

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

南條 浩輝 河原 達也

京都大学大学院 情報学研究科 知能情報学専攻

〒 606-8501 京都市左京区吉田本町

e-mail: {nanjo,kawahara}@kuis.kyoto-u.ac.jp

あらまし 大語彙の話し言葉音声認識における発話速度に関する問題に焦点をあてた認識手法について報告する。話し言葉音声では発話速度は一般に速く変動も大きいため、認識が困難である。実際に認識結果において、特に速い音声で認識率が低いこと、及び発話速度の速い音声と遅い音声では認識誤り傾向に明確な差があることを確認した。そこで、発話速度に応じて最適な音響分析フレーム・音素モデル・デコーディングパラメータを選択的に適用し認識を行う手法を提案する。発話速度の自動推定を組み合わせることにより認識率の向上を得た。さらに、発話速度情報を話者適応に用いる手法についても検討を行う。速い発話と遅い発話のそれぞれを指向した異なる話者適応モデルを構築しそれらを選択的に適用することで、速度情報を用いない適応よりも効率的な適応が行えることを確認した。

キーワード 音声認識、話し言葉、講演、発話速度、音響モデル、話者適応

Speaking-Rate Dependent Decoding and Adaptation for Spontaneous Lecture Speech Recognition

Hiroaki Nanjo Tatsuya Kawahara

Graduate School of Informatics, Kyoto University

Sakyo-ku, Kyoto 606-8501, Japan

e-mail: {nanjo,kawahara}@kuis.kyoto-u.ac.jp

Abstract This paper addresses the problem of speaking rate in large vocabulary spontaneous speech recognition. In spontaneous lecture speech, the speaking rate is generally fast and may vary a lot within a talk. We also observed different error tendencies for fast and slow speech segments. Therefore, we first present a speaking-rate dependent decoding strategy that applies the most adequate acoustic analysis, phone models and decoding parameters according to the speaking rate. Several methods are investigated and their selective application leads to accuracy improvement. We also propose to make use of speaking-rate information in speaker adaptation, in which the different adapted models are set up for fast and slow utterances. It is confirmed that the method is more effective than normal adaptation.

key words automatic speech recognition, spontaneous speech, lecture speech, speaking rate, acoustic model, speaker adaptation

1 はじめに

我々は、開放的融合研究「話し言葉工学」で構築されつつある大規模な日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese) [1][2] を用いて講演音声の認識の研究を行っている。その初期的実験評価 [3][4] においては、書き言葉・読み上げ音声で作成されたモデルではきわめて低い性能しか得られず、話し言葉のデータでモデル化することの重要性を示したが、60%程度の認識率しか得られなかった。さらに学習データを増加させた上で、話し言葉に対応したデコーディング手法を実装したものの、65%程度の認識率にとどまっており、音響モデルのより高精度なモデル化が必要であることを確認した [5]。

話し言葉に対する音響モデルに関する検討は、いくつかの機関で行われており [6][7]、発話速度や発話速度の変動が認識率低下の主要因の一つと報告されている。発話速度の変動に関しては、実際に人間による聴取実験の結果でも発話速度の変動が大きい発話では認識が困難であると報告されている [6]。CSJ においても同一話者の同一講演内でも発話速度の大きな変動が見られることが報告されている [1]。このため、同一のモデルやパラメータを一律に適用することには問題があると考えられる。実際の認識結果でも、発話速度による誤り傾向の違いは顕著であり、速い音声では挿入・置換誤りが多く、遅い音声では削除誤りが多い。

発話速度に関しては、特に速度の速い音声が問題である。速度の速い音声区間では調音が明確になされないため、同一音素であっても周波数パターンが変化する (アンダーシュート現象)。さらに、音素自体が消失している可能性もある。そこで速い発話速度を考慮した音響モデルの研究が行われてきており [8][9][10]、我々も既に検討を行ってきた [11]。これら先行研究においては話者独立の音響モデルの研究が行われているが、本稿では、話者依存の音響モデルについての発話速度の考慮を検討する。話者依存の音響モデル作成には話者適応手法がよく用いられるが、同一話者内でも速い音声での周波数パターンは異なるため、発話速度を考慮せずに全ての発話を用いて適応を行うことの妥当性に疑問が生じる。

本稿では、発話速度に依存したデコーディング手法及び話者と発話速度両方に音響モデルを適応させる枠組みについて検討を行う。

表 1: テストセット

講演名 (略称)	単語数	時間	認識率	PP
A01M0035 (AS22)	6294	28	58.9	133.5
A01M0007 (AS23)	4391	30	72.4	107.5
A01M0074 (AS97)	2508	12	72.5	117.2
A05M0031 (PS25)	5372	27	64.7	164.4
A02M0117 (JL01)	9833	57	62.7	186.8
A03M0100 (NL07)	2644	15	68.0	94.5
A06M0134 (SG05)	4460	23	58.6	111.8
KK99DEC005 (KK05)	6527	42	64.7	127.7
YG99JUN001 (YG01)	2759	14	61.5	125.5
YG99MAY005 (YG05)	3108	15	67.2	117.8
total	47896	263	64.2	135.1

認識率はベースラインモデルによる単語正解精度 (%)
時間は分, PP はテストセットパープレキシティを示す

2 学習セットとテストセット

CSJ は学会講演と特定の話題について独話した模擬講演から構成される。本研究では、音響モデルの学習に学会講演の音声 (224 講演, 37.9 時間) のみを用いる。これは、報告 [3] で、模擬講演のデータを加えても認識精度に効果がなかったためである。ただし現在、男性用の性別依存モデルしか作成していないため、女性話者のデータは用いていない。

言語モデルの学習には 2001 年 2 月の時点で使用できる全ての書き起こしテキスト (612 講演, 148 万形態素) を用いる。

テストセットは表 1 に示す 10 名の話者の講演である。いずれも講演に熟練した男性話者による講演であり、原稿を用いずに話している。

3 ベースラインシステム

音響モデルは混合連続分布 HMM (対角共分散) に基づいて作成する。音響分析は、フレーム長 25ms (ハミング窓)、フレーム周期 10msで行った。各フレームで 12 次元のメル周波数ケプストラム係数 (MFCC)、その一次差分 (Δ MFCC) とパワーの一次差分 (Δ Power) を計算し、計 25 次元の特徴量ベクトルを求める。音素カテゴリは 43 種類で /sil/ を除いて IPA モデル [12] と同一である。各音素は 3 状態 left-to-right (飛び越し遷移なし) HMM でモデル化した。ベースライン音響モデルとしては男性用の性別依存 PTM triphone モデル (128 混合 x129 コードブック; 2000 状態) を用いる。

言語モデルには、語彙数 19k の単語 3-gram モデルを用いる [5]。これは配布されているモニタ版に含まれているモデルと同一のものである。カバレッジは 97% で、テストセットパープレキシティは 135 である。形態素解析

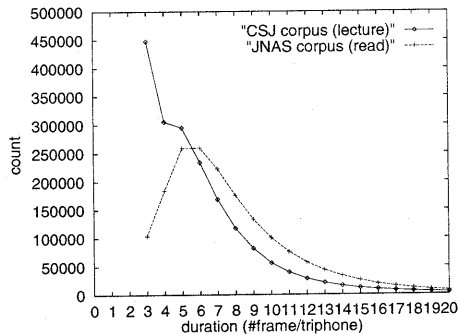


図 1: 講演音声 (CSJ) と読上げ音声 (JNAS) における音素の継続時間長分布の比較

システムとしては Chasen ver2.02 を用いており、単語 (形態素) の定義はそれに基づいている。デコーダには Julius-3.1 を用いた。

ベースラインシステムによる認識率の平均は 64.2% であり、話者ごとの認識率と共に表 1 に示されている。

4 発話速度の分析

4.1 音素継続時間長の分布

講演音声 (CSJ: 学会講演 35 時間) と読上げ音声 (JNAS: 約 40 時間) における各音素の継続時間長分布を図 1 に示す。継続時間長はビタビアルゴリズムに基づいて求めた。CSJ の学会講演音声は明らかに発話速度が速いことがわかる。特に 3 フレーム (30ms) にマッチングしている区間が多く、これらの中には、実際は 3 フレーム未満の時間長にも関わらず強制的に 3 状態のモデルにマッチングさせられているものも含まれると考えられる。

4.2 発話速度と認識率の関係

発話速度と認識率の関係をテストセットについて調査した。本稿では発話速度を、(各発話に含まれるモーラ数)/(各発話の時間)と定義する。ただし、ここでの発話 (入力単位) は、講演音声を一定のポーズで機械的にセグメンテーションしたものであり、言語的な文の単位と必ずしも一致しない。総発話数は 10 講演分で合計 2517 となった。

図 2 に発話速度と認識率の関係 (発話速度ごとの置換・脱落・挿入誤りの内訳) を示す。発話速度が速い音声は全体的に認識が困難であることが確認できる。また、置換・脱落誤りは発話速度が速い発話ほど多く、挿入誤りはそ

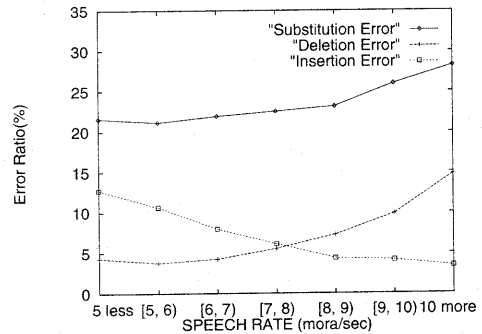


図 2: 発話速度ごとの置換・脱落・挿入誤りの内訳

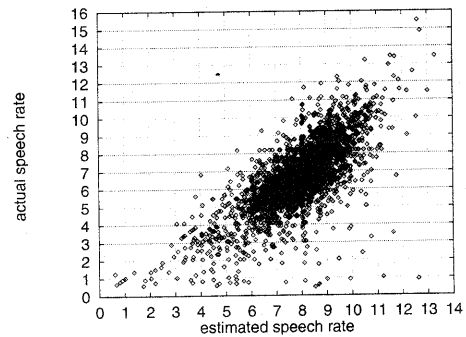


図 3: 実際の発話速度と推定された発話速度の関係

の逆の傾向にあることも確認できる。このことは、本実験で求めた発話速度が、局所的な発話変動を完全には反映していないものの妥当であることを示唆している。

4.3 発話速度の自動推定

次に、発話速度の自動推定を試みた。各発話に含まれるモーラ数は音節認識 (任意の音節の任意の接続を許す文法による) を行い、短母音と促音、撥音を 1 モーラ、長母音を 2 モーラとしてカウントし求める。

実際の発話速度と推定発話速度の関係を図 3 に示す。これら 2 者の相関係数は 0.74 であり、高い相関が確認された。これは、推定発話速度を用いる妥当性を示している。実際の発話速度と推定発話速度の平均と分散はそれぞれ 6.63, 2.05, 7.86, 1.62 であった。

5 発話速度別デコーディング

前章での分析に基づき、発話速度に依存して異なるデコーディングを適用する手法を提案する。これは、現在の発話区間に対し発話速度を推定し、最も適切な音響分

表 2: 種々のモデル・認識パラメータによる発話速度ごとの認識率 (%)

実際の発話速度 (発話数)	-5 (433)	5-6 (434)	6-7 (596)	7-8 (435)	8-9 (343)	9-10 (161)	10- (115)	合計 (2517)
ベースライン	61.3	64.5	65.9	65.9	65.3	60.1	53.6	64.2
1. (音響分析)	60.3	65.5	66.5	66.9	67.2	61.7	56.1	65.3
2. (飛び越し)	62.3	66.0	66.6	67.2	65.8	60.7	54.8	65.2
3. (音節モデル)	59.6	64.6	66.2	65.9	66.6	61.1	56.2	64.7
1.+2.	59.0	64.0	65.1	65.3	65.3	60.4	56.0	63.8
1.+3.	56.0	61.8	64.4	65.5	66.0	62.4	56.5	63.5
2.+3.	60.5	64.5	66.3	66.3	66.1	62.8	57.0	64.9
1.+2.+3.	54.3	60.7	63.4	64.9	66.2	62.0	57.9	62.9
4. (挿入ペナルティ)	64.3	67.3	66.4	64.7	62.8	55.8	50.1	63.7
発話速度 既知	64.3	67.3	66.6	67.2	67.2	61.7	56.1	65.9
推定発話速度 使用	62.6	66.4	66.7	66.9	66.4	60.6	55.8	65.4

析フレーム、音素モデル、デコーディングパラメータを選択する手法である。

本報告では、以下に示す 4 種類のモデル・パラメータの適用を行った。1. から 3. は速い発話を指向したものであり、4. は遅い発話を指向したものである。

1. 音響分析の変更

講演のような発話速度が速い音声認識においては音響分析のフレーム長やフレーム周期を変更することが有効である [13]。そこで、フレーム長を 20msec、フレーム周期を 8msec (共にベースラインの 80%) に変更する。

2. 飛び越し遷移を許すモデル

各音素 HMM (3 状態 left-to-right) の第 1 状態から第 3 状態に遷移を付加する。発話速度の速い講演音声認識に有効であることが確認されている [11]。

3. 音節単位モデル

速く発声され、かつ頻出する音素コンテキストを求め、それらを音節でモデル化する [11]。音節 HMM は音素 HMM を結合して作成し、飛び越し遷移を付与した上で遷移確率と重みのみ再学習を行うことによりモデル化を行う (図 4)。ただし、第 1 状態から第 5, 6 状態などの大幅な飛び越しは付与しない。

4. 単語挿入ペナルティの変更

遅い発話に対しては単語挿入誤りが増えるため、単語挿入ペナルティを厳しくすることにより対処する。ここでは -2 から -8 に変更した。

上記の 4 手法及びそれらを組み合わせたものを用いて認識を行った結果を表 2 に示す。全体的な認識率では、1. 音響分析のフレーム長・フレーム周期を短くする手法や、2. 飛び越し遷移を許すモデルがベースラインに比べて認識率を 1% 改善している。1. の手法は速い発話に対する認識率の改善が大きいのにに対し、2. の手法は全体的に認識率を改善しており、遅い発話に対する改善も大きい。3. の音節モデルは他の 2 手法に比べて認識率が若干低いも

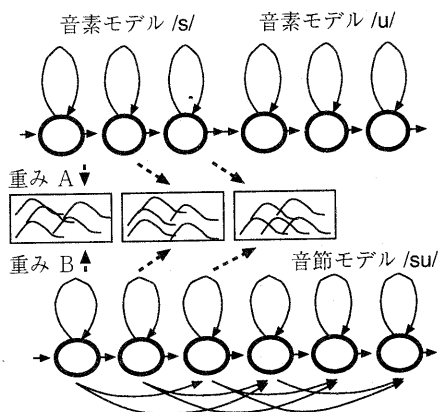


図 4: PTM 音節モデル

の速い発話に対してはよい性能を示している。

しかし、これらの 3 手法を組み合わせたものは、どのように組み合わせても認識率が低下した。これは主に遅い発話に対する精度が悪化したためであり、速い発話 (特に 9 モーラ/秒以上) に対する精度は向上している。また、遅い発話に関しては、4. の単語挿入ペナルティを厳しくする手法が効果的であった。

発話速度を求めて、これらを選択的に適用する場合、6 モーラ/秒未満の遅い発話に対しては 4. の手法を、8 モーラ/秒以上の速い発話に対して 1. の手法を、それ以外には 2. の手法を用いることで、認識率は 65.9% となる。自動推定した発話速度に基づいた場合は 65.4% であった。

6 発話速度を考慮した話者適応

次に、MLLR[14]を用いた話者適応に発話速度情報を利用することを考える。通常、講演は十分に長い (データとしてたくさんあるため) 教師なし話者適応の枠組み

表 3: 発話速度別の話者適応による認識率 (%) – 教師付き学習

実際の発話速度 (発話数)	-5 (433)	5-6 (434)	6-7 (596)	7-8 (435)	8-9 (343)	9-10 (161)	10- (115)	合計 (2517)
ベースライン	61.3	64.5	65.9	65.9	65.3	60.1	53.6	64.2
全体で適応	69.3	72.3	72.5	71.3	69.7	62.8	53.8	70.0
速い発話のみで適応	67.1	70.7	71.1	72.3	70.6	63.8	55.2	69.7
遅い発話のみで適応	70.4	72.6	72.9	69.8	67.8	59.7	51.1	69.1
選択的適用	70.4	72.6	72.9	72.3	70.6	63.8	55.2	70.7

速い発話速度の発話が多いテスト話者 5 名に対する認識率

(発話数)	(110)	(105)	(186)	(222)	(237)	(130)	(90)	(1080)
全体で適応	65.8	72.8	70.4	68.6	67.4	51.5	52.1	66.3
速い発話のみで適応	60.8	69.8	69.3	68.3	68.0	62.1	53.4	66.1
遅い発話のみで適応	69.5	73.5	72.3	66.1	66.5	58.1	49.1	64.7
選択的適用	69.5	73.5	72.3	68.3	68.0	62.1	53.4	67.0

速い発話速度の発話が少ないテスト話者 5 名に対する認識率

(発話数)	(323)	(329)	(410)	(221)	(106)	(31)	(25)	(1437)
全体で適応	70.0	72.2	73.2	74.1	76.0	71.5	63.5	72.9
速い発話のみで適応	68.3	70.9	71.7	76.3	77.9	75.1	65.6	72.5
遅い発話のみで適応	70.5	72.4	73.1	73.5	73.8	70.2	62.3	72.6
選択的適用	70.5	72.4	73.1	76.3	77.9	75.1	65.6	73.6

が有効に働くと考えられる。

まず、教師付き学習でその効果を調べた。以下に、発話速度を考慮した話者適応の手順を示す。

1. 人手による書き起こしから音素ラベルを作成する
2. 速い発話と遅い発話の境界を決定する。本実験では予備実験の結果に基づき境界を 7 モーラ/秒とした。
3. MLLR 適応を行う。速い発話 (7 モーラ/秒以上の発話) のみを用いて MLLR 適応を行い、速い発話に適応させたモデルを作成する。同様の処理を遅い発話 (7 モーラ/秒未満の発話) に対しても行い、遅い発話に適応させたモデルも作成する。比較のために、全ての発話を用いて適応させたモデルも作成する。

3. で得られた 3 種類のモデルを用いて認識実験を行った。結果を表 3 に示す。速い発話のみで適応を行ったモデルは単純に全ての発話を用いて適応を行ったモデルよりも速い発話に対して認識率を改善している。また、遅い発話に対しても同様の傾向が見られる。発話速度を考慮して適応を行い、認識時に選択的にモデルを適用することができれば単純に適応することに比べて認識率を 0.7% 改善できることがわかる。

テストセットを詳細に調査した結果、話者間で発話速度の分布に非常にばらつきが多いことがわかった。実際に、テスト話者 10 名中の 5 名で速い発話の大部分が占められており、残りの 5 名の発話には速い発話が少ない。速い発話をほとんど持たない話者とそうでない話者に対する発話速度別の話者適応の結果を表 3 の中段と下段に示す。どちらの話者群に対しても選択的適用を行った結

果、全体として 0.7% の認識率の向上であった。ただし、速い発話を十分に持つ話者 5 名では、速い発話と遅い発話両方に認識率の向上が見られるのに対し、速い発話が少ない話者 5 名では、遅い発話に対する認識率の向上がほとんど見られないが、速い発話に対する認識率は大幅に改善されている。このことは、後者では遅い発話が発話の大部分を占めるため、遅い発話に対する発話速度別の話者適応の効果が小さいものの、速い発話に関しては少量であるにも関わらず、そのみを用いることの効果が高いことを示唆している。

さらに、この認識率の差 (0.7%) の検定を行ったところ、有意水準 5% で有意であり、速度別の話者適応には効果があるといえる。

次に、教師なし学習での話者適応実験の結果を表 4 に示す。教師なし話者適応のプロセスは以下の通りである。

1. ベースラインの話者独立な音響モデルを用いた音声認識結果から音素列を生成しラベルを作成する。
2. このラベルを用いて MLLR 適応を行い話者依存の音響モデルを作成する (全体で適応 x1)。
3. この話者依存の音響モデルで講演を再認識し同様の処理を繰り返し、話者依存の音響モデルを更新する (全体で適応 x2)。

1 度目の適応で 4% の認識率の向上がみられ、2 度目の適応でさらに 0.6% の向上がみられた。この結果、68.8% の認識率を得た。教師付き話者適応 (人手で与えた正しい音素ラベルを用いた MLLR 適応) の結果では認識率は 70.0% であり、教師なしの話者適応でも MLLR 適応の効果は高いことがわかる。

表 4: 発話速度別の話者適応による認識率(%) – 教師なし学習

実際の発話速度 (発話数)	-5 (433)	5-6 (434)	6-7 (596)	7-8 (435)	8-9 (343)	9-10 (161)	10- (115)	合計 (2517)
ベースライン	61.3	64.5	65.9	65.9	65.3	60.1	53.6	64.2
全体で適応 x1	66.7	68.7	70.0	69.5	69.2	65.9	55.2	68.2
全体で適応 x2	68.6	69.7	71.0	70.1	69.4	63.5	54.6	68.8
速い発話のみで適応 x2	68.1	69.7	70.7	70.3	69.8	64.5	56.9	69.0
遅い発話のみで適応 x2	69.0	69.9	71.1	69.4	69.0	64.5	54.0	68.8
選択的適用(発話速度既知)	69.0	69.9	71.1	70.3	69.8	64.5	56.9	69.2
選択的適用(推定発話速度使用)	68.6	70.1	70.9	70.1	69.9	64.5	56.9	69.1

発話速度を考慮した話者適応は、上に示すプロセス中の 2. の MLLR 適応を行う際に、速い発話/遅い発話のみを使用する点のみ異なるだけである。ただし、発話速度の推定は全て話者独立の音響モデルを用いて行っている。同様に適応、認識、ラベル作成の処理を 2 度繰り返してモデルを作成した。

発話速度を考慮した教師なし話者適応によって、発話速度 9 モーラ/秒 以上の速い発話に関しては認識率が改善されているものの、発話数(単語数)が少なく、全体での認識率の改善はわずかであった(68.8% → 69.1%)。

教師ありと教師なしの話者適応の違いはラベルの正確さのみである。主に速い発話において、発話速度を考慮した適応の効果がみられることは、速い発話では周波数パターンが異なるため、精度の高いラベルが得られなくても分離して適応する効果があることを示唆している。

今回は推定発話速度による認識率低下は見られなかった。改善は小さいものの、適応を速い/遅い発話ごとに行うために、少ないデータ・短い時間で、全体を用いて適応を行うのと同様以上の効果が得られることがわかった。

7 まとめ

発話速度に依存したデコーディングと適応手法について検討を行った。発話速度に依存したデコーディングは、発話速度を推定し、最も適切な音響分析フレーム、音素モデル、デコーディングパラメータを選択することによって行った。種々の手法について実験を行い、これらを選択的に適用することの効果を示した。また、発話速度情報を話者適応に利用することを提案し、発話速度別の話者適応が妥当に機能することを示した。

謝辞 本研究は、開放的融合研究『話し言葉工学』プロジェクトの一環として行われた。アドバイスを頂きました東京工業大学の古井貞照教授をはじめとして、ご協力を頂いた関係各位に感謝いたします。本研究を進めるにあたって貴重な御意見を頂きました、京都大学教授 奥乃博 先生に感謝いたします。

参考文献

- [1] 前川喜久雄. 言語研究における自発音声. 日本音響学会研究発表会講演論文集, 1-3-10, 春季 2001.
- [2] 小磯花絵, 前川喜久雄. 『日本語話し言葉コーパス』の概要と書き起こし基準について. 情報処理学会研究報告, 2001-SLP-36-1, 2001.
- [3] 加藤一臣, 南條浩輝, 河原達也. 講演音声認識のための音響・言語モデルの検討. 電子情報通信学会技術研究報告, SP2000-97, NLC2000-49 (2000-SLP-34-23), 2000.
- [4] 篠崎隆宏, 斎藤洋平, 堀智織, 古井貞照. 話し言葉音声の認識を目指して. 電子情報通信学会技術研究報告, SP2000-96, NLC2000-48 (2000-SLP-34-22), 2000.
- [5] 河原達也, 加藤一臣, 南條浩輝, 李見伸. 話し言葉音声認識のための言語モデルとデコーダの改善. 情報処理学会研究報告, 2001-SLP-36-3, 2001.
- [6] 山本一公, 岩井直美, 中川聖一. 発話スタイルの違いが音声認識に及ぼす影響についての検討. 電子情報通信学会技術研究報告, SP99-31, 1999.
- [7] Thilo Pfau and Guenther Ruske. Creating Hidden Markov Models for Fast Speech. In *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, Vol. 2, 1998.
- [8] J.Zheng, H.Franco, and F.Weng. Word-level rate of speech modeling using rate-specific phones and pronunciations. In *Proc. ICASSP*, pp. 1775-1778, 2000.
- [9] C.Fugen and I.Rogina. Integrating dynamic speech modalities into context decision trees. In *Proc. IEEE-ICASSP*, pp. 1277-1280, 2000.
- [10] J.Nedel and R.Stern. Duration normalization for improved recognition of spontaneous and read speech via missing feature methods. In *Proc. ICASSP*, Vol. 1, pp. 313-316, 2001.
- [11] 南條浩輝, 河原達也. 講演音声認識のための話速別モデル化の検討. 日本音響学会研究発表会講演論文集, 1-3-18, 春季 2001.
- [12] 河原達也, 李見伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克巨, 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)の性能評価. 情報処理学会研究報告, 2000-SLP-31-2, NL2000-137-7, 2000.
- [13] 奥田浩三, 中嶋秀治, 河原達也, 中村哲. 講演音声の音響的特徴分析と音響モデル構築方法の検討. 情報処理学会研究報告, 2001-SLP-37-13, 2001.
- [14] C.J.Leggetter and P.C.Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, Vol. 9, No. 2, pp. 171-185, 1995.