しりとりAI

計数工学科数理情報工学コース B4 久保優騎

1. キーワード

- 音声認識
- 音響モデル
- 言語モデル

2. 理論解説

2.1. 音声認識とは

音声認識とは発話された言語音を入力として対応する言語 (単語，文等) の文字列を出力する技術である．

音声認識の根幹をなすのは音響モデルと言語モデルであり，この 2 つを組み合わせることで音声を認識する．というのも，音響モデルのみに従って音声を認識した場合に尤もらしい文字列が出力されたとしても，言語として成り立っていない場合が考えられるからである．音響的にも言語的にも尤もらしい文字列を出力することが必要であり，発話されている文字列は実際その 2 つの観点において最も尤もらしいはずである．

2.2. 音響モデル

音源信号 $x(t)$ に長さが一定の窓関数 $w(t)$ をずらしながら掛けてフーリエ変換 (短時間フーリエ変換) することで得られるものを，信号のスペクトログラムという．すなわち，

$$STFT_{x,w}(t,\omega) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-i\omega\tau}d\tau$$

で定められる関数 $STFT_{x,w}(t,\omega)$ のことである．スペクトログラムは 2 変数関数であるので, その値の大きさを色で示せば例えば図 1 のようになる．これは"あいうえお"と発話した音声のスペクトログラムである．

また，音声信号をフーリエ変換したものをその音声のスペクトルという．図 2 は上の音声のうち，"あ"に対応する部分のみを切り出した信号のスペクトルである．ただしここでは信号の両端が連続でないのでハニング窓という窓関数を掛けてから処理している．

よくスペクトルの対数をとったもの (対数スペクトルという) を用いるので，以下でも対数スペクトルを用いる．対数スペクトルは包絡と微細構造に分けられる．微細構造は対数スペクトルのうち周期の短い波の部分であり，ピッチを表す．一方で包絡は対数スペクトルから微細構造を除いた長い周期の部分であり，(話者に固有な) 声道の特徴を持っている．スペクトルにおいてはこの 2 つは乗算されていたが，対数を取ることで加算になり扱いやすくなる．図 3 に対数スペクトルとその包絡の一例を示す．

図 1　"あいうえお"のスペクトログラム

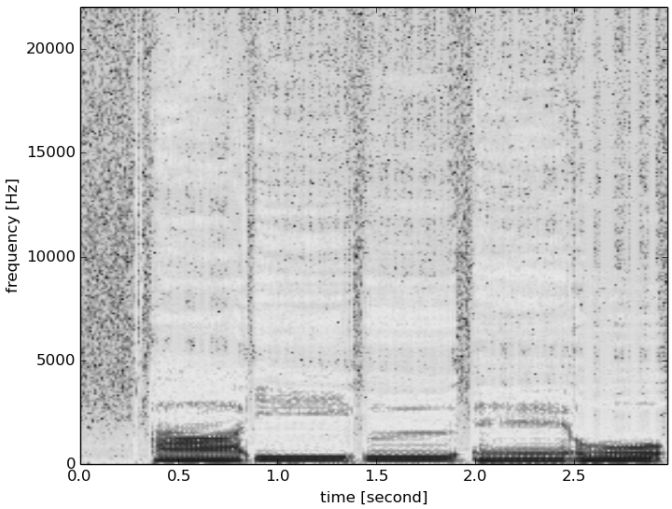
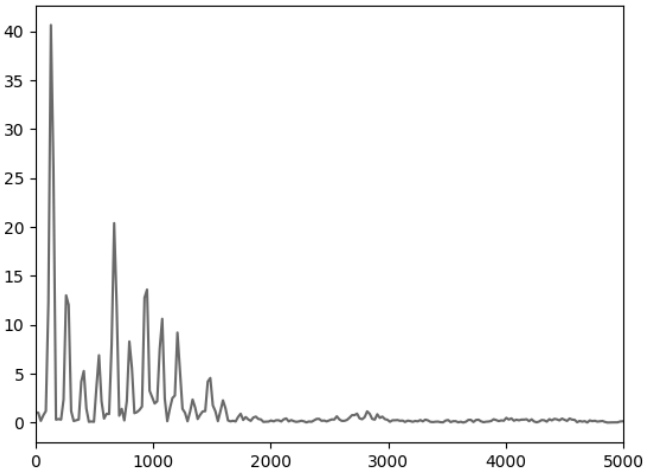


図 2　"あ"のスペクトル



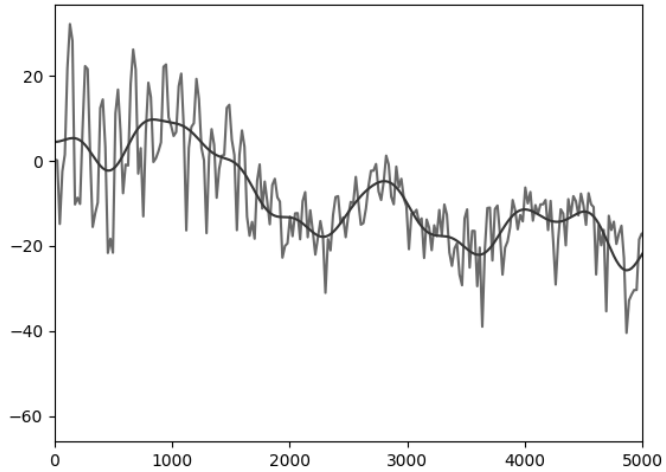
このスペクトル包絡から発話された音素の特徴量を取得し，それに対応する音素を対応付けることで音響モデル的な音声認識が行われる．特徴量の抽出においては隠れマルコフモデルを用いた方法などがあるが，ここでは割愛する．

2.3. 言語モデル

与えられた音声信号の音素と合致するような語の組み合わせ (=文) の数は膨大であるので，何らかの制約を入れて計算量を小さくする必要がある．

例えば，入力言語の文法を規定することで単語の繋がりに制約を入れることが出来る．"僕はギター"に続く

図 3　"あ"の対数スペクトルと包絡



単語としては助詞"が"や名詞"ケース"などはあり得る ("僕はギターが好きだ","僕はギターケースを買った"など) が，動詞"弾く"や形容詞"美しい"は繋がらない，などである．

文法の規定のみでは処理が難しい場合も多い．名詞"ギター"の後に名詞"鑑賞"がきて複合名詞を形成することはあるが，名詞"トマト"の後に名詞"フォーク"が来ることはまずあり得ない．このような場合を上手く扱うモデルが N-gram モデルである．これはある単語の生起確率が直前の $N-1$ 単語のみに依存するというモデルであり，マルコフモデルである．すなわち，単語の列 $w_1, w_2, \ldots, w_n$ に対して，その単語の列が並ぶ確率を

$$P(w_1, w_2, \ldots, w_n) = \prod_i P(w_i | w_{i-1}, w_{i-2}, \ldots, w_{i-N+1})$$

で表す．また余談ではあるが，文法構造が似通った言語間ではこのモデルを用いれば高精度な翻訳が出来る．[2]

2.4. まとめ

音声認識を行うには音声モデル・言語モデルを用いて，ただの波形から分離された音素を得て，元の言葉に対応していて言語的制約を満たす文を構成するという行程を踏まねばならない．そのためには事前に膨大なデータが必要となるため，システム全体を自作することは難しいが，一方で人的・金銭的資産や時間などが十分にある研究開発において開発された技術は高性能を誇り，4 章に挙げるように様々な応用例が存在する．

3. 実験・実装への応用

2.3 節で述べたような言語モデルを用いるには膨大なテキストのコーパス (文章や単語を品詞や統語構造といった文法的要素も含めて集めたデータベース) や音声のコーパスを用意する必要がある．時間的・労力的な制約もあり準備するに至らなかったので，本展示では docomo Developer support が提供する音声認識 API を用いて音声認識を行っている．

本展示のプログラムは入力された音声を API を用いて認識し，認識された文字列の末尾の文字 (拗音や促音，長音は適宜処理する) を語頭に持つ単語を返すという単純な仕様である．

4. 応用例

- Siri
  Apple 社の製品 iPhone,iPad,iPod touch に搭載されている音声対話ソフトウェア．
- Vocollect
  Vocollect 社が開発した音声物流業務支援システム．例えば倉庫におけるピッキングにおいて，システムからの音声指示を聞いて作業して回答するという一連の流れで作業を進めることが出来る．
- MMDAgent
  音声インタラクションシステム構築ツールキットであり，実用例として双方向音声案内システムが名古屋工業大学にある．[3]
- PVI
  音声を用いたパーキンソン病判別技術．集められたパーキンソン病発症に伴う発音障害の音声データを用いて，被験者の発話音声を入力としてパーキンソン病の重症度を予測する．[4]

参考文献

[1] Tom Hakamata. 自称・世界一わかりやすい音声認識入門, https://www.slideshare.net/c5tom/ss-56184353

[2] Takuya Nishimura. 対訳文対を用いた日英パターン翻訳器の自動作成法の検討, http://unicorn.ike.tottori-u.ac.jp/2009/s062044/paper/thesis/shuron.pdf

[3] メイちゃんとは？｜メイちゃん 公式ウェブサイト, http://mei.web.nitech.ac.jp/?page_id=12649

[4] *Parkinson's Voice Initiative*, http://www.parkinsonsvoice.org/science.php