

講演音声認識における発話速度の変動を考慮した音声認識手法

奥田 浩三[†] 河原 達也^{†,††} 中村 哲[†]

[†] ATR 音声言語コミュニケーション研究所
〒 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,nakamura}@slt.atr.co.jp, ^{††}kawahara@kuis.kyoto-u.ac.jp

あらまし 講演音声を対象とした音声認識を行う場合、発話速度の速い発声の認識率が劣化するという問題が生じる。この原因としては、発声のなまけなどの周波数領域における音響の特徴の変形が考えられるが、同時に時間領域における変形も生じていると考えられ、発話速度の正規化や補正が重要となる。本稿では、尤度基準により発話速度に応じた分析周期・分析窓長を自動選択することで、発話速度を補正する手法を提案する。本手法は分析周期・窓長を変更することで、発話速度の補正の効果をj得るものであるが、最適な分析周期・窓長は発話毎に異なると考えられる。そこで、複数の分析周期・窓長により認識した後、分析周期により正規化した音響尤度と言語尤度を用いて最も尤度が高くなる分析周期・窓長を選択する。「話し言葉工学」プロジェクトより配布されているモニターセットを用いた評価実験において、提案手法の有効性を確認した。

キーワード 音声認識, 講演音声, 発話速度, 分析周期, 分析窓長

Lecture speech recognition considering the speaking rate variation

Kozo OKUDA[†], Tatsuya KAWAHARA^{†,††}, and 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,nakamura}@slt.atr.co.jp, ^{††}kawahara@kuis.kyoto-u.ac.jp

Abstract In a lecture speech recognition, performance of speech recognition system degrades when a speaking rate is increased. The reason of this degradation is a change of acoustic characteristics not only in frequency domain but also in time domain. Because of these changes, normalization or compensation of the speaking rate is important. In this paper, we propose a speaking rate compensation method which selects an optimal frame period and frame length using a likelihood criterion. This method changes the frame period and length to compensate the speaking rate. However, the optimal frame period and length are different in each utterance. Therefore, our proposed method conducts speech recognition with various frame periods and lengths and determines the optimal frame period and length for the target speech using the acoustic likelihood normalized by the frame period and language likelihood. In a recognition experiment using CSJ corpus, this method improves the performance for high speaking rate speech.

Key words automatic speech recognition, lecture speech, speaking rate, frame period, frame length

1. はじめに

講演音声をはじめとする話し言葉の音声認識では、読み上げ音声などと比較し認識性能が劣化する傾向にある。この劣化の要因としては、言語的な複雑さや音響的な特徴の違いなどが挙げられるが、発話速度の変動も大きく影響していると考えられる[1]。講演音声においては、講演者のスキルや講演内容、講演のスタイル（原稿を読み上げた場合や暗記して講演する場合、原稿を準備をしない場合など）により、発話速度にばらつきが多く見られる。特に原稿を暗記して講演する場合や準備をしない場合などは、発話内に多くの間投詞や言い誤りが出現するとともに、発話速度のばらつきが大きくなると考えられる。また、同一講演内においても講演の前半と後半では発話速度に差が生じる傾向が観測されている[2]。

発話速度と認識性能の関係としては、発話速度の速い音声においては脱落誤りや置換誤りが多く発生し、発話速度の遅い発声においては挿入誤りが多く発生すると報告されている[3]。特に、発話速度の速い音声においては、なまけなどの影響による認識性能の劣化に加え、音響モデルの構造そのものが対応できない場合が考えられる。例えば、分析周期 10msec で構築した 3state left-to-right 型 HMM で構成された tri-phone 音響モデルの場合、状態スキップを許さなければ、継続時間長が 30msec 未満の音素に対しては十分な性能が得られない。話者の違いや講演のスタイルの違いによる音響的な特徴空間のずれや発声のなまけなどについては、話者適応により認識性能を改善することも可能と考えられるが、MAP や MLLR に代表される話者適応アルゴリズムは、音響モデルの構造については適応することができないため、問題となる。言い換えれば、発話速度を正規化や補正することで、話者適応なども効果的に働くと考えられる。

本稿ではこの問題を解決する手法として、音響分析における分析周期・分析窓長を変更することにより、発話速度を正規化する手法を提案する。認識対象音声の発話速度は事前にはわからない。そこで本手法では、入力音声に対して複数の分析周期・窓長で抽出した特徴ベクトル

に対してそれぞれ認識を行い、尤度基準で分析周期・窓長を選択するものである。本稿では「話し言葉工学」プロジェクト[4]より配布されたモニターセットを用い、提案手法の有効性について検証する。

本稿の構成は次の通りである。まず2章にて実験条件についてまとめる。次に3章にて評価セットにおける各話者の発話速度と認識率の関係について述べ、4章において分析周期・窓長を変更した場合の効果について述べる。この結果を踏まえ、5章にて尤度基準による分析周期・窓長の自動選択手法を提案し、その有効性を確認する。

2. 実験条件

2.1 評価セット

本報告で用いた評価セットには、プロジェクトより配布されたモニターセットから、男性話者7名を選択した。それぞれ表1に示す通りとなっている。評価実験の際には、配布データに含まれている書き起こしデータに記述されている時刻情報を元に、文章を分割している。

2.2 ベースラインシステム

本報告にて用いた、ベースラインとなる認識システムの概要は以下の通りである。

音響特徴パラメータには、16kHz サンプリング、分析周期 10msec、分析窓長 20msec 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、1 次 Δ 対数パワー）を用いている。

音響モデルは、状態共有化 HMM（HMnet [5]）により構築された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 1400 で表現されている。なお、本報告で用いた評価話者は全て男性話者のため、モデル

表1 評価セット

話者 ID	中心となるテーマ	講演時間	略称
A01M0007	音声処理、聴覚関係	30 分	M0007
A01M0035	音声処理、聴覚関係	28 分	M0035
A01M0074	音声処理、聴覚関係	12 分	M0074
A02M0117	日本語学関係	57 分	M0117
A03M0100	自然言語処理関係	15 分	M0100
A05M0031	音声学関係	27 分	M0031
A06M0134	社会言語学関係	23 分	M0134

は男性モデルのみを用いている。音響モデルの学習データには、モニターセットのうち評価話者7名を除く全ての学会講演、模擬講演男性話者データ200名（約34時間）のデータを用いた。なお音響モデルの学習は、配布された書き起こしデータに記述されている時刻情報、カタカナ書き起こし情報を元に行っている。

言語モデルは京都大学で作成され、モニターセットと共に配布された講演音声用言語モデル（2001-06；京都大学）の前向き単語バイグラム、後ろ向き単語トライグラムを用いている。認識辞書に関しても同様で19k単語のものを用いている。デコーダにはJulius3.2を用いている。

2.3 評価セットに対する認識性能

本認識システムを用い、評価セットの認識実験を行った。認識結果を表2にまとめる。話者により単語誤り率のばらつきが大きく、平均で35.8%の単語誤り率となっている。以降、この認識結果をベースラインとする。

3. 講演音声における発話速度と認識性能の関係

講演音声においては、話者の違いや講演内容、講演のスタイルにより発話速度が大きく変化する。図1に、評価話者7名の発話速度と単語誤り率との関係を示す。図中における発話速度は、文章毎に算出した1秒あたりのモーラ数の平均を示している。各文章の1秒あたりのモーラ数は、文章毎にビタビライメントにより算出した音声区間の時間長で文章中のモーラ数を割った値となっている。図1より、発話速度が比較的速い話者は単語誤り率が増加する傾向にあることがわかる。

次に、継続時間長の短い音素の単語認識率への影響を調査するため、話者毎に継続時間長が30msec以下の音素の出現頻度を算出した。各音素の継続時間長は、前述のビタビライメントの結果より抽出した。図2に、30msec以下の音素の出現頻度と単語誤り率の関係を示す。この図より、継続時間長の短い音素が多く出現する話者は、単語誤り率が増加する傾向にあることがわかる。

表2 ベースラインシステムによる認識実験結果

話者	単語誤り率 (%)
M0007	29.5
M0035	42.8
M0074	30.0
M0117	30.9
M0100	37.2
M0031	39.3
M0134	42.7
平均	35.8

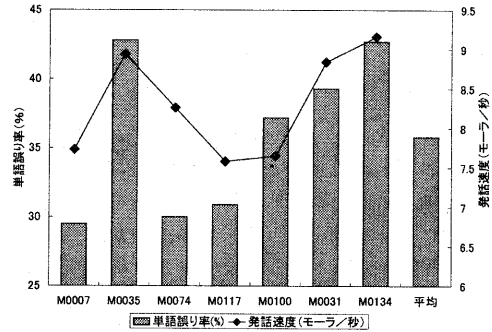


図1 発話速度と単語誤り率との関係

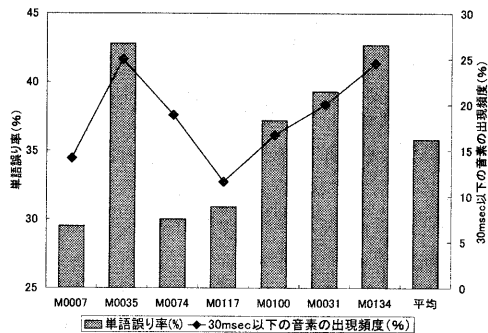


図2 30msec以下の音素の出現頻度と単語誤り率との関係

4. 分析周期・窓長の変更による発話速度の変更

発話速度の正規化や補正により、発話速度に起因する認識性能の劣化や、モデル構造とのずれを軽減することで、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[6]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[7]などが提案されている。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、後者の手法は発声内容が既知の場合には効果が得られるが、発声内容が未知の場合には十分な効果が得られていない。

そこで本稿では、音響分析における分析周期・窓長の両方を変更することで、相対的に発話速度を変更した場合と同じ効果を得る[8]ことを試みる。

ベースラインシステムを用い、評価セットに対して分析周期・窓長を変更した場合の効果について調査する。ベースラインシステムにおける分析周期(10msec)、窓長(20msec)に対して、分析周期・窓長をそれぞれ9msec・18msecと8msec・16msecに変更した場合について、認識実験を行った。図3に単語誤り率を示す。この結果より、発話速度の比較的速い話者は、分析周期が8msecから9msecの場合に単語誤り率が小さく、発話速度の比較的遅い話者は、分析周期が9msecから10msecの場合に単語誤り率が小さくなっていることがわかる。

5. 尤度基準による分析周期・窓長の自動選択

前章の結果より、発話速度に応じて分析周期・窓長を変更することで、発話速度の補正の効果があることが確認できた。しかしながら、発話速度は話者の違いだけではなく、同一話者においても発話毎や発話内で変動していると考えられる。本章では、発話毎に最適な分析周期・窓長を、尤度基準により選択する手法について提案する。

5.1 音響尤度を用いた選択手法

各発話における発話速度は事前にはわからない。そこで、それぞれの分析周期・窓長で認識した結果における音響尤度を比較することにより、分析周期・窓長を選択することを行った。分析周期を短くした場合、一発話当たり出力されるフレーム数が多くなるため、同一内容の認識結果を出力した場合でも、音響尤度は小さくなる(音響尤度の絶対値が大きくなる)。このため、分析周期により音響尤度を正規化が必要がある。音響尤度の正規化の方法として、次式より算出するものとした。

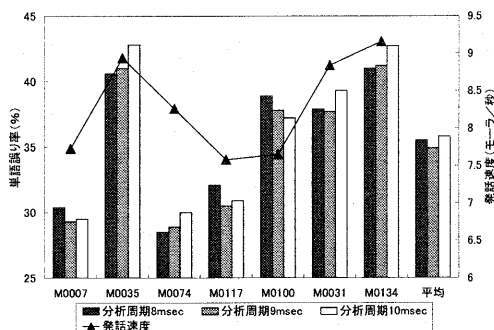


図3 分析周期・分析窓長と単語誤り率

$$AM' = AM * \frac{\text{フレーム周期(msec)}}{10}$$

AM: 発話毎の音響尤度

AM': 分析周期により正規化した音響尤度

発話毎に、上式より算出した音響尤度AM'を比較し、最も音響尤度の高い分析周期の認識結果を選択するものとした。本手法による認識率を図4に示す。

正規化した音響尤度を基準とすることで、平均単語誤り率はベースラインとなる分析周期10msec、窓長20msecの場合と比較し、1.0%改善した。

5.2 音響尤度・言語尤度を用いた選択手法

認識結果の音響尤度は、言語モデルによる制約を受けた状態での音響尤度となる。このため、デコード時の探索過程において言語尤度が大きくなる仮説が選択された結果、音響尤度単体で見えた場合、劣化している場合がある。そこで、分析周期で正規化した音響尤度に言語尤度(挿入ペナルティを含む)を加えたスコアを算出し、最もスコアの高い分析周期の認識結果を選択するものとした。本手法による単語誤り率を図5に示す。

この結果より、正規化した音響尤度より再計算したスコアを用いた場合、ベースラインと比較し1.5%の認識率の改善が見られた。また、音響尤度を用いて選択した場合と比較しても、認識率の改善が見られる。

6. 提案手法と話者適応の併用

話者適応を用いた場合、発話速度の影響によ

る認識性能の劣化がある程度改善され则认为られる。このため話者適応と併用した場合、提案手法による効果が得られない可能性もある。本章では、話者適応と併用した場合の提案手法の効果について報告する。

評価セットに対して、(1)分析周期・窓長を(8msec・16msec)、(9msec・18msec)、(10msec・20msec)のそれぞれで固定して教師あり話者適応を行った場合、(2)上記の分析周期・窓長の認識結果に対してスコアによる提案手法を用いた場合、(3)スコアを用いた提案手法により選択した分析周期・窓長により教師あり話者適応を行った音響モデルを用い、スコアによる提案手法を用いた場合、の3つの条件にて認識実験を行った。条件(2)は条件(1)の結果に対して提案手法を用いるため、各分析周期・窓長毎に話者適応したモデル(3種類)を切り替えた結果となるが、条件(3)では各分析周期・窓長で同一の音響モデルを用いている。適応アルゴリズムにはMAP-VFS[9]を用い、平均値、状態遷移確率の両方を適応している。また適応においては各話者の全ての音声を用い、話者毎に適応を行っている。教師あり話者適応を行った場合の単語誤り率を図6に示す。

この結果より、ベースラインとなる分析周期10msec、窓長20msec固定の場合と比較し、分析周期・窓長を(8msec・16msec)(9msec・18msec)固定で話者適応を行った場合の方が単語誤り率が小さくなっており、これらの認識結果に対して提案手法を用いることで、さらに単語誤り率が小さくなっていることがわかる。スコアを用いた提案手法により選択された分析周期・窓長により話者適応を行った音響モデルを用いた場合も、単語誤り率は小さくなっているが、分析周期・窓長を固定して話者適応を行った結果に対して提案手法を用いた方が良好な結果となっている。

7. 考察

提案手法ではベースラインと比較し、平均認識率で1.5%の改善となっているが、発話速度の速い話者M0035、M0031、M0134に関して見た場合、それぞれ3.0%、2.5%、2.5%、平均で2.7%認識率が改善している。このことより提案手法は、発話速度の遅い話者の認識率劣化

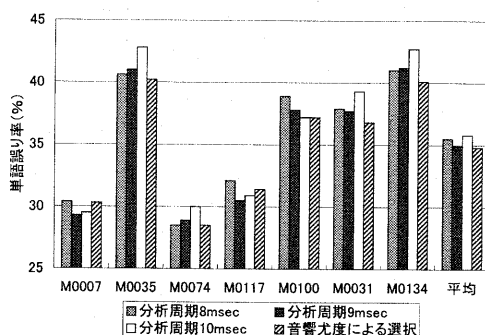


図4 分析周期を考慮した音響尤度を用いた分析周期・窓長の選択

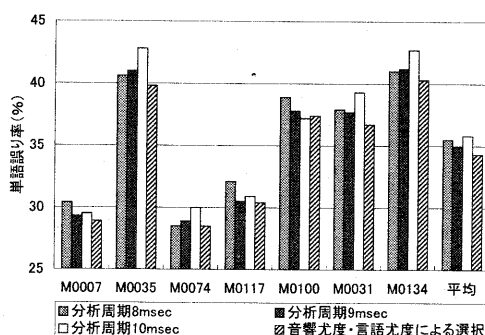


図5 スコアを用いた分析周期・窓長の選択

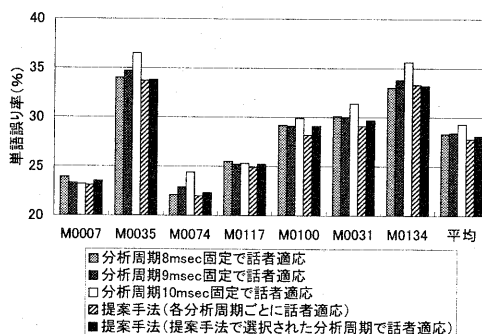


図6 話者適応と併用した場合の提案手法による評価実験

を抑えながら、発話速度の速い話者の認識率を改善しており、発話速度の変動の大きい講演音声に対して有効であると考えられる。

話者毎に最も認識結果の良い分析周期・窓長を選択した場合の単語誤り率を表3に示す。この結果より、話者毎に分析周期・窓長を選択するよりも、発話毎に選択した方がその効果が大きいことがわかる。

表3 話者毎に最適な分析周期・窓長を選択した場合の単語誤り率(%)

	提案手法	話者毎に最適な 分析周期・窓長を選択
M0007	28.9	29.3
M0035	39.8	40.6
M0074	28.5	28.5
M0117	30.4	30.5
M0100	37.4	37.2
M0031	36.7	37.7
M0134	40.3	41.0
平均	34.3	34.7

表4 認識結果より分析周期・窓長を選択した場合の単語誤り率(%)

	提案手法	認識結果より選択
M0007	28.9	24.6
M0035	39.8	36.0
M0074	28.5	25.3
M0117	30.4	26.7
M0100	37.4	32.7
M0031	36.7	33.1
M0134	40.3	36.2
平均	34.3	30.4

提案手法を話者適応と併用した場合にも、同様の効果が確認されており、分析周期・窓長をそれぞれ固定して話者適応を行った音響モデルに対して提案手法を用いた場合に最も良い結果が得られている。しかしながら、認識結果をもとに適応を行う教師なし話者適応の場合は、適応に用いる認識結果がその認識性能に大きく影響すると考えられる。この場合には、提案手法により選択された分析周期・窓長を用いて話者適応を行った音響モデルにおいて、提案手法の効果があまる可能性もある。

表4に、認識結果より発話毎に最も認識結果の良い分析周期・窓長を選択した場合の単語誤り率を示す。この結果は、認識率が最も良いものを人手で選択したものであり、本手法による認識性能の上限であると考えられる。この結果より、提案手法ではまだその上限には達していないことがわかる。

8. ま と め

音響尤度・言語尤度を用い、入力音声の発話速度に適した分析周期・窓長を選択する手法を提案した。本手法を用いることにより、分析周期・窓長をそれぞれ10msec、20msec固定の場合と比較して認識性能が改善し、特に発話速度

の速い話者に関して効果が得られた。また、話者適応と併用した場合にも同様の効果が得られていることから、話者適応などでは改善できない要因に対して提案手法の効果が得られていると考えられる。

しかしながら発話速度は同一発話内でも変動するため、単語や音素、フレーム毎に最適な分析周期・窓長を選択する必要があると考えられる。また音響モデルに関しても、発話速度毎に最適な分析周期・窓長やモデル構造を用いて構築することで、更なる改善が期待できる。

本提案手法は、分析周期・窓長の変更による発話速度の補正であり、発声のなまけや音響的特徴の変形に対して性能を改善するには至っていない。今後は、発話速度の補正や発話速度別音響モデルの構築を含め、発声のなまけや音響的特徴の変形、その他の認識性能劣化の要因を改善する手法の研究を進める。

文 献

- [1] 篠崎隆宏, 古井貞照. 話し言葉認識における決定木を用いた誤り要因の分析. 日本音響学会研究発表会講演論文集, 1-1-9, 2001-10.
- [2] 前川 喜久雄. 「日本語話し言葉コーパス」の構築. 話し言葉の科学と工学ワークショップ講演予稿集, p.7-12, 2001-2.
- [3] 南條浩輝, 河原達也. 発話速度に依存したデコーディングの検討. 日本音響学会研究発表会講演論文集, 1-1-6, 2001-10.
- [4] 龍宮隆之, 菊地英明, 小磯花絵, 前川喜久雄. 大規模話し言葉コーパスにおける発話スタイルの諸相 -書き起こしテキストの分析から-. 日本音響学会研究発表会講演論文集, 2-Q-9, 2000-10.
- [5] 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.
- [6] S.Tsuge, T.Fukada and K.Kita, Frame-period adaptation for speaking rate robust speech recognition, Proc.ICSLP2000, Vol.3, pp718-721, 2000.
- [7] 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.
- [8] 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, SLP-37-13, 2001.
- [9] 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.