音声言語情報処理 37−13
（２００１．７．１４）

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

奥田 浩三 † 中嶋 秀治 † 河原 達也 †‡ 中村 哲 †

† ATR 音声言語通信研究所
〒 619-0288 京都府相楽郡精華町光台 2-2-2

‡ 京都大学大学院 情報学研究科
〒 606-8501 京都府京都市左京区吉田本町

E-Mail:{kokuda,nakajima,tkawahar,nakamura}@slt.atr.co.jp

あらまし 自由発話の一つである講演音声の発話様式に頑健な音響モデルの構築方法について報告する．本稿ではまず，同一話者の講演音声，読み上げ音声を比較し，発話様式の違いによる音響的特徴の違いが顕著であることを明らかにする．また周波数的特徴だけでなく，発話速度の違いも認識率に大きく影響していることを示す．そこで発話速度の変動への対処として，(1) 音響モデルにおける発話速度の変動のモデル化，(2) 発話速度にあわせた分析窓長，フレーム周期の最適化を提案する．(1) は異なる継続時間長の音素セグメントで複数の音響モデルを構築し，同一の音素環境に割り当てることで一つの音素モデルに統合するものである．これにより，継続時間長の短い音素の出現位置を事前に知ることなく，音素継続時間長の変動を吸収することができる．また，(2) は分析窓長，フレーム周期の最適化により音響モデル学習データとの間の音素継続時間長分布のずれを修正するとともに，周波数分析における分析精度を改善するものである．これらの手法を併用することで，単語誤り率を 22.2 %改善することができ，さらに教師なし適応との相乗効果も確認できた．

Characteristics and Acoustic Modeling of Lecture Style Speech

Kozo OKUDA† Hideharu NAKAJIMA† Tatsuya KAWAHARA†‡ Satoshi NAKAMURA†

† ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 JAPAN

‡ School of Informatics, Kyoto University
Sakyo-ku Kyoto 606-8501 Japan

E-Mail:{kokuda,nakajima,tkawahar,nakamura}@slt.atr.co.jp

Abstract We propose an acoustic modeling method robust to the lecture style speech. We compare lecture speech with read speech spoken by the same speaker and make clear that lecture speech is different from read speech not only in spectral characteristics but also in the speaking rate. This difference of speaking rate degrades the performance of speech recognition system. To solve this problem, we propose (1)acoustic modeling method robust to the change of speaking rate, and (2)modifying the frame length and frame period. In the first method, we create a dedicated model for short phone segments, which is combined with the conventional acoustic model to compose a single acoustic model in a multi-path fashion. This model improves the recognition performance without prior knowledge of the position of fast phones. In the second method, we change the frame length and frame period to compensate the mismatch of duration distribution between training data and lecture speech data. By combining these methods, we obtain 22.2% improvement in word error rate.

1 はじめに

統計的手法の導入や大規模音声データベースの構築により，音声認識技術の認識性能は飛躍的に向上している．しかしながら統計的手法を用いた音声認識では，学習データと認識対象データが統計的に同分布であると仮定しているため，これらの分布が異なる場合，認識性能は大きく劣化する．現状の音声認識システムの多くは，読み上げ音声や擬似対話音声を用いて音響モデルを構築しているため，同時通訳や講演音声の書き起こし・要約技術などへ音声認識技術の応用を考えた場合，認識性能が大きく劣化する可能性がある．

一方，話し言葉音声の分析と認識を目指し，「話し言葉工学」プロジェクト [1] がスタートしており，大規模話し言葉コーパスの構築が進んでいる．このプロジェクトより配布されているデータの講演音声は対話音声同様，音声認識を前提としない自然な発話音声であるが，その発話様式は擬似対話音声とは大きく異なる．これらのデータを用い音声認識システムを構築することで，講演音声の発話様式に頑健な音響モデルの構築が可能であると考えられる．

講演音声においても，講演の内容や講演者の話者性・経験などにより，発話様式は多岐に渡る [2]．このため，プロジェクトのデータも話者や発話様式ごとに分析しなければ，十分な認識性能が得られない．このような背景のもと本報告では，プロジェクトのデータの一つである，話者 ID AS99SEP022 の音声データ（以降，AS002 と表記する）を対象に，音響的特徴を分析することで，講演音声に頑健な音響モデルの構築方法を考察する．

本報告ではまず，**2**章に本報告で用いた実験環境の概要を述べると共に，ベースとなる音響モデルによる単語認識実験結果をまとめる．次に **3**章で講演音声の音響的特徴について分析を行い，最後に **4**章で講演音声に頑健な音響モデルの構築方法について検討する．

2　ベースラインシステム

2.1　認識システム

本報告における認識実験では，当研究所で開発した連続音声認識エンジン，ATRSPREC[3] を用いた．音響特徴パラメータは，サンプリングレート 16kHz，プリエンファシス 0.98，分析窓長 20 msec，フレーム周期 10 msec で抽出した 25 次元の特徴ベクトル（12 次メルケプストラム，12 次 Δ メルケプストラムと Δ 対数パワー）を用いた．また，デコードの際，HMM の状態スキップは可能とした．

2.2　ベースライン音響モデル

ベースラインとなる音響モデル（ベースライン音響モデル）は性別依存モデルであり，各音素モデルは 3 状態，5 混合ガウス分布，総状態数 1,400 の状態共有化 HMM（HMnet[4]）で表現されている．音響モデルの学習データには，当研究所で収集した会議予約タスクデータベース [5] の，男性話者 1,321 人の擬似対話データ（約 19.3 時間）を用いた．

2.3　言語モデル

単語認識実験を行う場合，音響的要因と言語的要因の両方が含まれた認識結果が得られることとなる．しかしながら講演音声の発話様式に頑健な音響モデルを検討するためには，言語的要因による認識率の劣化を極力排除することが望ましい．本報告ではテストデータ AS022 を対象に音響的な分析を行うため，言語モデルには AS022 の発話内容のみで構築した単語バイグラムを用いることとした．

認識辞書に関しては，京都大学で作成されたものを用いるが，未知語による影響をなくすため，言い誤り以外の，AS022 の発話内容に出現する全ての単語を追加登録した．

2.4　講演書き起こしの読み上げ音声

本実験環境では言語モデルがクローズドモデルとなるため，他のテストデータとの比較が困難である．そこで本稿では言語的な要因を排除し音響的な比較を行うため，テストデータの発話内容の読み上げ音声を話者 AS022 から収録した．読み上げデータは，AS022 の講演内容（28 分，6305 単語）の内の前半（3272 単語）を使用し，間投詞，言い誤りもそのまま読み上げた．読み上げ単位は，プロジェクトより配布された書き起こしデータ内に記述されている時間情報を元に，人手で文らしくつなげた単位としている．本稿で述べる評価実験では読み上げ音声，講演音声共に，AS022 の講演内容の前半のみを用いる．

2.5　単語認識実験

ベースライン音響モデルを用いた単語認識実験の結果を表 1に示す．この結果から，講演音声より読み上げ音声の方が良好な認識率が得られていることがわかる．これは講演音声と比較し，読み上げ音声と音響モデルとのミスマッチが小さいためと考えられる．講演音声は自由発話音声であるが，音響モデルの学習データである擬似対話音声と比較するとその音響的特徴が大きく異なっていると思われる．

次に，話者適応を用いた場合の単語認識実験を行った．読み上げ音声，講演音声それぞれに対して，全ての発話データを用いたクローズドデータでの教師

表 1: ベースライン音響モデル及び，クローズド教師あり話者適応を用いた単語認識率

データ	ベースライン音響モデル	教師あり適応モデル (読み上げ)	教師あり適応モデル (講演)
読み上げ	81.9%	91.8%	59.8%
講演	74.3%	76.6%	81.2%
※言語モデルにはクローズドモデルを使用			

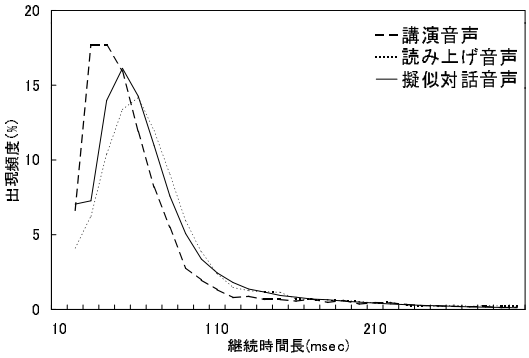


図 1: 各音声データの音素継続時間長分布

表 2: 講演音声，読み上げ音声，及び擬似対話音声における平均音素継続時間長 (msec)

データ	母音	子音	全体
講演音声	65.7	54.3	60.3
読み上げ音声	87.6	73.0	80.7
擬似対話音声	79.2	60.2	70.4

あり適応を，MAP-VFS[6] により行った．適応は平均値と状態遷移確率の両方に対して行っている．話者適応を用いた単語認識実験の結果を表 1に示す．

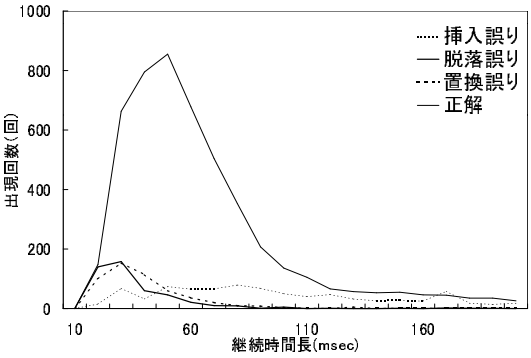
話者適応を行った場合，読み上げ音声で 91.8% の認識率が得られている．これに対し講演音声は，話者適応を行った場合でも認識率が 81.2% であり，読み上げ音声と比較すると劣化している．この結果より，音響モデルが良好に働いた場合，本実験環境においてはその認識率の上限が 91% 付近であると考えることができ，講演音声は話者適応だけでは十分な性能が得られていない．読み上げ音声で適応した音響モデルによる講演音声の認識，講演音声で適応した音響モデルによる読み上げ音声の認識においても，認識率の改善はほとんど見られない．話者性を適応するだけでは，十分な認識性能は得られないということであり，講演音声には発話様式の違いによる認識率劣化の要因が大きいと考えられる．

3 講演音声の音響的特徴の分析

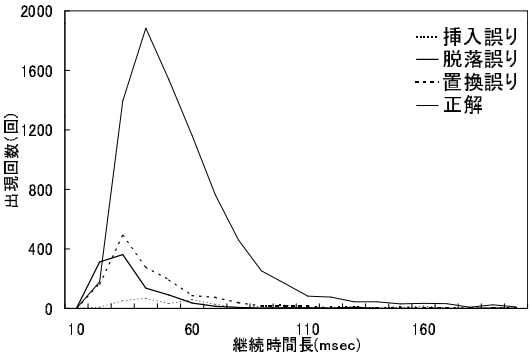
本章では，AS022 の講演音声の音響的特徴を分析することで，話者性の適応だけではカバーできない要因について調査する．

3.1 音素継続時間長における特徴

講演音声データは，読み上げ音声や擬似対話音声と比較し，発声速度が速くなる傾向がある．そこで，講演音声の音素継続時間長が，読み上げ音声や音響モデル学習データである擬似対話音声と比較し，どの程度異なっているかを分析した．分析は，ビタビ・アライメントによりそれぞれの音素区間を判定し，継続時間を算出している．音響モデルとのミスマッチによるアライメントのずれを極力減らすため，2.5 節で用いた話者適応モデルを使用した．



— 母音 —



— 子音 —

図 2: 音素継続時間長ごとの誤り分布

図 1に講演音声，読み上げ音声，擬似対話音声それぞれに含まれる各音素の継続時間長分布を示す．この図より講演音声は，音素継続時間長のピークが 30 〜 40 msec 付近に存在するとともに，擬似対話音声の分布から左（継続時間長が短くなる方向）にずれていることがわかる．読み上げ音声，擬似対話音声の音素継続時間長のピークはそれぞれ 60 msec，50 msec とずれているが，講演音声と比較すると，継続時間長分布は比較的近いと言える．

表 2に講演音声，読み上げ音声，擬似対話音声それぞれの音素継続時間長の平均値を示す．講演音声は，読み上げ音声や擬似対話音声と比較し，各音素の平均継続時間長が短くなっていることがわかる．

次に音素認識実験を通して，音素継続時間長と音素誤りの関係を調査した．ここでも，2.5 節で用いた話者適応モデルを使用している．

図 2に講演音声において誤りが生じた音素の，継

続時間長ごとの分布を示す．この図から，脱落・置換誤りは音素継続時間長の短い領域で多く発生していることがわかる．挿入誤りについては，母音に関しては全体的に生じているが，子音に関しては継続時間長の短い領域で多く発生している．

継続時間長の短い領域に誤りが多く発生する原因としては，音響モデル学習データである擬似対話音声の音素継続時間長分布が，講演音声の分布とずれており，継続時間長の短い音素に対する学習データが不足していたためと考えられる．また，音響モデルが 3 状態の left-to-right モデルであり，フレームシフトが 10 msec という制約上，音素継続時間長が 30 msec 未満の音素に対しては，モデルの構造そのものがミスマッチを生じていると考えられる．さらに，分析窓長が 20 msec のため，継続時間長が 20 msec 程度の音素に対しては，周波数分析の時点で分析精度が劣化している可能性がある．

3.2 周波数領域における特徴

ここでは周波数領域において，講演音声が読み上げ音声や擬似対話音声と，どの程度異なっているかを分析した．分析は，AS022 の読み上げ音声，擬似対話音声との比較により行った．分析において，F1, F2 は各母音セグメント（ビタビ・アライメントにより判定）の中央点において算出した．

図 3に講演音声，読み上げ音声，擬似対話音声における母音の F1-F2 の分布を示す．F1-F2 の分布は同一話者にも関わらず，講演音声と読み上げ音声の間で変動しており，また音響モデル学習データである擬似対話音声とも大きく異なっていることがわかる．2.5 節の認識実験において，講演音声による話者適応モデルで，読み上げ音声の認識率が大きく劣化している要因も，これらの周波数的特徴の変動によるものと考えられる．この変動は，発話様式の違いによって生じており，話者性だけではなく発話様式への適応も必要である．

4 講演音声認識のための音響モデル

前章までの結果より，AS022 の講演音声と認識システムの間で，(1) 発話速度が大きく異なるため，音素継続時間長の分布が学習データとずれていると共に，継続時間長の短い音素が多く出現する，(2) 周波数領域における特徴空間において，音響モデルの学習データとのずれが存在するといったことが明らかになった．これらを踏まえ，AS022 を通して講演音声に頑健な音響モデルの構築方法を検討する．

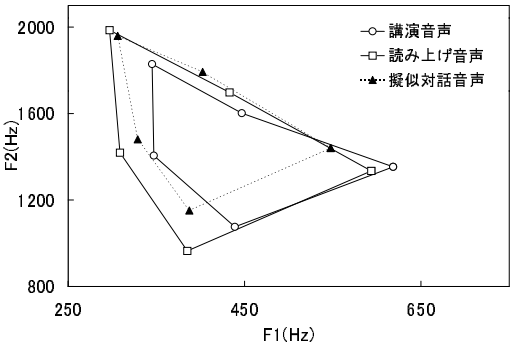


図 3: 各音声データの F1-F2 分布

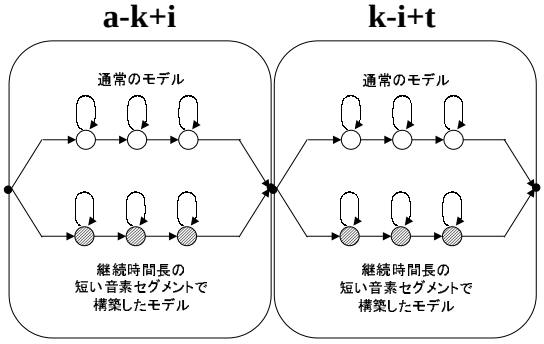


図 4: 複数のモデルを統合した音響モデル

4.1 音響モデルによる発話速度の変動のモデル化

3.2 節の結果より，講演音声では継続時間長が短い音素において，脱落・置換誤りが多く発生していることがわかった．これは学習データにおいて，継続時間長の短い音素データが量的には十分出現しているが，全体的には出現頻度が低いため，統計的アプローチで構築する HMM では，その領域の認識性能が劣化したと考えられる．そこで，継続時間長の短い音素セグメントのみを用いた音響モデルの構築を検討する．認識の際，継続時間長の短い音素の出現位置はわからないため，ベースライン音響モデルと音素継続時間長の短い音響モデルを一つの音響モデルに統合する [7]．図 4に示すように，一つの音素環境に対して 2 つのモデルを定義し，デコードの際それぞれのモデルに対して仮説を展開するものとした．デコードの結果，尤度の高い経路が選択されるため，継続時間長の短い音素の出現位置を事前に知る必要がなく，また認識辞書の変更も不要となる．

図 2より，脱落誤り，置換誤りの分布のピークが 30 msec に存在するため，継続時間長が 40 msec 以下の音素セグメントを用いて音響モデルを構築した．

表 3: 継続時間長の短い音素に対応した音響モデルによる単語認識率

	認識率	挿入誤り	脱落誤り	置換誤り
ベースライン音響モデル	74.3%	3.2%	8.1%	14.4%
音響モデルS	77.4%	3.7%	5.9%	13.0%

表 4: 分析窓長，フレーム周期を変更した場合の単語認識率

分析窓長	フレーム周期	認識率
20msec	10msec	74.3%
20msec	9msec	77.1%
	8msec	76.2%
	7msec	76.2%
	6msec	75.0%
18msec	10msec	74.9%
16msec		75.3%
14msec		75.0%
12msec		73.9%
18msec	9msec	75.9%
16msec	8msec	77.3%
14msec	7msec	75.9%
12msec	6msec	76.7%

構築した音響モデルは，総状態数 1400，5 混合ガウス分布とした．このモデルとベースライン音響モデルを統合したモデルを，音響モデルSと呼ぶこととする．

表 3に，音響モデルSによる単語認識率を示す．継続時間長の短い音素セグメントで構築したモデルを統合することで，認識率が 3.1% 向上した．提案手法では挿入誤りは増加するものの，脱落，置換誤りが減少していることから，継続時間長の短い音素セグメントで構築したモデルが効果的に働いていることがわかる．

4.2 発話速度に合わせた分析窓長とフレーム周期の最適化

講演音声の発話速度を正規化し，学習データとの音素継続時間長分布を近づけると共に，前処理やモデルの構造に適した発話速度にすることで，認識性能の改善が期待できる．発話速度の正規化に関しては，発話速度に応じてフレーム周期を変更する方法 [8] や，特徴パラメータの間引き，相関による補間を用いた方法 [9] が提案されている．しかしながら前者は発話速度が遅い発声に対しては効果が得られているが，発話速度の速い発声に対しては逆に劣化するという結果が得られている．また後者の手法は，発話速度の速い発声に対しても効果が得られているが，講演音声のように分析窓長より短い音素が多く出現する場合，前処理における周波数分析の精度が低くなるため，十分な性能が得られない可能性がある．前者の手法で，発話速度の速い発声に対して認識率が劣化する原因としても，周波数分析精度の劣化が影響している可能性がある．

表 5: 音響モデルSと分析窓長・フレーム周期変更の併用による単語認識率

分析窓長	フレーム周期	認識率
20msec	10msec	77.4%
18msec	9msec	79.2%
16msec	8msec	80.0%
14msec	7msec	79.0%
12msec	6msec	77.5%

そこで本稿では，分析窓長とフレーム周期を合わせて変更する．発話速度が速い場合，フレーム周期を短くすることで時間方向の分解能は向上する．しかしながら継続時間長の短い音素に対しては，分析窓長が長すぎる場合，一つの音素内の周波数の変化に対する周波数分析精度が劣化すると考えられる．このため，フレーム周期と合わせて分析窓長も短くする必要がある．表 4に，分析窓長，フレーム周期をそれぞれ変更した場合の単語認識率を示す．この結果より，分析窓長 16msec，フレーム周期 8msec としたものが最も良い結果となっている．これは，ベースラインシステムにおける分析窓長 20msec，フレーム周期 10msec と比較するとそれぞれ 1.25 倍となっており，学習データと講演音声データの音素継続時間長平均値の比率が 1.17 (= 70.4 msec/60.3 msec) であることから，全体的にマッチしたと考えられる．分析窓長 18msec，フレーム周期 9msec （ベースラインシステムと比較すると 1.1 倍）より認識率が向上しているのは，継続時間長が短い音素に対する周波数分析精度が向上したためと考えられる．

次に，音響モデルSと時間伸長を併用した場合の認識実験を行った．表 5に結果を示す．さらに認識率が向上し，ベースライン音響モデルと比較し認識率で 5.7%，認識誤り率で 22.2% 改善している．

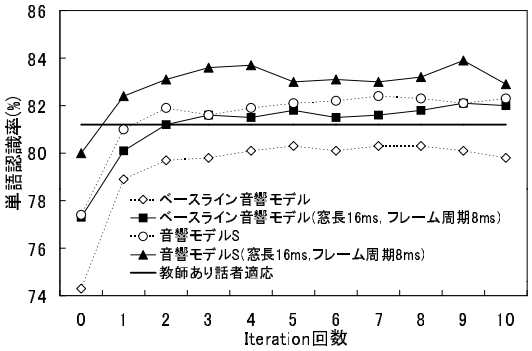


図 5: 教師なし話者適応による認識率

4.3 教師なし話者適応の導入

第３章において，学習データと講演音声の間で，F1-F2 の分布が異なっていることを述べた．この結果を踏まえ，教師なし話者適応を用いた評価実験を行った．教師なし話者適応は，講演音声を認識後，その認識結果を正解発話内容として，全てのデータを用いた MAP-VFS による適応を行うものとした．評価実験は，分析窓長・フレーム周期を変更しないものと，それぞれ 16msec， 8msec に変更したものを用い，ベースライン音響モデル，音響モデルＳそれぞれに対して行った．

結果を図 5に示す．比較として， 2.5 節で得られた教師あり話者適応の結果も示す．分析窓長・フレーム周期を変更し音響モデルＳを用いたものが最も良く，教師あり話者適応よりも良好な結果となっている．これらの結果より講演音声を認識するには，発話速度を考慮した前処理，音響モデルの構築が有効であり，話者適応によりさらに認識率が改善されることが示された．

5 まとめ

「話し言葉工学」プロジェクトのテストデータ，AS022 に対する音響的特徴分析を行い，講演音声に頑健な音響モデル構築方法について検討した．

AS022 の講演音声は学習データとの間で，音素継続時間長の分布にずれが存在していることがわかった．また継続時間長の短い音素が多く出現することも明らかになった．これに対し，継続時間長の短い音素セグメントのみを用いた音響モデルの構築，発話速度に合わせた分析窓長・フレーム周期の変更を提案し，認識誤り率を 22.2% 改善することができた．講演音声における認識率の劣化が発話速度の違いにより生じていることから，発話速度の正規化は，自然発話音声を認識するためには重要であると考えられる．発音の変形や怠けなどの対応も，発話速度を正規化した上で検討する必要がある．

本報告では，分析窓長・フレーム周期は固定値で行ったが，今後は動的に値を変更することでの発話速度の正規化の検討を進める．また今回，言語モデルに関してはクローズドモデルとして評価を行ったが，連続音声認識における最終的な認識性能は，音響的要因，言語的要因の両方が影響するため，オープンな言語モデルを用いた評価についても，今後進める予定である．

参考文献

[1] 古井貞熙，前川喜久雄，井佐原均．”『話し言葉工学』プロジェクトの概要と展望，”話し言葉の科学と工学ワークショップ,pp1-6,2001.

[2] 奥田浩三，中嶋秀治，松井知子，河原達也，中村 哲．”講演音声認識のための音響モデル構築方法の検討，”話し言葉の科学と工学ワークショップ,pp109-116,2001.

[3] 内藤正樹，山本博史，シンガー ハラルド，中嶋秀治，中村 篤，匂坂芳典，”対話音声を対象とした連続音声認識システムの試作と評価,” 信学論 (D-II),J84-D-II,No.1,pp31-40,2001-1.

[4] 鷹見淳一，嵯峨山茂樹，”逐次状態分割法による隠れマルコフ網の自動生成,” 信学論 (D-II),J79-D-II,10 pp.2155-2164,1993-10.

[5] 松井知子，内藤正樹，ハラルド シンガー，匂坂芳典．”大規模な日本語音声データによる音響モデルの分析,” 日本音響学会研究発表会講演論文集,1-Q-28,2000-3.

[6] 大倉計美，杉山雅英，嵯峨山茂樹．“混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式,” 信学技報, SP92-16,1992-06.

[7] 奥田浩三，松井知子，中村 哲．”音節強調発声に頑健な自然発話音声の認識法,” 信学技報,SP2000-98,Vol.100,No.523,pp19-24,2000.

[8] S.Tsuge, T.Fukada and K.Kita, “Frame-period adaptation for speaking rate robust speech recognition,” Proc.ICSLP2000,Vol.3,pp718-721,2000.

[9] Jon P.Nedel and Richard M.Stern, “Duration normalization for improved recognition of spontaneous and read speech via missing feature methods,” Proc.ICASSP2001,Vol.1,2001.