TR-S-0015

発話様式の変動を考慮した音響モデルの構築

高橋 伸寿　　奥田 浩三
Nobutoshi TAKAHASHI　　Kozo OKUDA

2001.2.23

近年、自然発話の音声認識技術の研究が盛んに行われるようになってきた。一般的に、読み上げ音声に比べて自然発話音声は認識率が低いとされている。本稿では、自然発話の言い淀みやなまけに注目し、そのような音声に対し機能語を加える事の有効性を検討する。

目次

1 はじめに

現在の音声認識技術は、読み上げ等のクリーンな音声についてはかなりの認識精度を達成している。我々は更に、自然な発話に対しても高い精度を達成すべく研究を行っている。自然な話し言葉に対して認識を行う事は困難であり、読み上げ音声に比べ急激に認識精度が低下する事が知られている。これには、言語的あるいは音響的特徴が書き言葉と話し言葉では大きく異なるといった理由が考えられる。

本稿では、この音響的特徴が異なる音声の認識精度を改善するために、音響モデルに対し機能語モデルを追加することを試みる。自然な発話ではある程度曖昧に発音しても発話内容や文脈上から許容される、一種の言い淀みやなまけのようなものが多数存在すると考えられる。このような音声に対し機能語を作成することによって、言語モデルや認識システムに対して手を加える事なく、音響モデルの改善のみで自然発話の認識性能を向上させる事を目的とする。

機能語モデルについては、すでに検討が行われている [1]。今回、認識タスクを独話音声としたため、更に顕著な傾向が見られることが期待できる。また、機能語の選択基準、モデルの作成法等についても別のアプローチを試みた。

2　ベースライン

2.1　認識タスク

近年、話し言葉の音声認識理解技術を高めることを目標とした、開放的融合研究『話し言葉工学』プロジェクト [2] が開始された。今回上記プロジェクトで収録されたコーパスから男性話者 4 名の講演音声を用いて評価を行った。各音声は 16kHz,16bit でサンプリングされている。テストセットの概要を表 1に示す。

表 1: テストセット

model	時間	単語総数
AS99SEP022	28 分	6305 語
AS99SEP023	30 分	4391 語
AS99SEP097	13 分	2508 語
PS99SEP025	27 分	5372 語

なお、簡易化のために以後各テストセットを下記のように表記する

```
AS99SEP022 -> AS22      AS99SEP023 -> AS23
AS99SEP097 -> AS97      PS99SEP025 -> PS25
```

2.2　ベースモデル

ベースとなる音響モデルは、 ATR の会議予約データベースの男性話者 1321 名分から作成されたものを用いた。この音声は 16kHz,16bit でサンプリングされている。フレーム幅 20msec でハミング窓を用い、フレームシフト 10msec で抽出された、ケプストラム 12 次元、デルタケプストラム 12 次元、デルタパワー 1 次元の計 25 次元の特徴ベクトルを入力とする。 HMM は連続出力分布型対角共分散行列 3 状態 10 混合の状態共有 HMM （HMnet）であり、音素数は 26、総状態数は 1400 である。

認識辞書は、融合プロジェクトから配布されているものを ATR のデコーダで

使用できるように変換したものを用いた。また言語モデルは同プロジェクトから配布されている、前向き単語バイグラム、前向き単語トライグラム、後ろ向き単語トライグラムの内、前向き単語トライグラムを ATR のデコーダで使用できるように変換したものを用いた。

ベースモデルでの認識率を表 2に示す。

表 2: ベースモデルでの認識率

speaker	acc	cor	netacc	ins	del	sub
AS22	53.26	57.69	79.52	271	635	1954
AS23	63.01	70.29	85.08	307	320	933
AS97	61.55	68.18	83.58	162	145	632
PS25	53.77	58.46	83.88	239	527	1588

2.3 デコーダ

デコーダは ATR で開発した ATRSPREC を使用した [3]。

3　機能語モデル

3.1　機能語の選択

機能語の選択基準として、以下の2種類の方法を用いた。

1. ベースモデルの認識結果を分析し、誤りやすいものを選出

2. 言語モデルから、出現尤度の高いものを選出

認識誤り傾向から機能語を選択する方法は前回行われているが、今回は別の評価式を用いて評価を行った。また、言語モデルの出現頻度から選択する方法についても試みた。

(3.1.1)　認識結果から選出した機能語

機能語は、それ単体では出現回数から見て認識率を大きく改善する事は望めない。しかしながら、言語モデルの制約によって芋蔓式に誤りを引き起こす（バースト誤り）きっかけとなった単語の認識を改善することによって、その前後の誤りが正しく認識されるといった事が期待できる。

以下の認識結果を例とする。

正解
です　けども　を　使う　　こと　が　あの
認識結果
です　よ　　こ　ます　か　こと　が　た

この例において、「けども」という単語は、一つ前の「です」という単語が正しく認識されているにも関わらず誤って認識されている。同様に、「使う」という単語も一つ後の「こと」が正しく認識されおり、言語的にアドバンテージがあるにも関わらず誤りとなっている。すなわち、このような単語については音響空間的にずれた発音がされている可能性の高い単語であると考えられる。

今回、このように前後どちらかが正解しているにも関わらず誤りとなった単語を集計して機能語とするモデルを選出した。このような単語は、全誤り単語の内約７０％であった。集計の対象となる話者はAS22、AS23、AS97とし、PS25は評価用として集計対象からはずした。

選出の尺度として、ある程度タスクに頻繁に出現し、かつ誤りやすい単語を選出するために以下に示す尺度を用いた。

（誤り数×（誤り数／出現回数））

これはすなわち、誤り率に誤りの出現頻度を掛けて正規化している。また、認識率に無関係なフィラーと、学習データにほとんど出てこない単語（出現回数60以下）は省いた。

表3が、以上によって選出された上位20単語である。

(3.1.2) 言語モデルから選出した機能語

言語モデルの出現尤度が高いもの上位２０単語を機能語として作成した。

表4が、言語モデルから選ばれた上位２０単語である。

3.2 monophoneによる認識実験

(3.2.1) 音響モデル

それぞれ選出された単語について機能語モデルを作成した。初期モデルとして、状態数が基本的に各単語の音素数の3倍のモデルを作成した。これは、ベースモデルが1音素3状態で表現されているからである。他の条件はベースモデルと同一としたmonophoneモデルであり、これに対し学習データの音素時間情報を元に切り出された音声データを用いて学習を行った。

前回の実験では状態数はすべて音素数の3倍であったが、今回状態数の調整を行った。慣れた発音の単語には発話速度が非常に速いものも見られ、このような単語はHMMの状態数よりも入力特徴量のフレーム数の方が短くなってしまう場

表 3: 認識誤りの尺度によって選ばれた上位２０単語

読み＋品詞	単語誤り数	単語正解数
と＋引用動詞	79	167
し＋本動詞	47	62
の＋連体助詞	91	334
が＋格助詞	61	204
は＋係助詞	56	184
う＋助動詞	13	0
という＋連体助詞	20	18
で＋助動詞	27	55
と＋接続助詞	30	76
に＋格助詞	41	172
よう＋助動詞	25	56
も＋係助詞	24	52
な＋助動詞	33	121
ちょっと＋副詞	15	18
て＋接続助詞	50	363
ね＋終助詞	22	59
お＋接頭辞	8	6
それ＋代名詞	18	55
か＋並立助詞	21	87
た＋助動詞	26	148

表 4: 言語モデルから選ばれた上位２０単語

読み＋品詞
の＋連体助詞
て＋接続助詞
は＋係助詞
に＋格助詞
が＋格助詞
を＋格助詞
た＋助動詞
です＋助動詞
で＋格助詞
ます＋助動詞
と＋引用助詞
ん＋準体助詞
な＋助動詞
し＋補助動詞
か＋終助詞
も＋係助詞
まし＋助動詞
こと＋普通名詞
の＋準体助詞
ね＋終助詞

合がある。ATR で用いられている HMnet では状態スキップを許していないため、そのような単語は確実に認識誤りを起こすことになる。

そこで、選出した各単語の学習データにおける発話時間を調査した。発話時間が短いもの上位 1 割の平均をとり、その時間分のフレーム数よりも状態が多いモデルに関しては、フレーム内におさまるように状態数に調整した。

作成した機能語の状態数を表 5に示す。

以上によって作成された機能語モデルをベースモデルに追加した。

(3.2.2)　言語モデル

ベースモデルの認識に用いたものと同じものを用いた。すなわち融合研究プロジェクトで配布している言語モデルのなかで、前向き単語バイグラムを ATR のデコーダで使用できるように変換したものである。

(3.2.3)　認識辞書

認識辞書に対しては、元々登録されていた単語に対して同一の ID で機能語を追加した。これにより、元々登録されている音素ならびと機能語モデルの両方で認識が行えるようにした。

(3.2.4)　実験条件

実験条件を表 6に示す。

(3.2.5)　実験結果

認識実験結果を表 7に示す。

ただし、モデルによって認識できた発話とできなかった発話があるため、共通して認識できた発話を対象として集計を行った。

表中で、ins は挿入誤り単語数、del は脱落誤り単語数、sub は置換誤り単語数である。acc と cor はそれぞれ正解精度と正解率であり、以下の式で計算した。

表 5: 機能語の音素数と状態数

読み＋品詞	音素数	状態数	読み＋品詞	音素数	状態数
と＋引用動詞	2	5	の＋連体助詞	2	5
し＋本動詞	2	5	て＋接続助詞	2	6
の＋連体助詞	2	5	は＋係助詞	2	5
が＋格助詞	2	5	に＋格助詞	2	5
は＋係助詞	2	5	が＋格助詞	2	5
う＋助動詞	1	3	を＋格助詞	1	2
という＋連体助詞	5	14	た＋助動詞	2	5
で＋助動詞	2	6	です＋助動詞	4	12
と＋接続助詞	2	6	で＋格助詞	2	6
に＋格助詞	2	5	ます＋助動詞	4	12
よう＋助動詞	3	7	と＋引用助詞	2	5
も＋係助詞	2	5	ん＋準体助詞	1	2
な＋助動詞	2	5	な＋助動詞	2	5
ちょっと＋副詞	6	18	し＋補助動詞	2	5
て＋接続助詞	2	6	か＋終助詞	2	6
ね＋終助詞	2	6	も＋係助詞	2	5
お＋接頭辞	1	3	まし＋助動詞	4	12
それ＋代名詞	4	12	こと＋普通名詞	4	12
か＋並立助詞	2	6	の＋準体助詞	2	6
た＋助動詞	2	5	ね＋終助詞	2	6

表 6: 実験条件

音響パラメータ	メルケプストラム 12 次元 + デルタメルケプストラム 12 次元 + デルタパワー 1 次元の計 25 次元ベクトル
サンプリング周波数	16kHz
フレーム幅	20msec
フレームシフト	10msec
時間窓	ハミング窓
認識タスク	男性 4 名の講演音声

$$acc = \frac{\text{総単語数} - ins - del - sub}{\text{総単語数}} \qquad cor = \frac{\text{総単語数} - del - sub}{\text{総単語数}}$$

モデル名の、base はベースモデル、emodel は認識誤り傾向から選択した機能語を加えたモデル、lmodel は言語モデルから選択した機能語を加えたモデルである。

(3.2.6) 考察

ベースモデルに比べ、機能語モデルを加えたモデルでは認識率はほぼ同程度となった。

まず見られるのが、AS22 と PS25 で脱落誤りが減っている事である。AS22 の例をあげると、

表 7: 認識実験結果

speaker	model	acc	cor	netacc	ins	del	sub
AS22	base	53.26	57.69	79.52	271	635	1954
	emodel	53.23	57.77	80.06	278	618	1966
	lmodel	53.15	57.80	80.14	285	623	1959
AS23	base	63.01	70.29	85.08	307	320	933
	emodel	62.72	70.17	85.49	314	320	938
	lmodel	62.79	70.22	85.46	313	309	947
AS97	base	61.55	68.18	83.58	162	145	632
	emodel	61.30	67.90	83.78	161	144	640
	lmodel	61.43	68.14	83.99	164	142	636
PS25	base	53.91	58.61	83.90	237	517	1568
	emodel	54.11	58.95	83.94	244	501	1567
	lmodel	54.17	58.99	83.80	243	504	1562

ベースモデル
いけ　ない　　いう　こと　が
機能語モデル
いけ　ない　と　いう　こと　が
正解
いけ　ない　と　いう　こと　が

のようにベースモデルでは脱落してしまった単語を正しく認識している事がわかる。この原因として、AS22 と PS25 の 2 話者は他の話者に比べて発話速度が早い事があげられる。機能語の状態数を減らした事によって、ベースモデルではフレームが短すぎて状態数内に収まり切らなかった単語を上手く分離できたのではないかと推測される。

挿入誤りは全体的に見て増える傾向にある。これには、機能語として選んだ単語に短いものが多いということが原因として考えられる。さらに、元々誤認識されている長い単語が機能語によって短い複数の誤りとなった事も原因の 1 つではないかと考える。

機能語によって認識が改善された例と、劣化した例を以下に示す。

改善された例

ベースモデル
お　見せ　し　ませ　ん
機能語モデル
お　見せ　し［まし］た
正解
お　見せ　し　まし　た

劣化した例

ベースモデル
なっ　て　き　た　訳
機能語モデル
［な］　でし　　た　訳
正解
なっ　て　き　た　訳

このように、改善されたものと劣化したもの両方が見られるが、全体として劣化したものが多かったために結果として認識率は低下している。そこで、機能語モデルとして選ばれた単語の正解数が、機能語モデルを加える前と後でどのように変化したかを集計した。結果を表8に示す。

表 8: 機能語とした単語の正解数の合計

model	total
base	4147
emodel	4185
正解文	6153

これを見ると、機能語モデルを加えた後でも機能語として選択した単語の正解数があまり増えていない。これは、機能語モデルが環境非依存なため、機能語のモデル精度が低いからではないかと推測される。

また、状態数の調整を行う際に参照した発話時間長は学習データの疑似対話音声から得られたものであり、自然発話における最適な状態数を得ることができなかった事や、状態数を減らした事による悪影響でモデルの表現力が低下したことも原因として考えられる。

正解率と正解精度では改善は見られなかったが、newacc は改善されている。すなわち、正解にはならなかったが候補として残っていた単語が多くなったということであり、機能語の可能性が垣間見られる。

3.3 biphone による実験

今回、より高精度な機能語モデルを作成するため、先行環境依存の biphone モデルを作成し評価を行った。

(3.3.1) 実験条件

言語モデル及び認識辞書は monophone モデルでの実験と同じものを用いた。

追加した機能語モデルは、先行環境依存の biphone モデルとした。これは、機能語は単語であるためにその前は必ず「母音、ん、語頭」のいずれかであるので、状態数や経路数が増大することを防げるからである。

今回、状態経路は先行音素ごとに独立とした。すなわち状態共有やトポロジー等は考えず、単純に先行音素ごとに独立したモデルとした。また、その他の条件は monophone モデルと同一とした。機能語モデルは monophone、biphone 両方とも認識結果の誤り傾向から選択された単語について作成したものである。

(3.3.2)　実験結果

認識実験結果を表 9に示す。表中の、 monophone は前項の emodel と同じものであり、 biphone は作成した biphone モデルである。

表 9: 認識実験結果

speaker	model	acc	cor	netacc	ins	del	sub
AS22	base	53.26	57.69	79.52	271	635	1954
	monophone	53.23	57.77	80.06	278	618	1966
	biphone	52.92	57.51	79.60	281	630	1970
AS23	baes	63.01	70.29	85.08	307	320	933
	monophone	62.72	70.17	85.49	314	320	938
	biphone	62.72	70.14	85.39	313	315	944
AS97	base	61.55	68.18	83.58	162	145	632
	monophone	61.30	67.90	83.78	161	144	640
	biphone	61.43	68.22	83.78	166	148	628
PS25	base	53.90	58.60	83.88	238	521	1575
	monophone	54.12	58.96	83.92	245	505	1573
	biphone	53.88	58.72	83.94	245	520	1570

(3.3.3) 考察

biphoneモデルを用いた場合若干の性能の劣化がみられた。これは、元々学習データに少数しか存在しない単語を環境依存ごとに振り分けて学習したために、学習データが不足しモデル性能の低下を引き起こしたからではないかと推測される。

各機能語の学習データ数を表10に示す。

表 10: 各機能語に対する学習データ数

機能語 No.	01	02	03	04	05	06	07	08	09	10
学習データ数	1116	209	6928	2006	2411	1790	512	7160	551	2441
機能語 No.	11	12	13	14	15	16	17	18	19	20
学習データ数	69	365	702	441	1198	1074	2022	278	75	2723

今後、学習データに自然発話コーパスを用いて、選択した機能語の学習データをより多く得るとともに、より自然発話の音声マッチした機能語を作成し、また発話速度を考慮に入れて入力特徴量の分解能を上げる事を試みる必要があると思われる。

4　まとめ

ベースとなるモデルにmonophoneモデル、 biphoneモデルの機能語を追加し、独話音声に対する性能の評価を行った。機能語の選択基準としては、誤り傾向から選択したものと、言語モデルから選択したものの2種類について評価した。さらに今回発話速度に注目し、状態数の調整を行った。性能はベースモデルとほぼ同等であった。

傾向としては、脱落誤りの減少と挿入誤りの増加が見られた。脱落誤りについては、状態数の調整が上手く機能したためであると思われる。挿入誤りについては、機能語が短い事が原因ではないかと思われ、今後更に長い機能語を選択する基準を検討すべきであると考える。

monophoneモデルに比べbiphoneモデルは性能が若干劣化した。これは、元々学習データの中に数の少ない機能語として選んだ単語の音声を更に環境依存モデルで振り分けたために、学習データが不足したからではないかと考える。

今後、学習データに自然発話の音声を用いて学習を行う事が必要であると考える。

参考文献

[1] 川本真一, 音声認識のための機能語モデルの作成と評価,TR-S-0011

[2] 龍宮隆之, 菊地英明, 小磯花絵, 前川喜久男. 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議講演論文集,2-Q-9, 秋季 2000.

[3] “ATRSPREC Home Page,” http://www.itl.atr.co.jp/sprec.