知的システム構成論課題
矢入先生担当分

航空宇宙工学専攻修士一年
37-186305 荒居秀尚

2018 年 8 月 10 日

1　選択したテーマ

今回の課題に向けて選択したテーマは「隠れマルコフモデルと EM アルゴリズムを用いた音声データの潜在状態の推定」としました。音声データとしては、研究室の研究会の様子を録音したもの、約 160 分程度のものの一部を使用し、潜在状態として話者を分類することを目的として学習を行いました。今回のデータはサイズが大きかったのとノイズが多い区間が多かったため実際のモデルの学習には先述のデータから 10 秒程度を切り出したものを用い、その検証にも別の区間から 10 秒程度切り出したものを使用しました。なお、学習用データに含まれるのは 3 人分の音声であり、検証用データに含まれるのは学習用データに含まれていた声のうち 2 人分でそれ以外の声は含まれていません。このことを考慮して、今回はクラスタ数を 3 として事前に与えました。

このテーマを選択した理由として、個人的に以前から研究会の文字起こしをしたいと考えており、特に話者を識別してアノテーションを加えたいという目的がありました。今回はそれを HMM を用いて出来ないか試してみたという例になります。

2　学習手順

音声データを扱うにあたって、メル周波数ケプストラム係数 (MFCC) に変換を行いました [1]。MFCC に変換する手順は本質的ではありませんが学習を行うにあたって問題の原因となっていると感じたため記載します。

MFCC に変換する手順は以下の 6 ステップからなります。

1. 広域強調
2. 音声フレーム
3. STFT
4. メルフィルタバンクの作成
5. ケプストラム特徴量の抽出

6. リフタリング

広域強調は人間の聴覚特性に合わせて広域のパワーを補償する作業です。その後ハミング窓などの窓関数を使って音声フレームを切り出し、短時間フーリエ変換を行います。その後メル尺度に合わせた特徴量セットを作成して、それをフーリエ変換後のデータに適用して逆離散コサイン変換をかけることでケプストラム特徴量を得ます。最後にこのケプストラム特徴量のうち目的に合わせて特徴量を選択します。

今回は最後のリフタリングにおいて 20 次元までに特徴量セットを絞りました。これを GaussianHMM に EM アルゴリズムを用いて学習させることで、クラスタリングを行います。

3　結果・考察

今回は時間がなかったため、アルゴリズム等のスクラッチ実装は行わず、ライブラリ関数を用いただけになってしまったのですが上記の手順を実装しました。MFCC の抽出は librosa[2] を用い、GaussianHMM は hmmlearn[3], と bnpy[4] を試しました。結果としては、検証用データには二人分の声しか含まれていないにもかかわらずクラスタリングのうえでは 3 クラスへ分類が行われていたため、おそらく発言した人物ごとのクラスタリングは失敗していると考えられます。以下ではこの原因に関して考察を行います。

まず、MFCC の可視化ですが、以下の図 3.1 のようになりました。 下から順に、MFCC の 1 番目の成分か

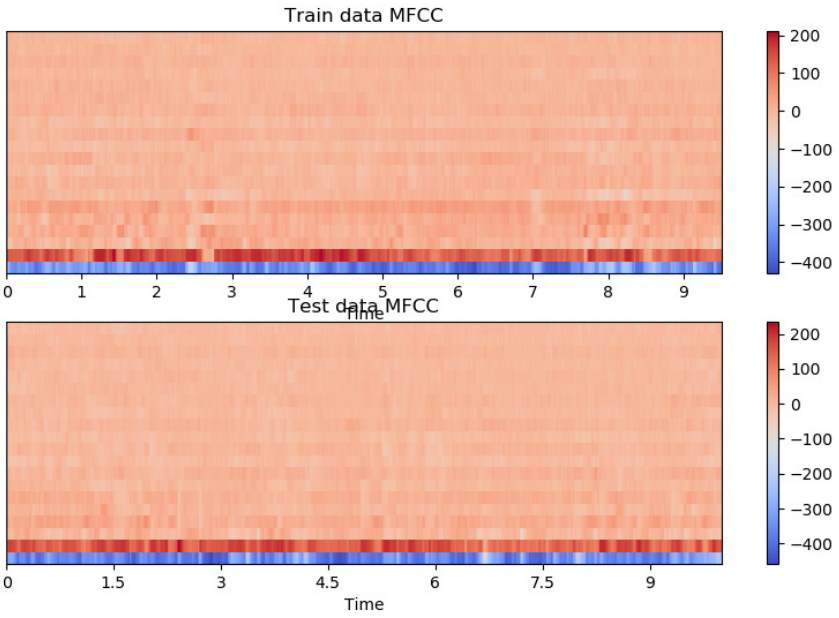

図 3.1: Train データと Test データを MFCC に変換した結果

ら 20 番目の成分が並んでいます。全体として音声パワーが学習データのほうが強く検証データのほうが弱いものになっていたのですがその影響がやや現れてしまっているように見られます。具体的には学習データのほ

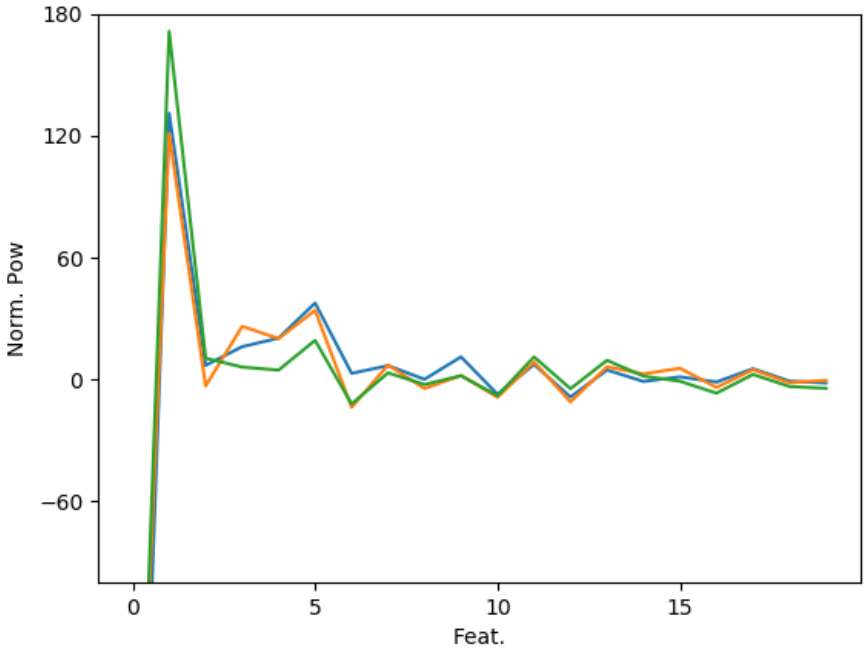

図 3.2: Train データのクラスタごとのパワー

うが、2.5 秒付近と 8 秒付近に強い発話があったことに対応する縦に伸びる構造が見えますが、検証データにはそれほど大きな同様の構造は見られません。また、第 6 特徴などは学習データでははっきりとした横に伸びる筋が見られますが、検証データではぼんやりしています。

また、bnpy を用いて学習データに対するクラスタリングの結果をクラスタごとのパワーとして表したのが以下の図 3.2 です。この図を見ると第 2 から第 6 特徴量に当たる部分が分類に大きく寄与していることがわかります。しかし、個人差が大きいと言われるのは高ケフレンシー領域、すなわち図 3.2 の左の方の特徴に集中するため、この部分の違いをクラスタリングに反映出来ていないことがわかります。この対策は、低ケフレンシー領域を捨てる、ということになりますが今回は時間が足らず MFCC を自分で実装することが出来なかったためこれは出来ませんでした。

また、モデルの面でも今回は、ライブラリを用いるのみとなってしまったため、遷移確率などに制約をかけられなかったことが 1 つ失敗の原因と考えられます。今回は学習データが 10 秒程度と短いため、正確な推論には的確に初期条件を与える必要があると考えられますが、この部分をしっかりと行わなかったことも原因の 1 つと考えられます。

今後は、この結果を考慮して

- MFCC において低ケフレンシー領域の特徴を捨てる

- モデルの初期値に対して制約を入れられるようにする

といった方向性で音声を人ごとに分離することができると考えられます。

参考文献

[1] 篠田浩一. 音声認識. 機械学習プロフェッショナルシリーズ. 講談社, 2017.
[2] Librosa. http://librosa.github.io/librosa/.
[3] hmmlearn. http://hmmlearn.readthedocs.io/en/latest/.
[4] bnpy. https://bnpy.readthedocs.io/en/latest/.