# 大語彙日本語連続音声認識研究基盤の整備
## － 汎用音素モデルの作成 －

武田一哉 (名大)　峯松信明 (豊橋技科大)　伊藤彰則 (山形大)　伊藤克亘 (電総研)
宇津呂武仁 (奈良先端大)　河原達也 (京大)　小林哲則 (早稲田大)　清水徹 (KDD)
田本真詞 (NTT)　荒井和博 (NTT)　山本幹雄 (筑波大)　竹沢寿幸 (ATR)
松岡達雄 (NTT)　鹿野清宏 (奈良先端大)

あらまし

大語彙連続音声認識研究の推進のためには、データベースと共に、標準 (ベースライン) となるソフトウェア (言語モデル・音響モデル・認識プログラム) が必要であり、著者らはその基盤整備を進めている。本稿では、音響モデルの構築について述べる。

# Common Platform of Japanese Large Vocabulary Continuous Speech Recognition Research
## – Construction of Acoustic Model –

Kazuya Takeda (Nagoya Univ.), Nobuaki Minematsu (Toyohashi Univ. of Tech.),
Akinori Ito (Yamagata Univ.), Katsunobu Itou (ETL), Takehito Utsuro (NAIST),
Tatsuya Kawahara (Kyoto Univ.), Tetsunori Kobayashi (Waseda Univ.),
Toru Shimizu (KDD), Masafumi Tamoto (NTT), Kazuhiro Arai (NTT),
Mikio Yamamoto (Tsukuba Univ.), Toshiyuki Takezawa (ATR), Tatsuo Matsuoka (NTT),
Kiyohiro Shikano (NAIST)

Abstract

For Japanese large vocabulary continuous speech recognition (LVCSR) research, we are developing standard baseline software repository that includes language models, acoustic models and recognition engines. In this report, construction of acoustic models is discussed.

# 1 はじめに

文として発声された音声を連続的に認識する,「連続音声認識」技術は，音声認識技術の中心的な課題である．特に数万単語をこえる語彙を扱う大語彙連続音声認識は，日常人間が発声する音声言語のほぼ全てを処理対象とすることから，音声認識技術の応用分野の拡大には極めて重要な課題である．

近年，米国や欧州では，音声／言語のデータやモデルを共通化した上で，大語彙連続音声認識システムの性能比較を行ない，これを通じて個別の要素技術の評価を進めることにより，システムの性能が飛躍的に向上してきた．近年の報告によれば，２万単語の語彙を持つ新聞読みあげ音声の認識を，標準的なＰＣを用いてほぼ実時間で実行し，単語誤り率１０％以下の認識精度を達成することが，標準的なシステムの性能である．

一方わが国では，要素技術の研究は高水準にあるが，その評価が（例えば，孤立単語認識による音声コントロール装置や，電話音声応答装置に代表される）応用システムに依存して行なわれているのが現状であり，大規模なタスクへの応用は少ない．また，共通データベースがなく，要素技術間の相互比較が極めて困難な状況にある．

このような背景の中で，情報処理学会音声言語研究会において，鹿野清宏奈良先端大教授をグループリーダとして，大語彙連続音声認識研究用データベースに関するワーキンググループが平成９年１１月に発足した．ワーキンググループ 設立の目的は，大語彙連続音声認識に含まれる様々な要素技術の性能を，迅速かつ厳密に評価しうる基盤を整備することであり，これまでに研究用新聞記事コーパスの設計や，ディクテーション基本技術講習会の開催などを行なってきた [1][2].

WGではさらに，要素技術の評価の基準を与えるために，標準的な言語モデルと汎用音素モデルの作成を行なった．本稿ではこのうち汎用音素モデルについて解説する．

表 1: 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] |
| 特徴パラメータ | MFCC(12 次) + ΔMFCC + ΔΔMFCC + ΔPow + ΔΔPow (計 38 次) |
| 周波数分析 | 等メル間隔フィルタバンク |
| フイルタバンク | 24 チャネル |
| CMS | 発声単位で実行 |

# 2 音響分析条件

作成した音素モデルの音響分析条件は表 1 に示すとおりである．特徴パラメーターにはMFCC[3] を用い，動的な特徴パラメータとしてΔMFCC とともにΔΔMFCC，を用いている．音響分析は HTK version2.0 [4] の HCopy コマンドを用いて行なった．

さらにマイクロフォン等，入力環境の異る様々なシステムの評価に利用することを考慮して，学習データの発声を単位としたケプストラム平均除去，(CMS) [5] を行なっている．

# 3 音素体系

音素体系及び仮名表記との対応は，日本音響学会連続音声データベース [6] のローマ字表記に準じた．作成された音素モデルは表 2に示す合計４１種類である．sil は文頭／文末の，sp は文節間の無音モデルを表しており，q には促音に伴う無音に対応させている ．

学校 => g a q k o u

表 2: 音素表

| |
|---|
| 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 sil |

```
0 2000000 sil
2000000 2500000 sp
2500000 3100000 a
3100000 3400000 r
3400000 4400000 a
   . . . .

9700000 10500000 j
10500000 10800000 i
10800000 12300000 ts # 音素 u が脱落
12300000 14600000 o
14600000 14900000 sp
14900000 16300000 s  # 音素 u が脱落
16300000 16900000 b
16900000 17900000 e
17900000 18400000 t
18400000 19200000 e
   . . . .
```

発声内容：あらゆる現実を全て自分....

図 1: 学習データの音素表記例

表 3: 音素毎の HMM 状態数

| | 男性 |
|---|---|
| 5 状態 | a y |
| 4 状態 | o o: w j ky by gy ny hy ry py p t k ts ch b d g z |
| 3 状態 | i u e a: i: u: e: N m n s sh h f r q sp sil |
| | **女性** |
| 5 状態 | a |
| 4 状態 | e o a: o: w y j my ky by gy ny hy ry py p t k ts ch b d g z |
| 3 状態 | i u i: u: e: N m n s sh h f r q sp sil |

また a:〜i:等は長母音を表している.

音素モデルの連結学習は，発声の音素表記（音素ラベル）に従って HTK version2.0 の HERest を用いて行なった．音素ラベルは，極力発声内容を忠実に表すものとするために，発声のバリエーション (無声化，ポーズの挿入位置など) を考慮した文法を用いて，話者依存HMMにより学習音声の認識を行なった結果に基づいて作成した．

一例を図 1に示す．このラベルは無声化に伴う母音（u）の脱落を反映している．ただし後述するトライフォンには子音の連続を認めていないため，上の例の音素ラベルは，トライフォンを用いたラベルに変換する場合には

sp s b → sp s+u u-b+e

と展開される.

## 4 HMM のトポロジー

HMMの状態数は，性別・音素別に表 3に示す3〜5状態とした．状態数 5のモデルでは，一部第 2，第 4状態の飛び越しを認めている．状態数は，当初 5状態から学習を開始し，学習につれて状態停留確率が極端に小さ

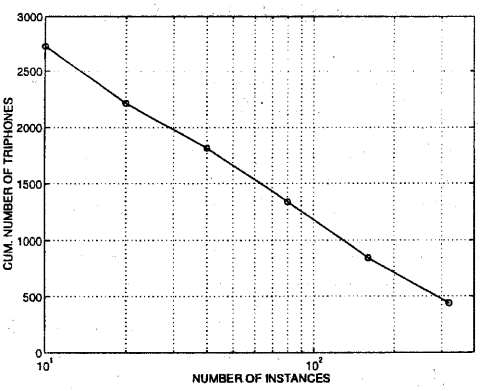


図 2: トライフォンの出現数と累積種類.

くなる状態が生じた場合に，当該状態を削除することで決定した．この結果，男性と女性とでは一部の音素でモデルの状態数が異っている．

## 5 Triphone 体系

上記の音素体系では，子音の連鎖を除外すると 19403 (monophone, biphone を含む) 種類のトライフォン (３つ組音素連鎖) の出現が可能である．これを以下のヒューリスティックな規則：

- 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+*`

を用いてマージすることで，7369 種類にまとめることが可能となる．

このように，存在可能な triphone 数が 7000 以上であるにも係わらず，例えば学習データにＡＳＪ連続音声コーパス [6] の全話者を用いた場合，出現する triphone の異り総数は 2904 種類であり，各々の出現回数は図2のとおりである．図より，ＡＳＪ連続音声コーパスを学習に用いる場合に 40 以上の学習データが確保されている triphone の種類は約 1800 種類，100 程度の学習データが確保されている triphone 数 は 1000 種類程度にすぎないことが分かる．

## 6 音素環境のトップダウンクラスタリングによる状態の共有

同一の中心音素を持つトライフォンモデルの各状態をクラスタリングし，クラスタ毎に状態間でパラメータを共有することでロバストなパラメータの推定を行なうことが一般的である．汎用の音素モデルを作成するという観点からは，語彙によらず精度の高いトライフォンモデルを提供することが望まれるため，トップダウンのクラスタリングにより文脈分類木を作成することが有効である [7]．トップダウンクラスタリングは，HTK version2.0 の HHEd コマンドにより行なった．基本的なアルゴリズムは，以下のとおりである．

1) 予め用意された分類条件にしたがって全ての状態クラスタを２分割する．

2) 1) で得られた分割のうち，２分割後の状態クラスタ間の距離が最も大きい分割を用い，状態クラスタ数を１つ増やす．

3) 2) の分割に伴う状態クラスタ間の距離が定められた閾値を下回らない限り 1) に戻り分割を続ける．

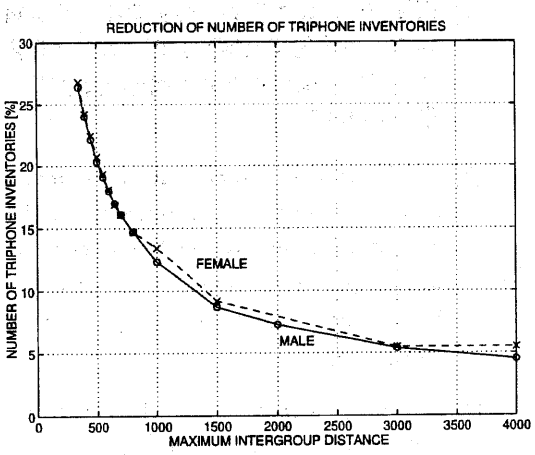


図 3: クラスタリングによるトライフォン数の削減: 縦軸は全ての可能なトライフォン (10835) に対する割合.

状態のクラスタ $i,j$ 間の距離は，分散で正規化された平均ベクトル間のユークリッド距離として，以下に従い計算した.

$$d(i,j) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{V}\sum_{k=1}^{Vs}\frac{(\mu_{isk}-\mu_{jsk})^2}{\sigma_{isk}\sigma_{jsk}}\right]^{\frac{1}{2}}$$

上式において $s$ はストリーム (MFCC,ΔMFCC など個々の特徴パラメータベクトル) に対応するインデックスであり，$S$ はストリーム総数 (本モデルの場合 5) である. $\mu$, $\sigma$ はそれぞれのクラスタに対応する正規分布の平均と分散を表している. またコンテキスト分類木作成に用いた，コンテキストの分類条件は主として調音位置に着目して設定した.

図 3に，トップダウンクラスタリングによるトライフォン状態数の削減の程度を示す. 今回作成したモデルは，しきい値を 1500.0 に設定しており，総状態数は 907 (男性)，934(女性) となっている.

状態毎に得られたコンテキスト分類木を用いて，未知の (学習データに出現しない) コンテキストを学習済みのトライフォンで置き換えることが可能となる. 学習データに存在しないトライフォン状態を，コンテキスト分類木を用いて既学習のトライフォン状態に割り当てて得られた結果が，最終トライフォン状態セットである.

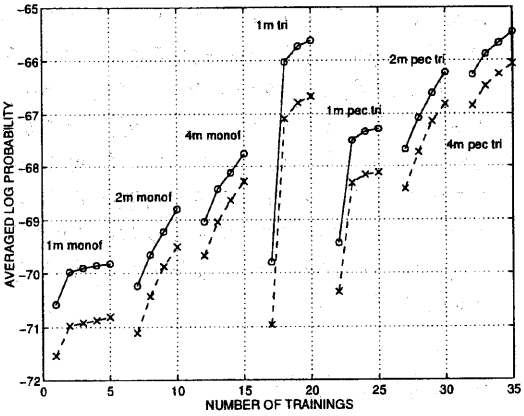


図 4: 学習回数毎の学習データに対する平均対数確率値 (フレーム当たり).

## 7 混合数の増加

トライフォン状態のクラスタリングまでは，単一ガウス分布によりモデルの作成を行なった. これらのモデルの混合数を 2，4 の順で増やし，最終的な汎用トライフォン HMM を作成した.

## 8 学習データに対する平均尤度

図 4に上記の学習過程それぞれの段階における，学習データに対する平均尤度を示す. 実線は男性モデルの，破線は女性モデルの平均対数確率値にそれぞれ対応しており，1m,2m 等は状態当たりのガウス分布の混合数を，tri は

triphone を，PEC はトップダウンのクラスタリング結果であることをそれぞれ表している．

トップダウンのクラスタリングに基づき状態共有を行なったモデルでは，状態あたり 4 混合のガウス分布を用いることで状態共有を行なわないモデルと同等の平均対数確率が得られた．また女性の対数確率値は常に男性に比べて低く，その差は混合数が少ない時ほど大きいことが分かる．

## 9 今後の課題

今回作成した音素モデルは，標準的な手続きにより作成されたものである．音素モデルの確率計算が大語彙連続音声認識全体の計算処理に占める割合は高く，Tied-Mixture モデル [8]，HMNet[9]，など確率計算の効率化を図る様々なモデル表現が研究されている．今後これらのモデル表現に基づく標準モデルを作成するとともに，表現形式間の比較を行なっていく必要がある．また，音素だけでなく音節や不均一な音素連鎖などの認識単位について検討を行なうことも重要な課題である．

### 謝辞

本WGの活動の一部は，ＩＰＡ(情報処理振興事業協会) の独創的先進的情報技術に係わる研究開発事業「日本語ディクテーション基本ソフトウエアの開発」の支援を受けて行なわれたものであり，関係各位の御協力に深謝いたします．

## 参考文献

[1] 伊藤克亘，武田一哉，竹沢寿幸，松岡達雄，鹿野清宏，" 大語彙連続音声認識のための読み上げ文コーパスの構築，" 情報処理学会第 54 回 (平成 9 年前期) 全国大会，5H-10, Vol.2, pp.225–226 (1997-03)

[2] 板橋秀一，山本幹男，竹沢寿幸，小林哲則，" 日本音響学会新聞記事読み上げ音声コーパスの構築，" 日本音響学会講演論文集 平成 9 年 9 月

[3] S.B.Davis and P.Mermelstein "Comparison of parametric representations for monosyllabic word recognition in countinuously spoken sentences," IEEE Trans. on Speech and Signal Processing, pp.357–366 (1980)

[4] S. Young, J.Jansen, J.Odell, D.Ollason, P.Woodland, "The HTK Book", Entropic Research Lab.

[5] B.Atal "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. Am., Vol.55, 6, pp.1304–1312 (1974).

[6] 小林哲則, 板橋秀一, 速水 悟, 竹沢寿幸, "日本音響学会研究用連続音声データベース", 日本音響学会誌 48 巻 12 号, pp.888–893 (1992).

[7] S.J. Young and P.Woodland, "The use of state tying in continuous speech recognition", Proc. of Eurospeech'93, pp.2203–2206, 1993.

[8] J.R.Bellegarda and D.Nahamoo, "Tied Mixture Continuous Parameter Modelis for Large Vocabulary Isolated Speech Recognition," Proc. of ICASSP'89, pp.13–16, 1989

[9] 鷹見淳一，嵯峨山茂樹，" 逐次状態分割による隠れマルコフ網の自動生成，" 信学論 (D-II), J76-DII, Vol.10, pp.2155–2164, (1993)