TR-S-0011

# 音声認識のための機能語モデルの作成と評価

# Study on function-word models for speech recognition

川本 真一
Shinichi Kawamoto
谷 智洋
Tomohiro Tani
松井 知子
Tomoko Matsui
中村 哲
Satoshi Nakamura

2000.11.15

ATR 音声言語通信研究所においては、自然発話音声を対象とした音声認識の研究を行う。自然発話音声は言い慣れた語／句を含むが、それらの発声は”なまける”ことが多い。そのために、他の部分のように音素列としてモデル化するよりも、語／句のモデルとして扱った方が高い認識性能が得られることが報告されている。本研究では、ATR の対話データベースを用いて、効果的な機能語モデルの作成法について検討する。

# 目 次

# 第1章　はじめに

自然発話音声において、言い慣れた語や句の発声はなまけることが多い。特に前置詞や接続詞、代名詞などの機能語は、頻繁に出現し、怠けた発声になりやすい。Waibel [1] は、内容語では93%にアクセントがみられるが、機能語では14%にしかアクセントがみられなかったと報告している。Lea [2] は、アクセントのない音節は認識しにくいと報告している。Lee [3] は、語彙のたった4%である機能語が約30%の出現頻度を占め、SPHINXの認識誤りのだいたい50%を占めると報告している。このように怠けた発声が多く、認識誤りになりやすい機能語を正確に認識することは、自然発話音声の認識において、重要な課題の1つである。

この問題に対して、言語モデルに対するアプローチは多く行われているが [4][5][6]、音響モデルに対するアプローチは少ない。Lee [3] は、機能語依存の音素モデルは従来の音素モデルと別に学習することで、多様な発話様式を吸収している。しかし、その機能語の選択法については、単語に依存して特に発声が怠けていると思われるものを経験的に選んでいる。この選択法や認識単位については検討の余地がある。

本研究ではATRの対話データベースを用いて、効果的な機能語モデルの作成方法について検討する。

# 第2章　助詞のみを考慮した機能語モデル

## 2.1　モデル化の対象となる助詞の選択方法

音響モデル学習用のデータベースとなるATRの旅行タスク(男性167話者、女性240話者)の品詞情報に着目し、学習データ内で助詞に分類される単語の頻度を調査した。学習データ内の助詞の出現頻度、学習データ内の全助詞に対する対象の助詞の出現頻度の割合、および最低出現頻度を表2.1および、図2.1,2.2に示す。

表2.1: 学習データ内の助詞の出現頻度、学習データ内の全助詞に対する対象の助詞の出現頻度の割合

| 頻度上位N単語 | 出現頻度の割合 [%] | 最低出現頻度 |
|---|---|---|
| 5 | 51.4 | 806 |
| 10 | 77.6 | 363 |
| 15 | 88.6 | 149 |
| **18** | **92.5** | **121** |
| 19 | 93.4 | 93 |
| 20 | 94.1 | 73 |
| 25 | 96.7 | 44 |
| 30 | 98.2 | 21 |

表2.1および、図2.1,2.2より、ある程度頻繁に出現し(学習データに最低100回は出現する単語)、学習データの助詞を十分にカバーすること(助詞の出現頻度の90%以上)を考慮し、助詞の頻度上位18単語を助詞モデルの対象として選択した。

選択した助詞の一覧を表2.2に示す。

表2.2: 助詞モデルの対象として選択した単語の一覧

| | | | | |
|---|---|---|---|---|
| ノ | ガ | ン | カ | デ |
| ハ | ニ | ト | ヲ | ケレドモ |
| ナ | ノデ | ネ | テ | バ |
| カラ | モ | ケド | | |

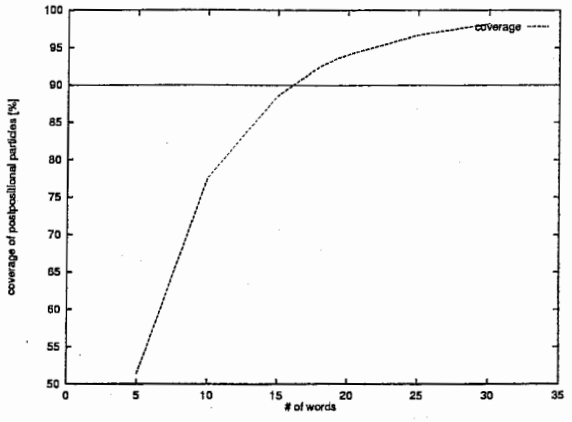


図 2.1: 学習データ内の全助詞の頻度総数に対する助詞の出現頻度上位 N 単語の頻度総数の割合

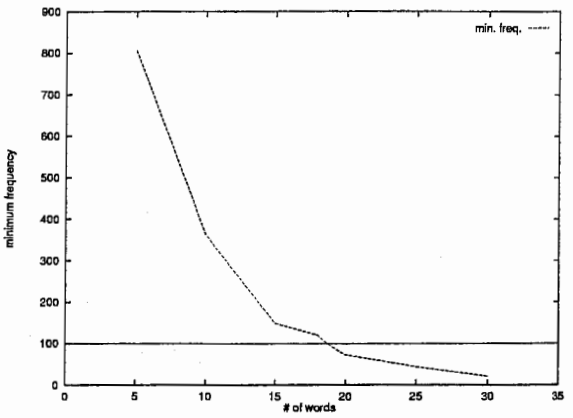


図 2.2: 助詞の出現頻度上位 N 単語における最低出現頻度

## 2.2 monophoneモデルに関する検討

### 2.2.1 音響モデル

学習データは ATR の旅行対話データベースの男性 167 話者、女性 240 話者を対象とする。対象のデータについて Viterbi 整列を与え、その出力を元に、助詞モデルの対象となるデータについて新たなラベルを割り当てる。つまり、助詞モデルの学習に用いたサンプルは音素モデルの学習には用いない。

全 6432 発話の内、言い直しなどを含まず、JMOR ファイルの単語数と ATRtrace 出力の単語数の一致のとれた 5197 発話を音響モデルの学習データとした。

助詞モデルのために割り当てた新たなラベルは、音響モデル内で音素と同じように扱うことで、助詞モデルを実現する。各音素モデルに割り当てる状態数は 3、各助詞モデルに割り当てる状態数は、助詞に含まれる音素数の 3 倍に設定した。また、状態共有は行わない。

比較対象として、従来の monophone 音素モデルも作成した。このモデルでは各音素モデルに割り当てる状態数は 3 に設定し、状態共有は行わない。

### 2.2.2 言語モデル

言語モデルとしては、多重クラス複合 N-gram を使用した。学習データは、旅行に関する対話 7195 片対話を使用した。また、クラス数は from-クラス 700、to-クラス 700 とした。

### 2.2.3 実験条件

monophone 助詞モデルの認識実験条件を表 2.3 に示す。

表 2.3: monophone 助詞モデルの実験条件

| | |
|---|---|
| 対話タスク | ATR 旅行対話データベース |
| 学習データ | 男性 167 話者、女性 240 話者 (5197 発話) |
| 認識データ | 男性 17 話者、女性 25 話者 (551 発話、open data) |
| サンプリング周波数 | 16kHz |
| 音響パラメータ | パワー、12 次メルケプストラム、Δパワー、<br>12 次Δメルケプストラムからなる計 26 次元ベクトル |
| フレーム幅 | 20 msec |
| フレーム周期 | 10 msec |
| プリエンファシス | 0.98 |
| 時間窓 | ハミング窓 |

### 2.2.4 実験結果

単語認識精度 (word accuracy) と単語正解率 (word correct) は、それぞれ式 (3.2) と式 (3.3) で計算した。

$$\text{word accuracy} = \frac{W - I - D - S}{W} \tag{2.1}$$

$$\text{word correct} = \frac{W - D - S}{W} \tag{2.2}$$

ここで、$W$ は総単語数、$I$ は挿入誤り単語数、$D$ は削除誤り単語数、$S$ は置換誤り単語数である。

monophone 助詞モデルを用いた認識実験結果を表 2.4 に示す。

表 2.4: monophone 助詞モデルの認識実験結果 (「phone」:従来の音素モデル、「phone+particle」:助詞を新たなラベルとして追加して学習したモデル、「phone−particle」:phone+particle モデルから助詞のラベルのみを削除したモデル)

| model | # of utterances | # of words | # of error words | | | word accuracy [%] | word correct [%] |
|---|---|---|---|---|---|---|---|
| | | | ins. | del. | sub. | | |
| phone+particle | 519 | 4437 | 160 | 199 | 591 | 78.59 | 82.20 |
| phone−particle | 536 | 4691 | 122 | 253 | 628 | 78.62 | 81.22 |
| phone | 543 | 4821 | 130 | 265 | 663 | 78.05 | 80.75 |

表 2.4 では、認識処理の完了した発話数が等しくない。そこで、共通して認識処理できた 502 発話についての認識結果を表 2.5 に示す。

表 2.5: 共通して認識処理できた発話に対する monophone 助詞モデルの認識実験結果 (「phone」:従来の音素モデル、「phone+particle」:助詞を新たなラベルとして追加して学習したモデル、「phone−particle」:phone+particle モデルから助詞のラベルのみを削除したモデル)

| model | # of utterances | # of words | # of error words | | | word accuracy [%] | word correct [%] |
|---|---|---|---|---|---|---|---|
| | | | ins. | del. | sub. | | |
| phone+particle | 502 | 4178 | 137 | 143 | 500 | 81.33 | 84.61 |
| phone−particle | 502 | 4178 | 99 | 189 | 480 | 81.62 | 83.99 |
| phone | 502 | 4178 | 112 | 182 | 500 | 81.00 | 83.68 |

### 2.2.5 考察

表 2.5 の認識実験結果では、助詞モデルを導入した音響モデルは従来の音響モデルとほぼ同等の性能を示している。助詞モデルを考慮することで若干の認識率向上は見られるが、今回の手法では助詞を新たなモデルとして扱う効果は薄い。この原因としては、学習サンプル数が不足していることや、学習や認識の対象とした発話音声が自然発話の中でも比較的明瞭であることが考えられる。

助詞モデルを導入して学習し、その後助詞ラベルを削除した音響モデル(「phone−particle」)は、従来の音素モデルと比べ、学習サンプル数は少ないにもかかわらず、若干認識率が向上している。これは、助詞のなまけた発話音声を音響モデル学習に用いないことで、より精密な音響モデルを学習できたのではないかと考える。つまり、学習データにも質があることを示唆する結果ではないかと考えている。

助詞モデルを導入した音響モデルの誤り傾向は、従来の音響モデルと比べ、挿入誤りが増加し、削除誤りが減少している。この考察については、後に述べる。

# 第3章　単語の誤りやすさを考慮した機能語モデル

## 3.1　モデル化の対象となる機能語の選択方法

### 3.1.1　単語の出現頻度

機能語モデルの予備的検討として全品詞を考慮し、頻度を調査した。品詞の種類は33種類。ATRの旅行タスクの全JMORファイルの総単語数335230語が調査対象である。

学習データ内の助詞の出現頻度、学習データ内の全助詞に対する対象の単語の出現頻度を表3.1, 3.2, 3.3に示す。

表3.2では、「格助詞」、「連体助詞」など助詞が頻度上位に来ており、助詞モデルのみでもかなりの割合を占めている。つまり、助詞のみを考慮した機能語モデルも出現頻度から見ると妥当なモデル化の1つと思われる。また、品詞の上位を占めているものとして、「助動詞」、「間投詞」、「接頭辞」などがあり、これらを考慮することで、さらなる性能向上が期待できる。

表3.3では、本動詞の「ネガイ」や普通名詞の「ヘヤ」、数詞の「ジュウ」などが上位に来ており、かなりタスクに依存した傾向が見られる。このことを考慮し、数詞や本動詞、普通名詞などタスク依存と思われる品詞はあらかじめ機能語の候補から除外する。

### 3.1.2　誤りやすい単語の分析

従来の音素モデルで認識実験を行った結果で誤り数の多い単語を調べた結果を表3.4に示す。

表3.4より、出現頻度の多い単語は誤り数も多くなっている。しかし、その単語の出現頻度に対する誤り率を見ると、必ずしも助詞が誤りやすいとは限らないことがわかる。例えば、接続助詞の「ケレドモ」、終助詞の「カ」、準体助詞の「ノ」などである。

### 3.1.3　機能語の選択基準

学習データにおける単語の出現頻度と単語の誤りやすさを考慮した、機能語の選択基準となる機能語尺度を提案する。この機能語尺度を式(3.1)のように定義する。

$$\text{機能語尺度 } S(w) = R1(w) \times R2(w) \times F(w) \tag{3.1}$$

表 3.1: 「読み」による単語出現頻度 (出現頻度上位 30 語)

| 読み | 出現頻度 | 割合 [%] |
|---|---|---|
| ノ | 14120 | 4.212 |
| オ | 12064 | 3.599 |
| マス | 12014 | 3.584 |
| デス | 10818 | 3.227 |
| ニ | 10062 | 3.002 |
| ガ | 9363 | 2.793 |
| カ | 8781 | 2.619 |
| ハイ | 8606 | 2.567 |
| デ | 8003 | 2.387 |
| エー | 6997 | 2.087 |
| ハ | 6143 | 1.832 |
| シ | 5311 | 1.584 |
| ン | 4785 | 1.427 |
| ゴ | 4508 | 1.345 |
| ト | 4120 | 1.229 |
| ネガイ | 3943 | 1.176 |
| ヲ | 3883 | 1.158 |
| ヘヤ | 3729 | 1.112 |
| ジュウ | 3714 | 1.108 |
| ゴザイマス | 3226 | 0.962 |
| デショウ | 3130 | 0.934 |
| ホウ | 3044 | 0.908 |
| マシ | 2800 | 0.835 |
| タ | 2758 | 0.823 |
| テ | 2647 | 0.790 |
| ネ | 2604 | 0.777 |
| サマ | 2510 | 0.749 |
| ノデ | 2496 | 0.745 |
| エ | 2404 | 0.717 |
| サン | 2281 | 0.680 |

表 3.2: 「品詞」による出現頻度 (出現頻度上位 30 品詞)

| 品詞 | 出現頻度 | 割合 [%] |
|---|---|---|
| 普通名詞 | 56052 | 16.720 |
| 助動詞 | 30055 | 8.965 |
| 格助詞 | 25978 | 7.749 |
| 本動詞 | 25236 | 7.528 |
| 数詞 | 24623 | 7.345 |
| 連体助詞 | 16189 | 4.829 |
| 間投詞 | 15952 | 4.759 |
| 接続助詞 | 15144 | 4.517 |
| 接頭辞 | 14802 | 4.415 |
| 感動詞 | 14230 | 4.245 |
| 判定詞 | 13806 | 4.118 |
| 補助動詞 | 11881 | 3.544 |
| 終助詞 | 9860 | 2.941 |
| サ変名詞 | 8656 | 2.582 |
| 係助詞 | 7354 | 2.194 |
| 副詞 | 6883 | 2.053 |
| 準体助詞 | 5504 | 1.642 |
| 接続詞 | 5368 | 1.601 |
| 形容詞 | 5076 | 1.514 |
| 準体助動詞 | 4478 | 1.336 |
| 代名詞 | 4368 | 1.303 |
| 形式名詞 | 2781 | 0.830 |
| 人称接尾辞 | 2685 | 0.801 |
| 副助詞 | 2529 | 0.754 |
| 形容名詞 | 2481 | 0.740 |
| 連体詞 | 1258 | 0.375 |
| 並立助詞 | 1038 | 0.310 |
| 接続副詞 | 378 | 0.113 |
| サ変形容名詞 | 298 | 0.089 |
| 接尾辞 | 196 | 0.058 |

表 3.3: 「読み＋品詞」による出現頻度 (出現頻度上位 30 語)

| 読み | 品詞 | 出現頻度 | 割合 [%] |
|---|---|---|---|
| ノ | 連体助詞 | 13210 | 3.941 |
| マス | 助動詞 | 12014 | 3.584 |
| オ | 接頭辞 | 11974 | 3.572 |
| デス | 判定詞 | 10079 | 3.007 |
| ハイ | 感動詞 | 8605 | 2.567 |
| ニ | 格助詞 | 7309 | 2.180 |
| エー | 間投詞 | 6985 | 2.084 |
| カ | 終助詞 | 6742 | 2.011 |
| ハ | 係助詞 | 6009 | 1.793 |
| ガ | 接続助詞 | 4876 | 1.455 |
| ン | 準体助詞 | 4748 | 1.416 |
| デ | 格助詞 | 4623 | 1.379 |
| ガ | 格助詞 | 4487 | 1.338 |
| シ | 補助動詞 | 4309 | 1.285 |
| ネガイ | 本動詞 | 3924 | 1.171 |
| ヲ | 格助詞 | 3883 | 1.158 |
| ヘヤ | 普通名詞 | 3729 | 1.112 |
| ジュウ | 数詞 | 3703 | 1.105 |
| デショウ | 準体助動詞 | 3130 | 0.934 |
| ホウ | 普通名詞 | 3044 | 0.908 |
| デ | 判定詞 | 3038 | 0.906 |
| ゴ | 接頭辞 | 2803 | 0.836 |
| マシ | 助動詞 | 2791 | 0.833 |
| ニ | 数詞 | 2753 | 0.821 |
| タ | 助動詞 | 2748 | 0.820 |
| ネ | 終助詞 | 2599 | 0.775 |
| サマ | 人称接尾辞 | 2507 | 0.748 |
| ノデ | 接続助詞 | 2496 | 0.745 |
| エ | 間投詞 | 2403 | 0.717 |
| ゴザイマス | 補助動詞 | 2291 | 0.683 |

表 3.4: 従来の音素モデルで誤り数の多い単語 (誤り数上位 30 語)

| 読み＋品詞 | 単語誤り数 | $\frac{\text{単語誤り数}}{\text{全単語誤り数}}$ [%] | $\frac{\text{単語誤り数}}{\text{単語出現頻度}}$ [%] | $\frac{(\text{第 3 列})\times(\text{第 4 列})}{100}$ [%] |
|---|---|---|---|---|
| ノ+連体助詞 | 43 | 3.749 | 30.935 | 1.160 |
| ヲ+格助詞 | 32 | 2.790 | 49.231 | 1.373 |
| アッ+間投詞 | 29 | 2.528 | 96.667 | 2.444 |
| エー+間投詞 | 24 | 2.092 | 63.158 | 1.322 |
| ガ+格助詞 | 24 | 2.092 | 53.333 | 1.116 |
| アノー+間投詞 | 23 | 2.005 | 92.000 | 1.845 |
| ニ+格助詞 | 19 | 1.656 | 17.593 | 0.291 |
| ハ+係助詞 | 18 | 1.569 | 33.962 | 0.533 |
| アノ+間投詞 | 17 | 1.482 | 45.946 | 0.681 |
| ア+間投詞 | 15 | 1.308 | 50.000 | 0.654 |
| オ+接頭辞 | 13 | 1.133 | 6.915 | 0.078 |
| エ+間投詞 | 13 | 1.133 | 56.522 | 0.641 |
| デス+判定詞 | 12 | 1.046 | 4.167 | 0.044 |
| エート+間投詞 | 11 | 0.959 | 91.667 | 0.879 |
| ノ+準体助詞 | 10 | 0.872 | 7.353 | 0.064 |
| デ+格助詞 | 10 | 0.872 | 13.889 | 0.121 |
| ト+格助詞 | 10 | 0.872 | 20.833 | 0.182 |
| ソウ+副詞 | 7 | 0.610 | 9.459 | 0.058 |
| テ+接続助詞 | 7 | 0.610 | 29.167 | 0.178 |
| カ+終助詞 | 7 | 0.610 | 5.600 | 0.034 |
| ネ+終助詞 | 7 | 0.610 | 17.949 | 0.110 |
| タイ+助動詞 | 6 | 0.523 | 10.000 | 0.052 |
| タ+助動詞 | 6 | 0.523 | 10.000 | 0.052 |
| ナイ+助動詞 | 5 | 0.436 | 31.250 | 0.136 |
| カラ+格助詞 | 5 | 0.436 | 22.727 | 0.099 |
| ケレドモ+接続助詞 | 4 | 0.349 | 4.651 | 0.016 |
| ナ+連体助詞 | 4 | 0.349 | 11.765 | 0.041 |
| モ+係助詞 | 4 | 0.349 | 50.000 | 0.174 |
| イタシ+補助動詞 | 3 | 0.262 | 13.043 | 0.034 |
| コト+形式名詞 | 3 | 0.262 | 50.000 | 0.131 |

ここで、各変数 $R1(w), R2(w), F(w)$ は、以下のようなものである。

$w$ ...対象単語
$R1(w)$ ...全誤りに対する対象単語 $w$ の誤りの割合
$R2(w)$ ...対象単語 $w$ のテストセット出現頻度に対する誤りの割合
$F(w)$ ...対象単語 $w$ の学習セット内の出現頻度

式 (3.1) に示す機能語尺度 $S(w)$ の値が大きい程、対象単語 $w$ が頻繁に出現し、かつ誤りやすい単語であることを示している。表 3.5 に機能語尺度上位 30 単語を示す。

表 3.5 では、やはり助詞や間投詞が多く上位の単語に含まれている。それとともに、接頭辞の「オ」、判定詞の「デス」、副詞の「ソウ」、助動詞の「タ」、形容詞の「ナイ」など、多くの品詞が上位に含まれていることがわかる。今回、これらの単語も機能語とみなしてモデル化することを試みる。また、以降の機能語モデルでは機能語尺度上位 10 単語、および上位 20 単語を機能語モデルの対象とする。

## 3.2 monophoneモデルに関する検討

### 3.2.1 音響モデル

学習データは ATR の旅行対話データベースの男性 167 話者、女性 240 話者を対象とする。対象のデータについて Viterbi 整列を与え、その出力を元に、機能語モデルの対象となるデータについて新たなラベルを割り当てる。つまり、機能語モデルの学習に用いたサンプルは音素モデルの学習には用いない。

全 6432 発話の内、言い直しなどを含まず、JMOR ファイルの単語数と ATRtrace 出力の単語数の一致のとれた 5197 発話を音響モデルの学習データとした。

機能語モデルのために割り当てた新たなラベルは、音響モデル内で音素と同じように扱うことで、機能語モデルを実現する。各音素モデルに割り当てる状態数は 3、各機能語モデルに割り当てる状態数は、機能語に含まれる音素数の 3 倍に設定した。また、状態共有は行わない。

比較対象として、従来の monophone 音素モデルも作成した。このモデルでは各音素モデルに割り当てる状態数は 3 に設定し、状態共有は行わない。

### 3.2.2 言語モデル

言語モデルとしては、多重クラス複合 N-gram を使用した。学習データは、旅行に関する対話 7195 片対話を使用した。また、クラス数は from-クラス 700、to-クラス 700 とした。

### 3.2.3 実験条件

monophone 機能語モデルの認識実験条件を表 3.6 に示す。

表 3.5: 機能語尺度 $S(w)$ 上位 30 単語 ($w$:対象単語, $R1(w)$:全誤りに対する対象単語 $w$ の誤りの割合, $R2(w)$:対象単語 $w$ のテストセット出現頻度に対する誤りの割合, $F(w)$:対象単語 $w$ の学習セット内の出現頻度)

| $w$ (読み + 品詞) | $S(w)$ | $R1(w)$ | $R2(w)$ | $F(w)$ |
|---|---|---|---|---|
| ノ + 連体助詞 | 4.848 | 3.749 | 30.935 | 1291 |
| ヲ + 格助詞 | 2.182 | 2.790 | 49.231 | 491 |
| ガ + 格助詞 | 1.687 | 2.092 | 53.333 | 467 |
| エー + 間投詞 | 1.588 | 2.092 | 63.158 | 371 |
| アッ+ 間投詞 | 1.266 | 2.528 | 96.667 | 160 |
| ハ + 係助詞 | 1.149 | 1.569 | 33.962 | 666 |
| アノー + 間投詞 | 1.111 | 2.005 | 92.000 | 186 |
| ア + 間投詞 | 0.637 | 1.308 | 50.000 | 301 |
| アノ + 間投詞 | 0.626 | 1.482 | 45.946 | 284 |
| ニ + 格助詞 | 0.557 | 1.656 | 17.593 | 591 |
| オ + 接頭辞 | 0.441 | 1.133 | 6.915 | 1746 |
| エート + 間投詞 | 0.410 | 0.959 | 91.667 | 144 |
| デス + 判定詞 | 0.357 | 1.046 | 4.167 | 2508 |
| デ + 格助詞 | 0.313 | 0.872 | 13.889 | 799 |
| エ + 間投詞 | 0.288 | 1.133 | 56.522 | 139 |
| ト + 格助詞 | 0.213 | 0.872 | 20.833 | 362 |
| ソウ + 副詞 | 0.134 | 0.610 | 9.459 | 713 |
| テ + 接続助詞 | 0.104 | 0.610 | 29.167 | 181 |
| タ + 助動詞 | 0.104 | 0.523 | 10.000 | 620 |
| ナイ + 形容詞 | 0.097 | 0.262 | 50.000 | 228 |
| カ + 終助詞 | 0.096 | 0.610 | 5.600 | 875 |
| ネ + 終助詞 | 0.088 | 0.610 | 17.949 | 247 |
| タイ + 助動詞 | 0.081 | 0.523 | 10.000 | 481 |
| ナイ + 助動詞 | 0.078 | 0.436 | 31.250 | 178 |
| モ + 係助詞 | 0.074 | 0.349 | 50.000 | 131 |
| コト + 形式名詞 | 0.045 | 0.262 | 50.000 | 105 |
| イタシ + 補助動詞 | 0.043 | 0.262 | 13.043 | 387 |
| カラ + 格助詞 | 0.039 | 0.436 | 22.727 | 121 |
| ナ + 連体助詞 | 0.035 | 0.349 | 11.765 | 262 |
| ノ + 準体助詞 | 0.033 | 0.872 | 7.353 | 159 |

表 3.6: monophone 機能語モデルの実験条件

| | |
|---|---|
| 対話タスク | ATR 旅行対話データベース |
| 学習データ | 男性 167 話者、女性 240 話者 (5197 発話) |
| 認識データ | 男性 17 話者、女性 25 話者 (551 発話、open data) |
| サンプリング周波数 | 16kHz |
| 音響パラメータ | パワー、12 次メルケプストラム、Δパワー、<br>12 次Δメルケプストラムからなる計 26 次元ベクトル |
| フレーム幅 | 20 msec |
| フレーム周期 | 10 msec |
| プリエンファシス | 0.98 |
| 時間窓 | ハミング窓 |

### 3.2.4 実験結果

単語認識精度 (word accuracy) と単語正解率 (word correct) は、それぞれ式 (3.2) と式 (3.3) で計算した。

$$\text{word accuracy} = \frac{W - I - D - S}{W} \tag{3.2}$$

$$\text{word correct} = \frac{W - D - S}{W} \tag{3.3}$$

ここで、$W$ は総単語数、$I$ は挿入誤り単語数、$D$ は削除誤り単語数、$S$ は置換誤り単語数である。

テストセット 551 発話の内、monophone 機能語モデルを用いて認識処理できた発話数について表 3.7 に示す。

表 3.7: 551 発話の内 monophone 機能語モデルを用いて認識処理できた発話数 (「phone」: 従来の音素モデル、「phone+f.w.10」:機能語 10 語を新たなラベルとして追加して学習したモデル、「phone−f.w.10」:phone+f.w.10 モデルから機能語 10 語のラベルのみを削除したモデル、「phone+f.w.20」:機能語 20 語を新たなラベルとして追加して学習したモデル、「phone−f.w.20」:phone+f.w.20 モデルから機能語 20 語のラベルのみを削除したモデル)

| | phone | phone+f.w.10 | phone−f.w.10 | phone+f.w.10 | phone−f.w.10 |
|---|---|---|---|---|---|
| 認識処理できた発話数 | 543 | 528 | 531 | 524 | 519 |

表 3.7 に示す発話数の内、各音響モデルで共通して認識処理できた 502 発話についての認識結果を表 3.8 に示す。

表 3.8: 共通して認識処理できた発話に対する monophone 機能語モデルの認識実験結果 (「phone」:従来の音素モデル、「phone+f.w.10」:機能語 10 語を新たなラベルとして追加して学習したモデル、「phone−f.w.10」:phone+f.w.10 モデルから機能語 10 語のラベルのみを削除したモデル、「phone+f.w.20」:機能語 20 語を新たなラベルとして追加して学習したモデル、「phone−f.w.20」:phone+f.w.20 モデルから機能語 20 語のラベルのみを削除したモデル)

| model | # of utterances | # of words | # of error words | | | word accuracy [%] | word correct [%] |
|---|---|---|---|---|---|---|---|
| | | | ins. | del. | sub. | | |
| phone+f.w.10 | 502 | 4178 | 114 | 170 | 509 | 81.02 | 83.75 |
| phone−f.w.10 | 502 | 4178 | 103 | 207 | 529 | 79.92 | 82.38 |
| phone+f.w.20 | 502 | 4178 | 129 | 154 | 501 | 81.17 | 84.27 |
| phone−f.w.20 | 502 | 4178 | 108 | 210 | 509 | 80.21 | 82.79 |
| phone | 502 | 4178 | 112 | 182 | 500 | 81.00 | 83.68 |

### 3.2.5 考察

表 3.8 の認識実験結果では、機能語モデルを導入した音響モデルは従来の音響モデルとほぼ同等の性能を示している。機能語モデルを考慮することで若干の認識率向上は見られるが、今回の手法では機能語を新たなモデルとして扱う効果は薄い。この原因としては、学習サンプル数が不足していることや、学習や認識の対象とした発話音声が自然発話の中でも比較的明瞭であることが考えられる。

機能語モデルを導入して学習し、その後機能語ラベルを削除した音響モデル(「phone−f.w.10」)は、従来の音素モデルと比べ、若干ではあるが、認識率が低下している。これは、機能語として、「エー」とか「アッ」などの間投詞が多く含まれており、これにより多くの母音の学習サンプルが音素モデルに使用されなかったため、音素モデルの頑健性が失われた結果ではないかと考えている。

機能語モデルを導入した音響モデルの誤り傾向は、従来の音響モデルと比べ、挿入誤りが増加し、削除誤りが減少している。この考察については、後に述べる。

# 第4章　誤り傾向ついての考察

従来の音素モデル (phone) と、音素＋機能語 20 語モデル (phone+f.w.20) の認識結果を比較し、その傾向について考察する。

## 4.1　機能語モデル導入によって性能が向上したケース

機能語として間投詞「あの」を追加したことで、以下のように性能が向上する場合が確認された。

TAC70301.0060.A

音素モデル

本日 は 十 九 日 まで この 部屋 とっ ていただい てる の です けれども

音素＋機能語 20 語モデル

あの じつは 十 九 日 まで この 部屋 とっ ていただい てる の です が

正解 (JTEXT ファイル)

[あの] じつは 十 九 日 まで この お 部屋 とっ ていただい てる の です が

「あの」という単語自体は認識率に影響しない単語として扱われている。しかし、「あの」が認識できることで、その後に続く「じつは」が正確に認識できている。一方、従来の音素モデルでは「あの」をうまく認識できない影響で、その後に続く「じつは」と連結した形で「本日 は」と語認識していると考えられる。図 4.1 に、発話「TAC70301.0060.A」を両モデルで認識したときの尤度を示す。この図は、音素＋機能語 20 語モデルにおける「あの」の部分が、音素モデルより高い尤度を示しており、追加した機能語モデルと良くマッチしたことを示している。

また、機能語として助詞「の」、「に」を追加したことで、以下のように性能が向上する場合が確認された。

TAC70203.0160.A

音素モデル

和室 は 人数 として は 値段 は お いくら な の です か

音素＋機能語 20 語モデル

和室 の 方 に 移る としたら お 値段 の 方 お いくら な の ですか

正解 (JTEXT ファイル)

和室 の 方 に 移る としたら 値段 の 方 は いくら (ん) なる の ですか

従来の音素モデルの認識結果において、このケースでは発話前半に現れる助詞の「の」、「に」が認識できなかったことで、多くの単語が「人数」という一語に連結されている。これは、従来の音素モデルによる認識において、削除誤りが多くなる一例である。図 4.2 に、発話「**TAC70203.0160.A**」を両モデルで認識したときの尤度を示す。この図は、音素＋機能語 20 語モデルにおける「の」の部分が、音素モデルより高い尤度を示しており、追加した機能語モデルと良くマッチしたことを示している。

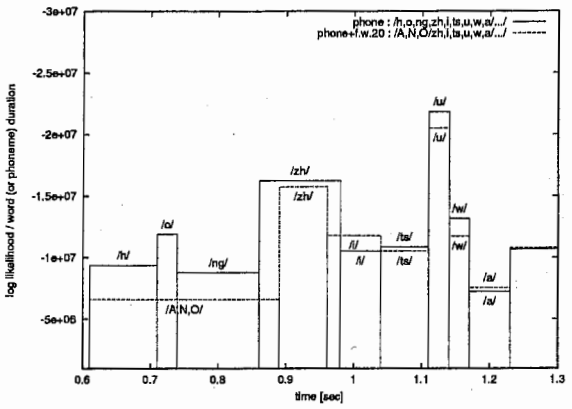


図 4.1: 発話「TAC70301.0060.A」の機能語 (もしくは音素) 継続時間と機能語 (もしくは音素) あたりの平均対数尤度 (phone: 従来の音素モデル, phone+f.w.20:音素＋機能語 20 語モデル, 小文字:音素モデル, 大文字:機能語モデル)

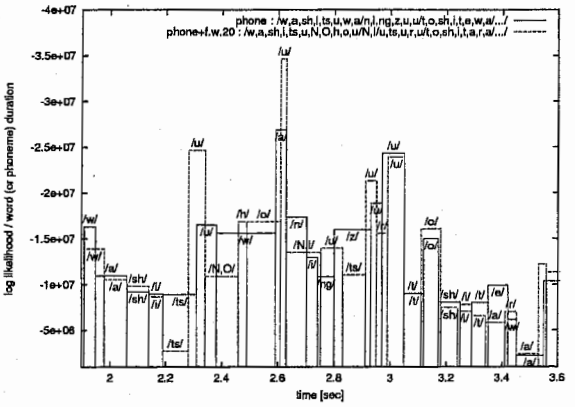


図 4.2: 発話「TAC70203.0160.A」の機能語 (もしくは音素) 継続時間と機能語 (もしくは音素) あたりの平均対数尤度 (phone: 従来の音素モデル, phone+f.w.20:音素＋機能語 20 語モデル, 小文字:音素モデル, 大文字:機能語モデル)

## 4.2 機能語モデル導入によって性能が低下したケース

機能語として副詞「そう」や接頭辞「お」を追加したことで、以下のように性能が低下する場合が確認された。

TAC70015.0220.A

音素モデル

承知 し まし た

音素＋機能語20語モデル

そう いたし まし た

正解

承知 し まし た

TAC70101.0120.A

音素モデル

おおいし ようこ です

音素＋機能語20語モデル

お 衣装 です

正解

おおいし ようこ です

副詞「そう」、接頭辞「お」の誤認識の影響で、その後に続く単語にも影響を及ぼしていることが確認される。

図4.3に、発話「**TAC70015.0220.A**」を両モデルで認識したときの尤度を、図4.4に、発話「**TAC70101.0120.A**」を両モデルで認識したときの尤度を示す。図4.3における機能語「そう」の尤度をみると、音素モデルに比べて高くなっている。また、図4.4における機能語「お」についても、「おお」との区別ができていない。これらは、学習データ不足が原因の1つではないかと考えられる。つまり音素モデルの学習サンプルに対して、機能語モデルの学習サンプルは少なく、そのためなまけた発話に対して機能語モデルの分散は小さくなったのではないかと考える。また、機能語として選択した語「お」は、音素モデル「お」と同じ音素表記であり、探索過程において曖昧な候補となる。これも認識を難しくする原因の1つと考えられる。

さらに機能語として助詞を追加したことで、以下のように性能が低下する場合が確認された。

TAC70021.0200.A

音素モデル

六 時 頃 お 願い いたし ます

音素＋機能語20語モデル

六 時 頃 を お 願い いたし ます

正解

六 時 頃 お 願い いたし ます

TAC70102.0240.A

音素モデル

だいたい ゆうがた の

音素＋機能語20語モデル

だいたい に は あの

正解

だいたい ゆうがた の

発話「**TAC70021.0200.A**」では、助詞「を」が挿入誤りとして、発話「**TAC70102.0240.A**」では、助詞「に」、「は」が挿入誤りとなっている。このような助詞などの比較的短い機能語が誤った発話とマッチし、挿入エラーが増えていると考えられる。

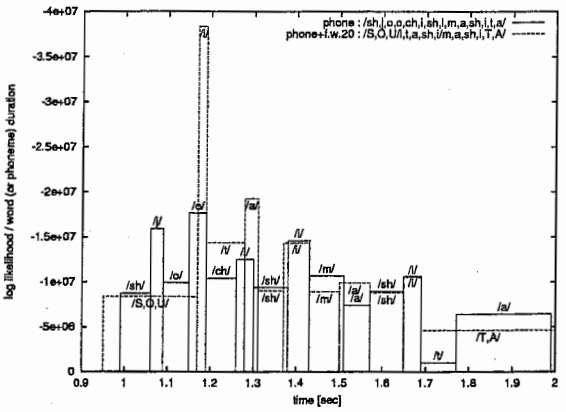


図 4.3: 発話「TAC70015.0220.A」の機能語 (もしくは音素) 継続時間と機能語 (もしくは音素) あたりの平均対数尤度 (phone: 従来の音素モデル, phone+f.w.20:音素＋機能語 20 語モデル, 小文字:音素モデル, 大文字:機能語モデル)

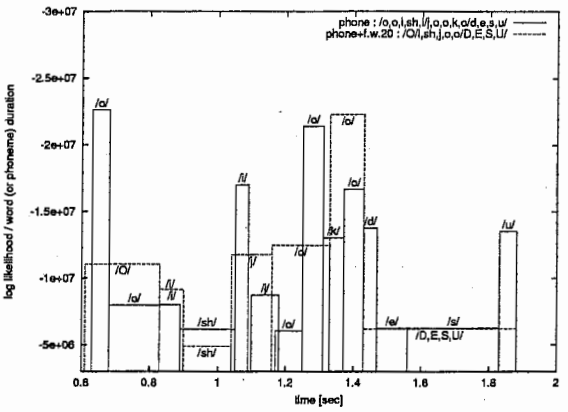


図 4.4: 発話「TAC70101.0120.A」の機能語 (もしくは音素) 継続時間と機能語 (もしくは音素) あたりの平均対数尤度 (phone: 従来の音素モデル, phone+f.w.20:音素＋機能語 20 語モデル, 小文字:音素モデル, 大文字:機能語モデル)

図 4.5 に、発話「**TAC70021.0200.A**」を両モデルで認識したときの尤度を、図 4.6 に、発話「**TAC70102.0240.A**」を両モデルで認識したときの尤度を示す。図 4.5 における助詞「を」は、他の音素モデルに比べ高い尤度を示している。また、図 4.6 における助詞「に」、「は」についても同様に他の音素モデルに比べ高い尤度を示している。これらにの原因としては、学習サンプルが不足していることが考えられる。

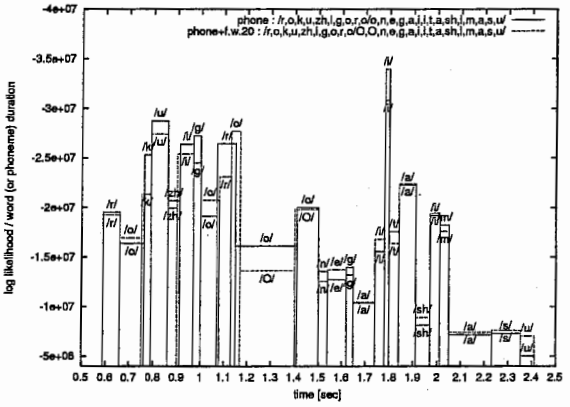


図 4.5: 発話「TAC70021.0200.A」の機能語 (もしくは音素) 継続時間と機能語 (もしくは音素) あたりの平均対数尤度 (phone: 従来の音素モデル, phone+f.w.20:音素＋機能語 20 語モデル, 小文字:音素モデル, 大文字:機能語モデル)

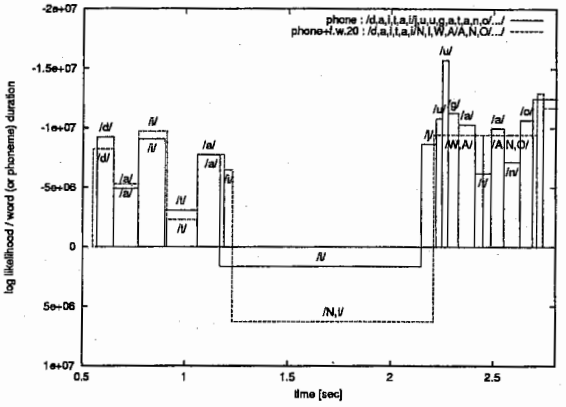


図 4.6: 発話「TAC70102.0240.A」の機能語 (もしくは音素) 継続時間と機能語 (もしくは音素) あたりの平均対数尤度 (phone: 従来の音素モデル, phone+f.w.20:音素＋機能語 20 語モデル, 小文字:音素モデル, 大文字:機能語モデル)

# 第5章　まとめ

機能語として助詞18語を採用した機能語モデルを作成し、その性能を評価した。性能は従来の音素モデルとほぼ同等であった。

また、機能語の選択基準として学習データの量と単語の誤りやすさを考慮した尺度 (機能語尺度) を提案した。この機能語尺度によって選択した機能語10語、および20語で機能語モデルを作成し、その性能を評価した。性能は従来の音素モデルとほぼ同等であった。

機能語20語を選択した機能語モデルの誤り傾向について解析した。その結果、間投詞や助詞などで機能語モデルの効果が見られた。また、追加した機能語モデルが誤った単語とマッチし、挿入誤りが増加していることが分かった。この原因としては、学習サンプルが不足していることが考えられる。

今後の課題としては、今回の実験タスクが比較的きれいなデータである可能性があるので、別タスクによる評価実験が必要と考える。機能語選択基準として音響モデルの尤度なども考慮することが考えられる。さらには、今回の実験ではmonophoneでの実験であったが、triphoneについても検討が必要と考える。

# 謝辞

約一ヵ月間にわたって、研究をはじめさまざまな面でお世話になった、音声言語研究所第一研究室、および第二研究室の皆様に心から感謝致します。また、研究環境を整えて頂いたTSGの方々にも感謝致します。

# 関連図書


[1] A. Waibel, "Prosody and Speech Recognition," Morgan Kaufmann Publishers, Inc., San Mateo, California, 1988.

[2] W. A. Lea, "Trends in Speech Recognition," Prentice-Hall, Englewood Cliffs, NJ, 1980.

[3] K. F. Lee, "Automatic Speech Recognition: The Development of the SPHINX System," Kluwer Academic Publisers, Boston, 1989.

[4] 伊藤 彰則, 牧野 正三, 木村 正行, 城戸 健一, "機能語予測 CYK 法による連続音声認識とその評価," 電子情報通信学会 技術報告 SP90-22, pp.25-32, 1990.

[5] Ryosuke Isotani, Shoichi Matsunaga and Shigeki Sagayama, "Speech Recognition Using Function-Word N-Grams and Content-Word N-Grams," Trans. IEICE Inf. & Syst., Vol.E78-D, No.6, pp.692-697, 1995.

[6] Kazuyuki Takagi, Rei Oguro, Kenji Hashimoto and Kazuhiko Ozeki, "Performance Evaluation of Word Phrase and Noun Category Language Models for Broadcast News Speech Recognition," Proceedings of ICSLP'98, Vol.6, pp.2507-2510, 1998.