第7回：音声と統計モデル（HMM・GMM）＋自己教師あり表現学習

音声処理と認識

担当：小野原彩香｜北星学園大学｜3年後期

この回は動画回です。Google Formsでテストを受けてください。

本回の到達目標

1. 従来型音声認識システムの全体アーキテクチャを説明できる
2. HMMの3要素を理解し、音声認識への応用を説明できる
3. GMMが連続値の特徴量の出力確率密度をどう表すか理解する
4. 自己教師あり学習の概念と、wav2vec 2.0・HuBERTの違いを説明できる

従来型音声認識の全体アーキテクチャ

要素	役割
特徴量抽出	波形 → 短時間の特徴量列（MFCCなど）
音響モデル（AM）	特徴量列と音響状態の対応を評価
発音辞書	単語と音素列の対応を与える
言語モデル（LM）	単語列の事前確率・文脈を評価
デコーダ	各情報を統合し、候補の単語列を探索

$$W^* = \arg\max_W P(W \mid O) = \arg\max_W p(O \mid W) P(W)$$

O：観測列、W：単語列。音素列を先に一つへ確定する単純な直列処理ではない。

Kaldiのデコードグラフ解説：実装ではモデルの重みや挿入ペナルティなども使う。

マルコフ連鎖の概念

マルコフ性（Markov Property）：現在の状態は直前の状態だけに依存する、という仮定

$$P(q_t \mid q_{t-1}, q_{t-2}, \ldots, q_1) = P(q_t \mid q_{t-1})$$

天気のマルコフ連鎖（具体例）

状態集合 $Q = \{\text{晴れ}, \text{曇り}, \text{雨}\}$

$$A = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

（行：現在の状態、列：次の状態）

> 音声では「状態 = 音素の一部（音素サブユニット）」と対応づける。音声は時間軸上で状態が遷移する過程として捉えられる。

HMMの3要素

$$\lambda = (A, B, \pi)$$

記号	名前	意味
π_i	初期確率	時刻1で状態iから始まる確率
a_{ij}	遷移確率	状態iから、次の時刻の状態jへ進む確率
$b_i(o)$	出力分布	状態iでの観測の確率、または確率密度

- **観測できるもの**：音声の特徴量列
- **隠れたもの**：どの音響状態がどのフレームに対応するか

ASRの最終目標は**文字・単語列**。その探索の中でHMMの状態系列も扱う。

> MFCCのような連続値には確率密度を使う。密度値は1を超える場合もある。

音素HMMの構造

説明用に、音素/a/へ3状態の左から右へ進むHMMを割り当てる。

遷移	確率	意味
S1 → S1、S2 → S2、S3 → S3	a_{11}、a_{22}、a_{33}	自己ループ
S1 → S2、S2 → S3	a_{12}、a_{23}	次の状態へ
S3 → 終端	$a_{3,\mathrm{end}}$	音素HMMを出る

各状態から特徴量を出力する。出る遷移の確率は終端も含めて合計1。自己ループ確率が高いと継続時間の期待値が長くなる。

HMMの3つの問題

1. **評価**：観測列の確率（密度） → 前向きアルゴリズム
2. **復号**：最良状態系列 → Viterbiアルゴリズム
3. **学習**：パラメータ推定 → Baum-Welch（EMの特殊版）

Viterbiアルゴリズムの直感的理解

目標：観測列Oに対して、最も確率が高い状態系列を求める。

$$Q^* = \arg\max_Q P(Q, O \mid \lambda)$$

$$\delta_t(j) = \left[\max_i \delta_{t-1}(i) a_{ij}\right] b_j(o_t)$$

$\delta_t(j)$：時刻tで状態jにいる、最良パスの値。

処理	内容
初期化	$\delta_1(i) = \pi_i b_i(o_1)$
更新・記録	各状態へ入る最大値と、直前状態の番号を残す
経路の復元	最後の最大値から、記録した番号を逆にたどる

N状態・T時刻の全探索N^Tを、密な遷移なら$O(N^2T)$へ削減。連続観測では出力密度を用いる。

ガウス混合モデル（GMM）の基礎

多次元ガウス分布：例として13次元MFCCに差分2種類を加えた39次元特徴量を扱える。

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}$$

この密度式では共分散行列Σを正定値とする。

GMM：複数のガウス分布を重み付き和で組み合わせる。

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \quad w_k \geq 0, \quad \sum_k w_k = 1$$

- 多峰性を含む、複雑な特徴量分布を近似できる
- 2成分でも必ず2つの山になるわけではない
- 成分が性別・感情などの群と一対一に対応するとは限らない

GMM-HMMからDNN-HMMへ

GMM-HMM：連続観測の出力密度をGMMでモデル化。

$$b_i(\mathbf{o}) = \sum_{k=1}^{K} w_{ik} \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$$

EMのEステップで状態・成分の事後確率を計算し、Mステップでパラメータを更新する。大域最適解や汎化性能の改善は保証しない。

DNN-HMM：DNNが推定する状態事後確率を、状態の事前確率で割って利用。

$$p(\mathbf{o}_t \mid q_t = i) \propto \frac{P(q_t = i \mid \mathbf{o}_t)}{P(q_t = i)}$$

DNNの出力そのものは、GMMの出力密度と同じ量ではない。

評価の注意：TIMITの音素認識でのPERと、単語認識でのWERを区別する。学習・評価条件をそろえて比較する。

Hinton et al. (2012)

自己教師あり学習とは

文字起こしの作成には人手が必要。利用できる量や作業時間は、言語・分野・音質などに依存する。

自己教師あり学習：文字起こしのない音声から、入力自体を用いて学習目標を作る。

段階	データと目的
事前学習	大量の文字起こしなし音声から表現を学ぶ
ASRへの追加学習	文字起こし付き音声で対象タスクを学ぶ
評価	学習に用いなかったデータで性能を測る

Masked Prediction（マスク予測）

- BERT：一部のトークンを隠すなどして、そのトークンを予測
- wav2vec 2.0・HuBERT：潜在特徴量をマスクし、対応する量子化表現やクラスタIDを予測

単純な次の音の予測や、元波形の復元そのものではない。

wav2vec 2.0 のアーキテクチャ

波形 → CNN特徴量z。原論文では16 kHz入力から約20 ms間隔、25 msの受容野で特徴量を出す。

経路	処理と出力
量子化側	マスク前のzから、コードブックを使って量子化ベクトルqを作る
文脈側	一部のzをマスクし、Transformerで文脈表現cを作る
比較	マスク位置のcに対応するqを、負例から識別する

$$\mathcal{L}_t = -\log \frac{\exp(\mathrm{sim}(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \in Q_t} \exp(\mathrm{sim}(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

Q_t：正例と、同じ発話の別のマスク位置から選ぶ負例。simはコサイン類似度、κは温度。

コードは正解の音素ラベルではない。コードブックの多様性損失も加える。Baevski et al. (2020)

HuBERT：オフラインクラスタリングによる学習

核心：クラスタIDを擬似ラベルとして作り、マスク位置で予測する。

段階	処理
最初の目標作成	MFCCにk-means → クラスタID
表現学習	マスク位置のIDを分類するようモデルを学習
目標の更新	学習した特徴量に再びk-means → 更新したID

項目	wav2vec 2.0	HuBERT
離散化	学習可能な量子化器	オフラインk-means
主な目標	量子化ベクトルを識別	クラスタIDを分類
損失	コントラスト＋多様性など	マスク位置のクロスエントロピー

クラスタIDは正解の音素ラベルではない。反復するだけで必ず改善するわけではない。Hsu et al. (2021)

ラベル付きデータ量と性能：条件をそろえて比較

wav2vec 2.0：**BASE・LS-960で事前学習・4-gram LMで復号**。LibriSpeechテストセットのWER（%）。

追加学習のラベル付き音声	test-clean	test-other
10分	9.1	15.6
1時間	5.5	11.3
10時間	4.3	9.5
100時間	3.4	8.0

Baevski et al. (2020), Table 1の同一条件の行を比較。

- 10分は**追加学習に使う文字起こし付き音声**の量
- 事前学習には960時間の音声、言語モデルには別途テキストを使う
- 英語読み上げ音声での値。別言語・雑音条件に同じ性能を保証しない

Google Forms テスト（全5問）

締め切り：翌週末 23:59

1. **HMMの3要素**：初期確率・遷移確率・出力分布の正しい説明の組み合わせはどれか
2. **Viterbi**：最も確率の高い状態系列を動的計画法で求めるアルゴリズムの説明はどれか
3. **GMM**：「成分数Kを大きくすれば常に汎化性能が向上する」は正しいか
4. **wav2vec 2.0**：量子化ベクトルを正例とするコントラスト学習の説明はどれか
5. **自己教師あり学習**：紹介した原論文で、事前学習と文字起こし付きの追加学習をどう組み合わせるか

まとめ

概念	ポイント
従来型ASR	音響モデル・発音辞書・言語モデルをデコーダで統合
HMM	初期確率・遷移確率・出力分布の3要素
Viterbi	最良状態系列を動的計画法で求める
GMM	複数ガウス分布の重み付き和で確率密度を表す
DNN-HMM	DNNの状態事後確率をHMMの系列構造と組み合わせる
wav2vec 2.0	量子化表現をコントラスト学習の目標にする
HuBERT	オフラインのクラスタIDをマスク予測する

次回（第8回）：ConformerとWhisperを例に、ニューラル音声認識の構成と使い方を学ぶ。

Google Forms テスト

第7回の理解度テストを受けてください

- 全5問・各1点
- 締め切り：**翌週末 23:59**
- Google Formsのリンクはメールまたは授業ページで確認してください

お疲れさまでした！