# 音声認識のための音響モデルと言語モデルの仕様

**Ver.1.0 （2004/3/23）**

**南條 浩輝 †，河原 達也 †，篠崎 隆宏 ‡，古井 貞熙 ‡**

†: 京都大学大学院 情報学研究科（言語モデル担当）

‡: 東京工業大学大学院 情報理工学研究科 （音響モデル担当）

## 目 次



## 1 はじめに

本マニュアルは,『日本語話し言葉コーパス（CSJ）』を用いて学習した講演音声認識のための標準的なモデル（音響モデルと言語モデル）の仕様を解説したものである[1]．CSJを用いた音声認識の標準モデルについては [1] などで発表を行ってきたが，その後，修正などを行っているので，本マニュアルを参照されたい．

また，本マニュアルでは，CSJ における音声認識のテストセットについても述べる．

[1]最終版が確定する前に学習を行ったため，細部において異なる部分がある．

## 2　音響モデル

音響モデルは混合連続分布 HMM（対角共分散）であり，HTK[2] で作成した．音素ごとに 3 状態 left-to-right HMM（飛び越し遷移なし）でモデル化を行い，音素環境依存モデル（状態共有 triphone モデル）を学習した．その際，決定木に基づく状態共有を行い，状態数 3000 のモデル（16 混合）を学習した[2]．各モデルには MLLR 適応のための回帰クラス情報が付加されている．

表 1 に，音響モデルの学習データの一覧，すなわち講演の種別と性別ごとのデータ量の一覧を示す．学習データには，CSJ における音声認識のテストセットの講演（後述：表 10）及びテストセットの話者の他の講演は含まれていない．

提供する音響モデルは，汎用的と考えられる性別非依存モデル（表 1 のうち，モデル名が書いてあるものもの: 3 種類）である．

表 1: 音響モデルの学習データの内訳と作成モデル

| 学習データ | | | モデル名 （ファイル名） |
|---|---|---|---|
| 講演種別 | 性別 | データ量 | |
| 学会講演 | 男性 | 787 講演／ 186 時間 | |
| | 女性 | 166 講演／ 42 時間 | |
| | 男性+女性 | 953 講演／ 228 時間 | 学会 GID モデル （AM/CSJ-APS/hmmdefs.gz） |
| 模擬講演 | 男性 | 721 講演／ 124 時間 | |
| | 女性 | 822 講演／ 134 時間 | |
| | 男性+女性 | 1543 講演／ 258 時間 | 模擬 GID モデル （AM/CSJ-SPS/hmmdefs.gz） |
| 学会+模擬 | 男性 | 1508 講演／ 310 時間 | |
| | 女性 | 988 講演／ 176 時間 | |
| | 男性+女性 | 2496 講演／ 486 時間 | 学会+模擬 GID モデル （AM/CSJ-APS,SPS/hmmdefs.gz） |

以下，詳細に音響モデルの説明を行う．

[2] これらの音素モデルは，CSJ の分節音ラベルを付与するために用いたモデルとは異なるものである（「『日本語話し言葉コーパス』の分節音ラベリング」（segment.pdf）参照）．

### 2.1 音響分析

音声データ（16kHz，16bit）をフレーム長 25msec のハミング窓，フレーム周期 10msec で音響分析を行った．各フレーム毎に MFCC（12 次元），ΔMFCC（12 次元），ΔPower（1 次元）を計算し，計 25 次元の特徴量ベクトルを求め，音響モデルの学習に利用した．ただし，発話ごとにケプストラム平均除去（CMS）を行っている．

詳細な音響分析条件を表 2 に示す．

表 2: 音響分析条件

| | |
|---|---|
| サンプリング周波数 | 16 kHz |
| プリエンファシス | 0.97 |
| 分析窓 | Hamming 窓 |
| 分析窓長 | 25 ms |
| 窓間隔 | 10 ms |
| 特徴パラメタ | MFCC（12 次）+ΔMFCC（12 次）+Δ パワー（計 25 次） |
| 周波数分析 | 等メル間隔フィルタバンク |
| フィルタバンク | 24 チャネル |
| CMS | 発話単位 |

ここで，パワーは式 (1) に基づいて求め，デルタパラメータは式 (2) に基づいて求める．

$$Power = \log \sum_{n=1}^{N} {s_n}^2 \tag{1}$$

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \tag{2}$$

ただし，$\Theta = 2$

表 3 に，使用した HTK config file を示す．

表 3: 使用した HTK config file

| |
|---|
| SOURCEFORMAT=NOHEAD |
| SOURCEKIND = WAVEFORM |
| SOURCERATE = 625 |
| TARGETKIND = MFCC_E_D_Z |
| TARGETRATE=100000.0 |
| SAVECOMPRESSED=F |
| SAVEWITHCRC=F |
| WINDOWSIZE=250000.0 |
| USEHAMMING=T |
| PREEMCOEF=0.97 |
| NUMCHANS=24 |
| NUMCEPS=12 |
| ZMEANSOURCE=T |
| ENORMALISE=F |
| ESCALE=1.0 |
| TRACE=0 |
| RAWENERGY=F |

## 2.2 音素体系

使用した音素は表 4 に示す 42 種類である．ここで，q は促音に伴う無音，sp は音声中の短い無音である．また，silB は発話の先頭の無音，silE は発話の終端の無音であり，発話は基本的に 500 ミリ秒以上の無音区間で区切ったものと定義している（次章（2.3 章）参照）．また，N は撥音，a:〜o:は長母音を表す．

表 4: 音素セット

| |
|---|
| a i u e o a: i: u: e: o: |
| N w y j my ky by gy ny hy ry py |
| p t k ts ch b d g z m n s sh h f r |
| q sp silB silE |

## 2.3 学習に用いるラベルの作成

音響モデルの学習には音声とそれに対応する音素列が必要である．音素列は，CSJ の書き起こしテキストの発音形（カナ）から作成した．CSJ の書き起こしに含まれるタグのうち，?タグや W タグなど発音形に複数候補がある場合（併記されている場合）は，前のエントリを使っている．

例： (W ソエ; ソレ) デ --> ソエデ， (? ホーコー, ホーホー) --> ホーコー

カナは表 5 に示す対応表に従って音素列に変換を行い，モノフォンラベルを作成した．ここでは，原則として 500 ミリ秒以上の無音区間を発話の始終端とみなして silB および silE を割り当てた．ただし，500 ミリ秒以上の無音区間が存在せず，音声区間が 20 秒以上続いた場合は，直後のポーズ（500 ミリ秒未満でも）を発話の始終端とみなしている．

このラベルを用いて初期音響モデル（モノフォンモデル）を作成し，母音直後にショートポーズ sp が入りうるというルールを適用して強制アライメントを行うことでラベルに sp を挿入した．提供する音響モデルは，このショートポーズ sp 入りのラベルを用いて学習されたものである．

表 5: かな音素対応表

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ア | a | イ | i | ウ | u | エ | e | オ | o |
| カ | k a | キ | k i | ク | k u | ケ | k e | コ | k o |
| ガ | g a | ギ | g i | グ | g u | ゲ | g e | ゴ | g o |
| サ | s a | シ | sh i | ス | s u | セ | s e | ソ | s o |
| ザ | z a | ジ | j i | ズ | z u | ゼ | z e | ゾ | z o |
| タ | t a | チ | ch i | ツ | ts u | テ | t e | ト | t o |
| ダ | d a | ヂ | j i | ヅ | z u | デ | d e | ド | d o |
| ナ | n a | ニ | n i | ヌ | n u | ネ | n e | ノ | n o |
| ハ | h a | ヒ | h i | フ | f u | ヘ | h e | ホ | h o |
| バ | b a | ビ | b i | ブ | b u | ベ | b e | ボ | b o |
| パ | p a | ピ | p i | プ | p u | ペ | p e | ポ | p o |
| マ | m a | ミ | m i | ム | m u | メ | m e | モ | m o |
| ラ | r a | リ | r i | ル | r u | レ | r e | ロ | r o |
| ワ | w a | ヲ | o | | | | | | |
| ヤ | y a | ユ | y u | ヨ | y o | | | | |
| キャ | ky a | キュ | ky u | キョ | ky o | | | | |
| ギャ | gy a | ギュ | gy u | ギョ | gy o | | | | |
| シャ | sh a | シュ | sh u | ショ | sh o | | | | |
| ジャ | j a | ジュ | j u | ジョ | j o | | | | |
| チャ | ch a | チュ | ch u | チョ | ch o | | | | |
| ニャ | ny a | ニュ | ny u | ニョ | ny o | | | | |
| ヒャ | hy a | ヒュ | hy u | ヒョ | hy o | | | | |
| ビャ | by a | ビュ | by u | ビョ | by o | | | | |
| ピャ | py a | ピュ | py u | ピョ | py o | | | | |
| ミャ | my a | ミュ | my u | ミョ | my o | | | | |
| リャ | ry a | リュ | ry u | リョ | ry o | | | | |
| イェ | i e | シェ | sh e | ジェ | j e | ティ | t i | トゥ | t u |
| チェ | ch e | ツァ | ts a | ツィ | ts i | ツェ | ts e | ツォ | ts o |
| ディ | d i | ドゥ | d u | デュ | d u | ニェ | n i e | ヒェ | h e |
| ファ | f a | フィ | f i | フェ | f e | フォ | f o | フュ | hy u |
| ブィ | b i | ミェ | m e | ウィ | w i | ウェ | w e | ウォ | w o |
| クヮ | k a | グヮ | g a | スィ | s u i | ズィ | j i | テュ | t e y u |
| ヴァ | b a | ヴィ | b i | ヴ | b u | ヴェ | b e | ヴォ | b o |
| ン | N | ッ | q | ー | : | | | | |

## 2.4 音素環境依存モデル

このモノフォンラベルから，前後の音素環境を考慮したトライフォンラベルを作成し，音素コンテクスト依存音響モデル（トライフォンモデル）の学習を行った．ただし，silB，silE，sp に関しては，音素環境の依存化は行っていない．トライフォンラベル作成の際には，情報処理振興事業協会（IPA）の補助で開発された「日本語ディクテーション基本ソフトウェア」[3]を参考にして，表 6 に示す縮訳規則を適用した．

表 6: —縮約規則—

- 文脈において長母音と通常の母音との違いを無視する
  a:-k+a → a-k+a
- 右音素文脈では拗音を区別しない
  -a+ky → *-a+k
- 拗音の左音素文脈を共通化する
  ky-a+* → y-a+*

## 2.5 トライフォンの状態共有

日本語に出現する全てのトライフォンを統計的に学習するためには，膨大な学習データが必要であり，現実的には不可能である．そこで，音響的特徴が類似したトライフォン（の各状態）に対して決定木に基づくクラスタリングを行い，状態共有トライフォンを作成した．具体的には，同一中心音素を持つトライフォンの状態位置毎に行った．

決定木に基づくクラスタリングは，以下の手順で行った．

（1）全ての状態を一つの集合にまとめ，最もゆう度が高くなるように分割を行う質問を 1 つ選択し，分割を行う．ここで，質問は,「後続音素が鼻音か？」や「先行音素は母音 “あ” か？」などである（表 7 参照）．

（2）再帰的に質問を行い，ゆう度の上昇がしきい値を下回れば終了し，同じ集合に残った状態を共有化する．

CSJ で提供する音響モデルは，このようにして総状態数が約 3000 になるように分割を行い，学習したものである．

[3]http://www.itakura.nuee.nagoya-u.ac.jp/~ takeda/IPA/

表 7: クラスタリングに用いた質問（分類規則）

| 質問名 | 共有化するコンテクスト |
|---|---|
| L_Nasal | N-∗, n-∗, m-∗ |
| R_Nasal | ∗+N, ∗+n, ∗+m |
| L_Bilabial | p-∗, b-∗, f-∗, m-∗, w-∗ |
| R_Bilabial | ∗+p, ∗+b, ∗+f, ∗+m, ∗+w |
| L_DeltalAlveolar | t-∗, d-∗, ts-∗, z-∗, s-∗, n-∗ |
| R_DeltalAlveolar | ∗+t, ∗+d, ∗+ts, ∗+z, ∗+s, ∗+n |
| L_PalatoAlveola | ch-∗, j-∗, sh-∗ |
| R_PalatoAlveola | ∗+ch, ∗+j, ∗+sh |
| L_Velar | k-∗, g-∗ |
| R_Velar | ∗+k, ∗+g |
| L_Glottal | h-∗ |
| R_Glottal | ∗+h |
| L_YOUON | y-∗ |
| L_SOKUON | q-∗ |
| R_SOKUON | ∗+q |
| L_R | r-∗ |
| R_R | ∗+r |
| L_N | N-∗ |
| R_N | ∗+N |
| L_A | a-∗ |
| R_A | ∗+a |
| L_I | i-∗ |
| R_I | ∗+i |
| L_U | u-∗ |
| R_U | ∗+u |
| L_E | e-∗ |
| R_E | ∗+e |
| L_O | o-∗ |
| R_O | ∗+o |

## 3 形態素解析と単語辞書

形態素は，国立国語研究所で定義された短単位 [3] [4]に基づいており，形態素解析システムは，通信総合研究所で最大エントロピー法により CSJ を用いて統計的に学習されたもの [4] [5]を用いている．

単語辞書は，語彙エントリ-表記-音素列 の集合で HTK 形式 [2] で構成した（ファイル名：LM/csj.htkdic）．語彙エントリには句読点は含まれていないが，2 種類のポーズ記号，すなわち，発話の始終端のポーズに対応するロングポーズ記号<sil>とそれ以外のポーズに対応するショートポーズ記号<sp>が含まれている．ただし，<sil>は 1000msec 以上のポーズに，<sp>はそれ未満のポーズに割り当てている．

表 8: 単語辞書の例

| 語彙 | 表記 | 発音 |
|---|---|---|
| <sil> | [<sil>] | silB |
| <sil> | [<sil>] | silE |
| <sp> | [<sp>] | sp |
| ．+名詞 | [．] | t e N |
| １０ｄＢ+名詞/数詞 | [１０ｄＢ] | j u: d e: b i: |
| １６ＰＰＳ+名詞/数詞 | [１６ＰＰＳ] | j u: r o k u p i: p i: e s u |
| １６ＰＰＳ+名詞/数詞 | [１６ＰＰＳ] | j u: r o k u p i: p i e s u |
| １Ａ+名詞 | [１Ａ] | w a N w e: |
| １Ａ+名詞 | [１Ａ] | w a N e: |
| １ＥＲＢ+名詞/数詞 | [１ＥＲＢ] | i ch i i: a: r u b i: |
| ２０Ｋ+名詞/数詞 | [２０Ｋ] | n i j i q k e: |
| ２０Ｋ+名詞/数詞 | [２０Ｋ] | n i j u q k e: |
| ２Ａ+名詞 | [２Ａ] | ts u: e: |
| ２ＤＫ+名詞/数詞 | [２ＤＫ] | n i: d e: k e: |
| ２ＤＫ+名詞/数詞 | [２ＤＫ] | n i: d i: k e: |

発音（読み）には，CSJ の発音形から取得された実際の発音を付与している．CSJ では基本的に文節ごとに，表記（基本形）とその発音（発音形）がペアで記述されているため，単語単位での自動マッチングを行って，割り当てた．読みが複数ある場合は，それらを辞書のエントリに登録している．ただし，ある語彙エントリに対し，可能なすべての発音を割り当てた場合，認識時にわき出し誤りが増加するため，各語彙エントリに対して，各発音エントリの生起確率を求め，その値がしきい値（0.2）以下のものは除いた．

発音表記（カナ）から音素列への変換は，音響モデルを作成した際に用いたものと同じルールで行っている．

語彙は，CSJ における出現頻度の高いもので構成した．具体的には，CSJ の学会講演と模擬講演からなるテキスト集合（2596 講演，6.67M 単語）で 4 回以上出現した形態素で構成した（カットオフ 3）．語彙サイズは 25,300，発音エントリ総数は 27,249 である．

[4]「『日本語話し言葉コーパス』の形態論情報の概要」（pos.pdf）参照.
[5]「短単位・長単位データマニュアル」（wdb.pdf）参照.

# 4　言語モデル

3 章で定義した語彙を用いて単語 N-gram 言語モデルを作成した．CMU-Cambridge SLM toolkit ver.2[5][6]を用いて順向きの単語 2-gram モデル（csj.2gram.gz）と逆向きの単語 3-gram モデル（csj.3gram.gz）を作成した．back-off 平滑化には Witten-Bell 法を用いており，N-gram エントリのカットオフは行っていない．語彙に含まれているポーズ記号<sil>及び<sp>は，通常の単語と同様に扱っている．

学習データは，語彙を作成したものと同一の講演であり，CSJ における音声認識のテストセットの講演（30 講演: 表 10–後述）は含んでいない．また，従来，用いられていた音声認識のテストセットの講演[7]も含んでいない．

表 9 に提供する言語モデルの詳細をまとめる．学習データは，CSJ の 2592 講演（6.67M 単語）であり，ユニグラムエントリ数（語彙サイズ）は 25K，バイグラムエントリ及びトライグラムエントリ数は，それぞれ 0.7M，2.6M である．

表 9: 言語モデルの詳細

| | | |
|---|---|---|
| 学習データ量 | （講演数） | 2,592 |
| | （単語数 *） | 6,671,844 |
| 1-gram エントリ数 | | 25,300 |
| 2-gram エントリ数 | | 731,728 |
| 3-gram エントリ数 | | 2,611,952 |

*: <sil>及び<sp>を含まない

# 5　CSJ における音声認識のテストセット

テストセットは『日本語話し言葉コーパス（CSJ）』のモニタ版に含まれるものから選定した．講演の種類と性別のバランスを考慮して表 10 に示す 3 セットを構成した．その際，学会講演では男性が非常に多いため（表 1 参照），男性依存のモデルの評価を行えるように男性のみのセットも用意した．具体的には，男性話者の学会講演 10 講演のセット（test-set 1），男性話者 5 名・女性話者 5 名の学会講演 10 講演のセット（test-set 2），男性話者 5 名・女性話者 5 名の模擬講演 10 講演のセット（test-set 3）を構成した．

各セットの 10 講演は，文献 [6][10] を参考にして，音声認識性能に影響を与える要因と考えられるパープレキシティ・言い直し率・発話速度の 3 つの尺度を用いて，母集団である講演集合（2002 年 10 月時点で利用可能な CSJ の講演）をよく表現するように選択した．具体的には，講演ごとのパープレキシティ，言い直し率，発話速度の分布が正規分布に従うと仮定し，その分布形状に従ってバランスよく講演を選択した．

[6] http://mi.eng.cam.ac.uk/~ prc14/toolkit.html

[7] A01M0007, A01M0035, A01M0074, A02M0117, A03M0100, A05M0031, A06M0134, 及びその他の 3 講演．文献 [6][7][8][9] 参照．

これらのテストセット 30 講演の話者は全て異なり，また，それ以外の CSJ の講演にも基本的に含まれていない．**test-set 2 の A01M0056 と同一話者の講演が存在し，それらの ID は，S05M0613, R00M0187, D01M0019, D04M0056, D02M0028, D03M0017 である．この点のみを注意することで，話者独立な評価が可能である．**

表 10: CSJ における音声認識テストセット講演一覧

**（test-set 1） 学会講演 10 講演（男性 10）**
A01M0097 A01M0110 A01M0137 A03M0106 A03M0112
A03M0156 A04M0051 A04M0121 A04M0123 A05M0011

**（test-set 2） 学会講演 10 講演（男性 5，女性 5）**
A01M0056 A01M0141 A02M0012 A03M0016 A06M0064
A01F0001 A01F0034 A01F0063 A03F0072 A06F0135

**（test-set 3） 模擬講演 10 講演（男性 5，女性 5）**
S00M0008 S00M0070 S00M0079 S00M0112 S00M0213
S00F0019 S00F0066 S00F0148 S00F0152 S01F0105

## 参考文献


[1] T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui. Benchmark Test for Speech Recognition using the Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 135–138, 2003.

[2] P.C.Woodland, C.J.Leggetter, J.J.Odell, V.Valtchev, and S.J.Young. The 1994 HTK Large Vocabulary Speech Recognition System. In *IEEE Int'l Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, Vol. 1, pp. 73–76, 1995.

[3] 小椋秀樹. 話し言葉コーパスの単位認定基準について. 話し言葉の科学と工学ワークショップ講演予稿集, pp. 21–28, Feb. 2001.

[4] 内元清貴, 井佐原均. 話し言葉コーパスの形態素解析. 話し言葉の科学と工学ワークショップ講演予稿集, pp. 33–38, Feb. 2002.

[5] P.R.Clarkson and R.Rosenfeld. Statistical Language Modeling using the CMU-Cambridge Toolkit. In *Proc. European Conf. Speech Communication & Technology (EUROSPEECH)*, pp. 2707–2710, 1997.

[6] 篠崎隆宏, 古井貞熙. 日本語話し言葉コーパスを用いた講演音声認識. 情処学論, Vol. 43, No. 7, pp. 2098–2107, 2002.

[7] T.Shinozaki and S.Furui. Towards Automatic Transcription of Spontaneous Presentations. In *Proc. European Conf. Speech Communication & Technology (EUROSPEECH)*, pp. 491–494, 2001.

[8] H.Nanjo and T.Kawahara. Speaking-Rate Dependent Decoding and Adaptation for Spontaneous Lecture Speech Recognition. In *IEEE Int'l Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, pp. 725–728, 2002.

[9] 南條浩輝, 加藤一臣, 李晃伸, 河原達也. 大規模な日本語話し言葉データベースを用いた講演音声認識. 信学論, Vol. J86-DII, No. 4, pp. 450–459, 2003.

[10] T.Shinozaki and S.Furui. Analysis on Individual Differences in Automatic Transcription of Spontaneous Presentations. In *IEEE Int'l Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, Vol. 1, pp. 729–732, 2002.

## 連絡先

**〒 606-8501　京都市左京区吉田二本松町**
**京都大学 学術情報メディアセンター南館 4F**
**河原達也**
**kawahara@i.kyoto-u.ac.jp**