沢市城南 4-3-16
Tel: 0238-26-3365 FAX: 0238-26-3365

あらまし　現在，音声認識は不特定話者大語彙連続音声認識の枠組みにおいて，新聞記事などの読み上げ音声で実用化レベルまで精度が向上している．しかし，講演音声などのいわゆる話し言葉からなる自然発話は，認識を困難にさせる要素が多数存在し，まだまだ実用化の域には達していない．本研究の目的は，自然発話の認識が困難な要因について，音響的な観点と言語的な観点に着目し，自然発話用の音響モデルと言語モデルを作成することである．音響モデルに関しては，音声学習データ選択を行い，言語モデルに関しては，発音変形依存の形態素で学習することで高精度なモデルを作成した．適応を行わない認識率では31.4%の WER を得た．また，第2パス言語モデルに様々なカットオフや4gram を導入した．4gram による効果は無く，カットオフの効果は0.6%であった．更に，音響モデルを話者適応し，3.2%の改善を得た．

キーワード　講演音声，音響モデル，音声学習データ選択，言語モデル，発音変形依存，話者適応．

Acoustic and Linguistic Modeling for Lecture Speech Recognition

Ryousuke TSUTSUMI†, Masaharu KATOH†, Tetsuo KOSAKA†, and Masaki KOHDA†

† Faculty of Engineering, Yamagata University
4-3-16 Jonan Yonezawa-shi, 992-8510, Japan
Tel: 0238-26-3365 FAX: 0238-26-3365

Abstract　Large vocabulary continuous speech recognition (LVCSR) became practical application level for a newspaper read-speech. However, in LVCSR of spontaneous utterance such as a lecture speech etc. there exist many problems which make recognition difficult. The purpose of this paper is to make precise acoustic and linguistic models for LVCSR system of spontaneous utterance. Speech training data selection is performed to make acoustic model, and morpheme analysis of pronunciation variant dependency is executed to make linguistic model. WER of 31.4% was obtained without speaker adaptation of acoustic model. Trigram and 4-gram with various cutoffs were introduced to rescore language likelifood. WER inprovement by introducing cutoffs was 0.6%. Furthermore, speaker adaptation of acoustic model was carried out, and WER improvement of 3.2% was obtained.

Key words　lecture speech, acoustic model, speech training data selection, linguistic model, pronunciation variant dependency, speaker adaptation.

1. はじめに

日本語話し言葉コーパス (CSJ) の講演音声は自然発話に近いので，発音の怠け・言い淀み・非言語音の共起など，人が聞いても判別が困難な音声を多く含んでいる．それらは音響的特徴が曖昧で，音響モデル作成の学習データとしては相応しくない[4]．本稿では，タグに基づいて音声学習データを分類し，高精度な音響モデルを作成する．また，話し言葉は発音変形が多いので，発音変形を考慮した言語モデル・単語辞書を作成することで，更なる認識精度向上を目指す．さらに予備実験として，音響モデルの話者適応を行った．

2. 認識システム

第1パスで triphoneHM-Net 及び単語 bigram を用いて単語グラフを生成し，第2パスで単語 trigram を用いて単語グラフをリスコアする，2-Pass デコーダを用いる．第1パスでは，単語間の音素環境を考慮する．また，音素グラフに基づく仮説制限法[1]によって，単語グラフの精度を劣化させることなく，処理時間を約30%減少させる．この手法は，音素グラフを利用して，単語列の探索における仮説数を削減する．音素グラフから脱落した音素を含む単語が認識不能になることはない．

講演音声を適当な単位に分割して認識を行なう．ここでは，「5秒経過後にあらわれる最初の200msec以上の無音区間」で分割する．これは，文の単位と必ずしも一致しない．

3. 講演音声の学習・評価セット

『日本語話し言葉コーパス』モニター版2001のデータを用いる．

3.1 学習セット

音響モデルに関しては男性話者の計115の学会講演 (約25時間：全データ30時間中，発話とは独立した長い非言語音5時間を除いたもの) を用い，言語モデルに関しては男性・女性話者の計384の学会講演および模擬講演 (単語総数913214) を用いる．

3.2 評価セット

男性話者4名の学会講演で，学習セットに含まれない AS22(A01m0035)，AS23(A01m0007)，AS97(A01m0074)，PS25(A05m0031) である．

4. 音響モデルの作成

4.1 音響モデルの構成

音響モデルは，始めに4〜6状態の長い音素環境依存モデルを学習し，3状態音素環境独立 HMM を並列に連結した HM-Net を初期 HM-Net として，それらの状態対応確率を用いた状態クラスタリング (コンテキスト方向+時間方向の分割) を行い，各 triphone 3〜6状態，総状態数2000，1状態あたり16混合ガウス分布の HM-Net を作成した[2]．音素カテゴリーは34音素+無音の計35音素とする．

★タグの詳細

比較的きれいな音声 α(17時間)

なし	タグが付いていない部分
(F)	フィラー，感動詞
(D)	言い直し
(M)	音や言葉に関する引用
(A)	漢字仮名以外の文字の使用
(K)	漢字表記できなくなった場合
(S)	未登録の口語表現
(H)	母音の引き延ばし

音響的に曖昧な音声 β(6時間)

(W)	言い間違い，発音の怠け
(O)	外国語や古語，方言など
(笑)	非言語音との共起
(L)	ささやき声などのかなり小さな声
(Q)	子音の引き延ばし
(FV)	母音不確定音

発話とは独立した非言語音(5時間)

〈雑音〉	ベル音・拍手・長い咳などの雑音

発話中の非言語音 γ(2時間)

(R)	差別語などで音声自体が削除
〈笑〉	短い笑いや咳などの非言語音
(?)	聞き取り，語彙同定に自信なし

図1 学習データのタグによる分類

表1 状態数毎の triphone モデルの数

	状態数					
	3	4	5	6	合計	平均状態数
AM0 (23h)	13453	5010	306	66	18835	3.31
AM1 (17h)	14083	4479	218	55	18835	3.27
AM2 (6h)	14968	3531	292	44	18835	3.22

CSJ 書き起しの発音形を参照して音素ラベルを作成する．その際に，タグ (W) など「言語的に正しい発音」と「聞こえたまま」の2種類が付いているような曖昧性がある場合は，「聞こえたまま」を使用している．音響モデルの学習では，単語間に無音が入ることも考慮した音素ラベルを用いる．

4.2 音声学習データ選択

講演音声には，発音の怠けや言い淀み等，音響的に曖昧な要素が多い．また，雑音などの非言語音は学習データに使えない．それらを学習データから除くことで，精度の良いモデルが学習できると考える．そこで学習データ選択のために，CSJ 書き起こしに付加されているタグ[3]を用いて，学習データを図1のように分類する．

作成した音響モデルは，図1の学習データ ($\alpha+\beta$) の23時間モデル AM0，学習データ α の17時間モデル AM1，学習データ β の6時間モデル AM2 である．AM2 は，AM1 と比較のために作成した．

モデル AM0，AM1，AM2 において，各 triphone モデルの状態数の集計結果を表1に示す．これらのモデルの特徴は3状態固定ではなく時間方向に延ばしたモデルを含んでいる点である．話し言葉音声は，長音化が起こりやすいので，長母音など時間的に長い音素には大きな状態数がうまく対応している．また5〜6状態の HMM にはコンテキストに無音を含むモデルが9割以上対応しており，言い詰まりや言い止まりによる無音の多発化が起こりやすい講演音声に特に有効である．

5. 言語モデルの作成

5.1 言語モデルの構成

言語モデルは単語 N-gram で，語彙 20k，カットオフなし，バックオフは Witten-Bell descounting を使用した．第1パス用として bigram，第2パス用として trigram を作成した．

表 2　言語モデルの性能

	茶筅のみ LM1		発音変形依存 LM2	
講演名	PP	Cover	PP	Cover
AS22	143.7	97.1 %	137.8	98.0 %
AS23	129.6	98.3 %	121.8	99.1 %
AS97	122.6	97.0 %	104.3	97.2 %
PS25	147.6	97.7 %	128.4	98.1 %

5.2　発音変形依存モデル

講演音声などの話し言葉においては，話し言葉に忠実な読みを持ち，発音変形を考慮した形態素解析データで学習することにより，高精度な言語モデルを作成できる．ここでは，これを発音変形依存モデルと呼ぶ．CSJ 書き起こしには，基本形（漢字・仮名で綴った表記に揺れがない文）と発音形（音声を忠実にカタカナで綴った正解文）が存在する．そこで，発音形を用いて発音変形を考慮した形態素解析データを作成し，そのデータで言語モデル及び単語辞書を作成する．例として，単語「音素」に対して，「音素+オンソ+2」「音素+オーンソ+2」「音素+オンソー+2」の様に読みの異なる複数の形態素解析データを作成する．発音変形依存形態素解析のアルゴリズムを以下に示す．

アルゴリズム

(1)　基本形を "茶筅+茶碗" で形態素解析

(2)　形態素の読み部を正解 (発音形) の読みに訂正

(a)　複数読みを一意に決定

Ex. 一日+{イチニチ/ツイタチ}+2

⇒ 一日+ツイタチ+2

(b)　母音の長音化

Ex. 音声+オンセイ+2 ⇒ 音声+オンセー+2

(c)　1 文字単位でマッチング

Ex. 声質+セイシツ+2 ⇒ 声質+コワタチ+2

この場合では，[声] と [質] が持つ全ての読みの組み合わせに関してマッチング．

[セイシツ][セータチ][コエシツ][コワタチ]...

(3)　マッチングがとれない場合は，茶筅の読み部をそのまま用いる．

茶筅のみの形態素解析データで作成した言語モデルを LM1，提案法の発音変形を考慮した発音変形依存モデルを LM2 とする．単語辞書に関しては，LM1 を用いるときは茶筅形態素，LM2 を用いるときは発音変形依存形態素で作成し，語彙は 20k である．語彙のカットオフは行わず，上位 20k を語彙とした．その内，約 4k は出現頻度 1 の単語である．

LM1,LM2 のパープレキシティ(PP) とカバレージ (Cover) を表 2 に示す．評価データの 4 講演すべてにおいて，パープレキシティとカバレージが改善されており，モデル LM2 が精度良く作成できていることが分かる．

表 3　音響分析条件

標本化・量子化	16kHz,16bit
フレーム長	32ms (ハミング窓)
フレーム周期	8ms
特徴ベクトル	12 次元の LPC メルケプストラム係数と対数パワー，その 1 次と 2 次の回帰係数 計 39 次元
正規化	発話単位のケプストラム平均正規化

6.　評価実験

6.1　音響分析

音響分析条件を，表 3 に示す．

6.2　評価方法

デコーダの性能は，WER(Word Error Rate)，GER(Graph Error Rate)，WGD(Word Graph Density) の 3 つの値で評価する．

(1)　**WER**：認識精度を計る指針となる．

$$WER\,[\%] = \frac{S+I+D}{N} \times 100$$

ここで，N は正解単語列の単語数，S は置換誤り数，I は挿入誤り数，D は脱落誤り数を示す．

(2)　**GER**：単語グラフ内に正解単語列が含まれない割合 (単語グラフ内の最も正解系列に近い候補の WER) で，単語グラフの精度を示す指針である．この値は第 2 パス WER の下限値を表す．

(3)　**WGD**：正解 1 単語当たりの単語仮説数を表し，単語グラフの大きさを計る指針となる．

$$WGD = \frac{\text{単語グラフの単語数}}{\text{正解単語列の単語数}}$$

誤り率を計算する際は，単語の表記部分のみを比較した．正解と認識結果の間で単語の切れ目がずれることがあるが，誤りとみなした．また，同じ意味で表記が異なる単語，例えば，"ホルマント" と "フォルマント"，フィラーの "えーと" と "えっと"，言い誤りなどは全て誤りとした．

6.3　実験結果

音響モデル AM0，AM1，AM2 と，言語モデル LM1，LM2 の種々の組み合わせによる評価データの認識実験を行った．

音響モデル AM1 を用いて，言語モデルを変えた実験結果を表 4，表 5 に示す．表 4 は発音変形依存モデルの効果をみるために形態素解析データによる性能比較，表 5 は種々のカットオフによる trigram および 4gram を用いたリスコアの性能比較である．また，言語モデル LM2 を用いて，音響モデルを変えた実験結果を表 6，表 8 に示す．表 6 は音声学習データ選択による性能比較，表 8 は話者適応による性能である．

更に，モデル (AM1,LM2) について，評価データに付加されているタグ別にまとめた認識結果を，表 9 に示す．

表 4　形態素解析データによる言語モデルの性能比較 (%)

音響モデル		AM1			
言語モデル		LM1		LM2	
講演名	単語数	WER	GER(WGD)	WER	GER(WGD)
AS22	6328	45.5	17.5 (176)	40.6	13.5 (172)
AS23	4437	37.7	15.2 (186)	25.4	6.7 (156)
AS97	2530	32.7	12.6 (354)	24.7	8.1 (346)
PS25	5503	40.9	14.7 (479)	31.3	8.4 (419)
4 講演	18798	40.6	15.5 (299)	32.1	9.7 (273)

6.4　言語モデルの比較

6.4.1　発音変形依存モデルの効果

表 4 において，モデル (AM1,LM1) と (AM1,LM2) の 4 講演 WER を比較すると，8.5%(相対値で 21%) の精度向上となり，提案法の発音変形依存モデルが有効に働いている．一般に，形態素解析では形態素分割の比重が大きいが，講演音声などの話し言葉においては環境や話者の違いで読みが多様に変化するため，形態素の読み部を如何に適切に解析するかが重要になる．

第 1 パス WER を表には記さなかったが，第 2 パスにおける trigram リスコアリングの効果が (AM1,LM1) と (AM1,LM2) で，それぞれ 1.8%(42.4%→ 40.6%) と 0.4%(32.5%→ 32.1%) しかない．しかし，単語グラフの精度は，(AM1,LM2) の場合，GER 9.7%(WER より 22.4%減)・WGD 273 と非常に良いグラフが作成できている．第 2 パスのリスコアリングで精度が上がらない要因は，言語モデル用の学習データが約 91 万語と少なく，trigram の十分な統計量が得られていないことが大きい．

6.4.2　種々のカットオフによる trigram および 4gram を用いたリスコア

これまでは，第 2 パス言語モデルとして trigram のカットオフなしを用いてきた．カットオフをするとエントリ数が激減するためである．具体的には，カットオフなしでは約 55 万，カットオフ 1 では約 8 万となる．4gram は認識時間が非常に大きくなるため用いなかった．

しかし，第 1 パス WER から第 2 パス WER の改善は殆ど見られず，第 2 パスにおける何らかの対策が必要である．そこで，種々のカットオフによる trigram および 4gram を用いたリスコアにより，精度向上を目指す．認識実験の結果を表 5 に示す．単語グラフはモデル (AM1,LM2) から生成したものを使用した．

種々のカットオフを用いた比較では，trigram と 4gram のカットオフを適切に設定することにより精度の向上が見られた．ただし，bigram までカットオフすると逆に精度が劣化する傾向が見られる．bigram，trigram，4gram のカットオフの最適値はそれぞれ，0，1，2 である．

最終的に最も認識率の良かった言語モデルは，trigram で WER 31.5%，4gram で WER 31.4%であり，ベースラインのカットオフなし trigram との比較で，それぞれ 0.6%(32.1%→ 31.5%) と 0.7%(32.1%→ 31.4%) の改善を得た．

しかし，GER との差が未だに大きいので，十分な結果とは言い難い．第 2 パスにおける更なる改善が望まれる．最も改善が期待できるのは，言語モデル用の学習データの増加と考える．

表 6　学習データ選択による音響モデルの性能比較 (%)

音響モデル	AM0		AM1		AM2	
言語モデル	LM2					
講演名	WER	GER	WER	GER	WER	GER
AS22	39.1	12.7	40.6	13.5	41.8	14.2
AS23	24.8	6.5	25.4	6.7	28.4	8.4
AS97	25.0	8.6	24.7	8.1	28.2	9.9
PS25	30.7	8.2	31.3	8.4	33.1	9.9
4 講演	31.4	9.4	32.1	9.7	34.2	11.0

表 7　音声学習データにあらわれる triphone の種類数

	23h(AM0)	17h(AM1)	6h(AM2)
triphone 数	7818	6856	6720

6.5　音響モデルの比較

6.5.1　音声学習データ選択の効果

表 6 において，学習データが比較的少ないにも関わらず，30%程度の誤りと高精度な認識ができている．

音響的に曖昧な部分を含めた学習データで作成した音響モデル AM2 は AM1 と比べると，4 講演 WER で 2.1%劣化している．特に，AS23 や AS97 などの認識性能が良い講演ほど，WER の劣化が大きい傾向が見られる．これは音響的に綺麗な講演音声である．逆に AS22 や PS25 などの認識性能が悪い講演は，WER の劣化が小さい．これは音響的に曖昧な部分が多い講演音声である．これらの音声に対しては，音響的に曖昧な音声を用いた音響モデルを効果的に作成することによって，精度向上が期待できる．

しかし，音声学習データ選択を行わない 23 時間モデル AM0 が最も精度が良く，データ選択の効果は見られなかった．データ量の違いが一番の要因であると考えられるため，HM-Net の構造を決定する triphone の数 (学習データにあらわれる triphone の種類数) を，表 7 に示す．23 時間のデータが，17 時間および 6 時間と比較して，約 1000 個だけ triphone の数が多い．このことが 17 時間音響モデルが 23 時間音響モデルより WER が劣化している大きな原因だと考えられる．比較的きれいな音声データと全てのデータで，triphone の種類数に差がなくなるまで学習データの量が増加した場合に，提案法の比較的きれいな音声データから作成の音響モデルが有効に働くことが期待できる．

6.5.2　音響モデルの話者適応

講演音声などの自然発話では，ベースラインのモデルでは認識率の大幅な改善は得られず，話者性や発話速度などの適応により効果的にモデルを作成できる [5]．

そこで本節では，音響モデル適応の予備実験として，MLLR を用いた話者適応を行った．MLLR の回帰クラスタ数はフレーム閾値によって自動的に決定する．つまり，全てのモデルを木の根に位置付け，各モデルに対応する学習データの総フレーム数が閾値より大きい場合に，質問の条件をもとに葉に分類する．

表 5　種々のカットオフによる trigram および 4gram の性能 (%).

LM	cut-off (n-grams)			2-pass WER (%)					PP
	n=2	n=3	n=4	AS23	AS22	AS97	PS25	average	
trigram	0	0	—	25.4	40.6	**24.7**	31.3	32.1	126.3
	0	1	—	**24.6**	39.8	25.2	**30.3**	**31.5**	**114.0**
	0	2	—	24.8	40.0	25.5	30.5	31.7	114.2
	1	1	—	25.1	39.7	25.8	30.9	31.8	118.0
	1	2	—	25.0	**39.5**	26.2	31.1	31.8	118.3
	2	2	—	25.9	40.2	25.8	31.6	32.3	123.6
4gram	0	0	0	26.0	41.0	25.1	31.4	32.5	138.3
	0	0	1	25.4	40.6	24.7	31.1	32.1	125.6
	0	0	2	25.6	40.7	**24.5**	31.2	32.2	124.9
	0	1	1	24.7	**39.7**	24.9	**30.2**	**31.4**	113.4
	0	1	2	**24.6**	**39.7**	25.1	**30.2**	**31.4**	**112.7**
	0	2	2	**24.6**	**39.7**	25.7	30.5	31.5	113.0
	1	1	1	25.1	**39.7**	25.6	30.8	31.7	117.4

※ モデル（AM1,LM2）を使用して作成した単語グラフをリスコア.

表 8　音響モデルの話者適応の性能 (%)

	Baseline	教師なし	教師有り
AS22 PER	—	19.9	0
WER	39.1	36.2	31.2
AS23 PER	—	10.9	0
WER	24.8	22.2	18.0
AS97 PER	—	9.6	0
WER	25.0	21.5	16.6
PS25 PER	—	14.7	0
WER	30.7	26.9	21.9
Ave PER	—	14.8	0
WER	31.4	28.2	23.4

※ Baseline は AM0 を使用.

フレーム閾値は 2000 とした．適応データは各評価データの全てを用いた．その結果，回帰クラスタ数は教師有り・なし共に 4 講演すべてで 15〜16 個になり，1 個のクラスタに対して 2 個の変換行列 (平均用 1 個，分散用 1 個) でパラメータを推定する．変換パラメータは，混合分布の重み・平均ベクトル (フル変換行列)・分散ベクトル (対角変換行列) の 3 つである．MLLR の学習回数は 1 回とした．

教師有り適応は，CSJ 書き起しから得られる正解単語列を音素に変換した音素ラベルを用いて適応し，教師なし適応は，ベースラインモデルから得られる単語認識結果を音素に変換した音素ラベルを用いて適応する．

音響モデルとして，Baseline(AM0)・教師なし適応・教師有り適応を用いた認識結果を表 8 に示す．表中の PER は，適応に用いた音素ラベルの音素誤り率 (Phoneme Error Rate) である．教師なし適応モデルの 4 講演 WER は 28.2%となり，ベースラインから 3.2%の向上を得た．教師有り適応モデルは，8%の向上であった．PER が平均 14.8%と比較的良いが，期待したほど教師なし適応の効果は見られなかった．今後は，教師なし適応を Iteration の枠組みに導入し，さらに MAP 適応と組み合わせることで更なる改善を計る．

6.6　タグ別の認識結果

表 9 に，モデル (AM1,LM2) を使用した評価セット 4 講演の認識結果を，タグ別に分類した結果を示す．

タグ (F) に関しては GER で 3.2%となり，単語数も多いことから，改善が望まれる．GER が低いタグに関しては，第 2 パス言語モデルで，ある程度の補正ができる．しかし，タグ (D),(W),(L) などは GER が 30%以上とかなり悪い．GER が悪いタグに関しては，第 1 パスにおいて音響的な観点での対策が必要である．タグが付いていない音声の WER は 28.8%となり，精度良い認識ができている．タグも含めた WER 32.1%と比べて，3.3%の精度改善がなされている．タグの付いている音声を精度良く認識する重要性がわかる．

α と β と γ の比較では，α が精度良く，β と γ が非常に悪い．β と γ はタグそのものの認識率も悪いが，それによって二次的に周りの単語も認識できなくなると予想される．よって，β と γ を精度良く認識することが，講演音声における最大の課題と言えよう．

7.　ま　と　め

音響モデルの学習データ選択法によって，データの質に焦点を当てた音響モデルを作成し，その可能性を示した．また，発音変形を考慮した形態素解析データを用い，言語的な要素だけでなく音響的な要素も加えた言語モデル・単語辞書を作成し，認識精度が向上することを確認した．適応をしないベースラインモデルでは 4 講演平均で 31.4%の WER であった．

今後は，音響的に曖昧な要素を多く含むデータに対して有効なモデルを検討し，また音響モデルや言語モデルの話者適応を行い，総合的な認識精度の向上を目指す．

表 9　評価セットのタグ別認識結果 (%)

タグの種類	図 1 の分類記号	単語数	WER	GER
タグなし	α	15651	28.8	9.4
(F)	α	1743	39.9	3.2
(D) , (D2)	α	399	77.4	31.3
(M)	α	100	77.0	20.0
(A)	α	256	49.2	12.1
(K)	α	4	75.0	25.0
<H>	α	93	32.2	11.8
<C>	α	18	61.1	11.1
計	α	18164	31.3	9.6
(W)	β	250	63.6	38.0
(笑) , (泣) , (咳) , (あくび)	β	87	51.7	31.0
(L)	β	117	85.4	40.1
<Q>	β	27	40.7	3.5
<FV>	β	7	42.8	14.2
計	β	488	65.1	35.0
<息>, <笑>, <泣>, <咳>	γ	3	33.3	0.0
(?)	γ	43	81.3	27.9
計	γ	46	78.2	26.1
合計	α+β+γ	18798	32.1	9.7

※ 但し，モデル (AM1,LM2) を使用.

文　　献

[1] 堀 貴明，岡 直生，加藤 正治，伊藤 彰則，好田 正紀："大語彙連続音声認識のための音素グラフに基づく仮説制限法の検討"，情報処理学会論文誌,Vol.40,No.4,PP.1365-1373,1999.

[2] 堀 貴明，加藤 正治，伊藤 彰則，好田 正紀："状態クラスタリングによる HM-Net の構造決定法の検討"，信学論 (D-II),J81-D-II,No.10,pp.2239-2248(1998).

[3] 小磯 花絵："日本語話し言葉コーパスの書き起こし基準"，話し言葉の科学と工学ワークショップ講演予稿集,pp.13-20,2001.

[4] 南條 浩輝，河原 達也："講演音声認識のための種々の形態素解析及び音響モデルの評価"，話し言葉の科学と工学ワークショップ講演予稿集,pp.47-52,2002.

[5] 南條 浩輝，河原 達也："発話速度に依存したデコーディングと音響モデルの適応"，信学技報 Vol.101 No.523, SP2001-103, pp.7-12 (2001-12)