Deep Learning を用いた シャドーイング音声の自動評価

峯松信明，楽俊偉，齋藤大輔（東大）
山内豊（東京国際大），伊藤佳世子（京大）
mailto: mine@gavo.t.u-tokyo.ac.jp

本発表の流れ

- はじめに
- シャドーイングコーパスの構築
- シャドーイング音声の手動評価
- Deep Learning を用いた自動評価
  - 学習者音声と（モデル話者が読み上げた）テキストとの比較
- Deep Learning + DTW を用いた自動評価
  - 学習者音声とモデル音声との比較
- まとめと今後の課題

はじめに

- 伊藤科研
  - 自律的な英語シャドーイング学習を目指した自動評価と教材データベースの開発研究
    - 代表者：伊藤佳世子（京大国際高等教育院）
    - 2016年度～2019年度
    - goo.gl/tM22xl
- 山内科研
  - 多言語に対応できるシャドーイング自動評価システムの開発と外国語教育への応用研究
    - 代表者：山内豊（東京国際大学商学部）
    - 2016年度～2019年度
    - goo.gl/QlyJie
- シャドーイング音声の自動評価技術の構築
  - 峯松・齋藤研究室（東京大学大学院工学系研究科）

シャドーイング音声の収録

- 複数のサイトで行われるシャドーイングを一括収録
  - 各文のシャドーイング直後にサーバー（東大）に転送させる
  - 聴取した音声がマイクに混入しないような工夫
  - 教室環境では周りの学生の音声が混入しないような工夫
  - 分かりやすいインタフェースでの収録

収録音声とその手動評価

- 3大学，合計125名の学生の音声を収録
  - 4パッセージ，55文のシャドーイング
  - 4回のシャドーイング，合計，27,500発声を収録
- 手動評価のための文選定
  - 文の複雑さ，発音の難しさを考慮し10文を選定
    - 各文は2〜3の節により構成，10文＝27節＝合計3,375節
  - 4回目のシャドー音声を手動評価の対象に。節単位での評価
  - 評価の着眼点
    - 音素の生成が適切に行なわれているか（音素）？
    - 韻律の生成が適切に行なわれているか（韻律）？
    - 各単語を語の同定を伴って発声しているか（正確さ）？
    - 5段階評価（1〜5），合計すると3〜15点
  - 評価者
    - 発音教育を実践する博士学生（教育学専攻，日米のハーフ）

手動評価結果の統計

- 各評価尺度の統計
  - 3,375 節の平均と標準偏差

	音素	韻律	正確さ
平均	1.9	4.2	4.1
標準偏差	0.62	0.57	0.54

- 節の位置に基づく統計
  - 節位置＝第一，第二，第三

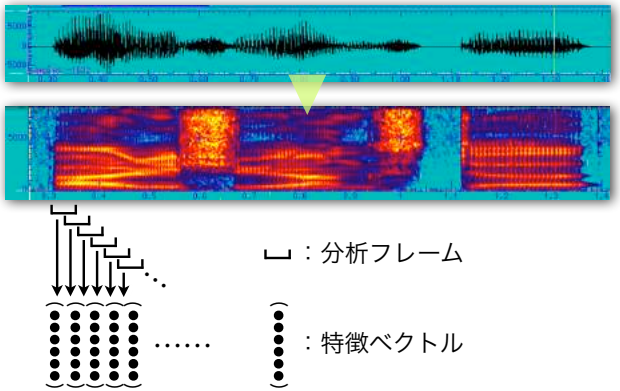
	第一	第二	第三
平均	10.5	9.8	10.2
標準偏差	1.3	1.7	1.9

本発表の流れ

- はじめに
- シャドーイングコーパスの構築
- シャドーイング音声の手動評価
- Deep Learning を用いた自動評価
  - 学習者音声と（モデル話者が読み上げた）テキストとの比較
- Deep Learning + DTW を用いた自動評価
  - 学習者音声とモデル音声との比較
- まとめ

本題に行く前に

- 音声の分析について
  - 音声 → 声紋パターン → フレーム分割 → 特徴ベクトル時系列



  - フレームのずらし＝10msec ほど，１音節＝100〜200msec
    - １音節あたり，10〜20個の特徴ベクトル系列へ

本題に行く前に

事後確率（条件付き確率）について

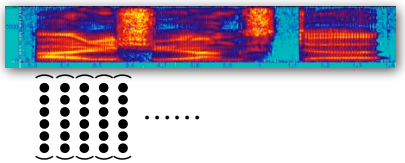
- 「NYCはどんな天気ですか？晴？曇？雨？」
  - 「いつの天気のことを聞いているのですか？」「教えない。」
  - 一年を通して NYC は 晴：曇：雨＝3：2：1 くらかな？
  - 事前に持っている知識に基づく確率：事前確率
    - P(天気=晴) = 3/6，P(天気=雨) = 1/6
    - P(天気=晴) + (天気=曇) + (天気=雨) = 1

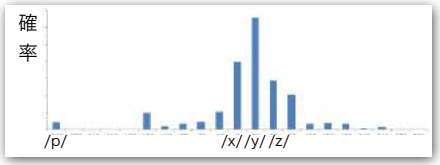
- 「教えて欲しいのは 4/1 の天気です」
  - 「4月上旬の NYC は，よく晴れるよなぁ」
  - 晴：曇：雨 = 7：2：1 くらいかな？
  - ヒント，場面設定，条件づけがされた確率：事後確率（条件付確率）
    - P(天気=晴 | 日付=4/1) = 7/10，P(天気=雨 | 日付=4/1) = 1/10

本題に行く前に

音素事後確率（条件付き確率）について

- 「ある人が喋りました。ある時刻の音素は何でしょう？」
  - P(音素=/a/) = ?，P(音素=/k/) = ?
  - そもそも，人の喋りの中に，/a/ と /k/，どっちが多いかな？
- 「その時刻の特徴ベクトルはこれでした。」
  - P(音素=/a/ | 特ベ=⋮) = ?，P(音素=/k/ | 特ベ=⋮) = ?
  - 音の様子（特徴ベクトル）が分かれば，音素は一意に決まるのでは？
    - 物理的に同じ音であっても，
      話者Aの声なら /i/ になり，話者Bの声なら /e/ になる。
    - 結局，音が与えられても，どの音素かは確率的にしか議論できない
  - 特徴ベクトルが与えられた時の，音素事後確率 = GOP

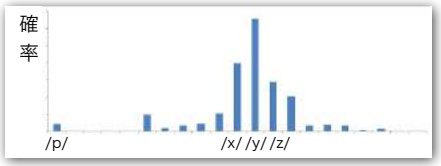




本題に行く前に

特徴ベクトルが与えられた時の音素事後確率

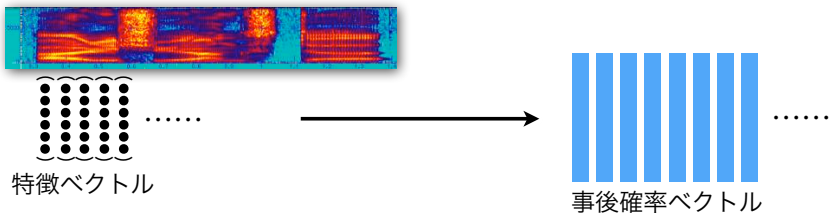
- 音素の数 = 数十
- その確率値を並べれば
  数十次元のベクトルになる
  - 事後確率ベクトル



音素から音（素）クラスへ

- 音素 = ある言語の母語話者がもつ言語音の種類数
  - ある音素の音響的実現は年齢・性別によって異なる
  - aiueo は音色が連続的に変化する。母音の中間音というのが存在
    - beat it の it は「い」と「え」の中間音 ↔ いえいえ・・
- 種類数を拡大する。人の違いよる変化，気付かない中間音も考慮
  - 言語学が定義する音素数：数十
  - 音声工学が考える音（素）クラス数：数百〜数千
    - 特べが与えられた時の音クラスの事後確率ベクトル：数百〜千次元

本題に行く前に



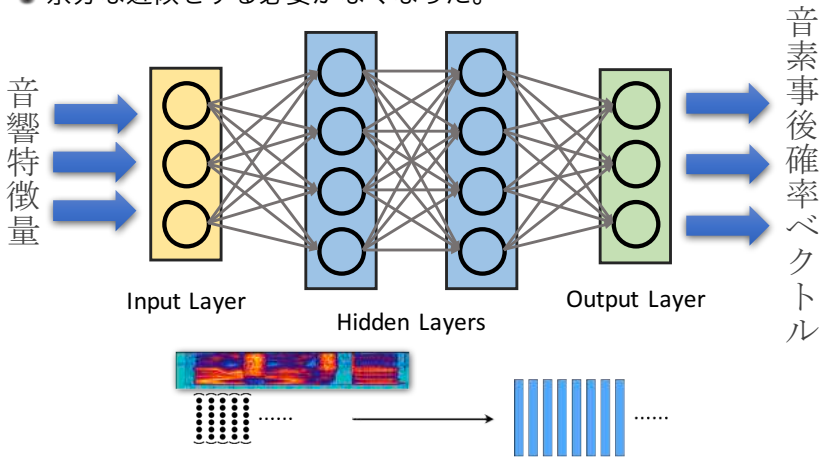
本発表の流れ

- はじめに
- シャドーイングコーパスの構築
- シャドーイング音声の手動評価
- Deep Learning を用いた自動評価
  - 学習者音声と（モデル話者が読み上げた）テキストとの比較
- Deep Learning + DTW を用いた自動評価
  - 学習者音声とモデル音声との比較
- まとめと今後の課題

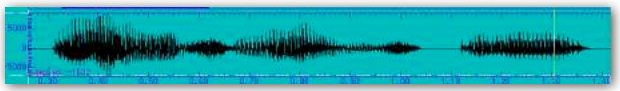
DL（深層学習）と事後確率

Deep Neural Network を用いた事後確率計算

- ここ数年の音声認識精度の向上の主要因はこれ
- HMM（隠れマルコフモデル）→ DNN
  - 余分な近似をする必要がなくなった。



HMM時代の技術も利用して



HMM
+意図した音素列

Frame	Phoneme
1	a
2	a
3	u
...	...
1232	sil

DNN

音(素)クラス

時間

Frame	sil	a	i	u	...
1	0.01	0.8	0.1	0.02	...
2	0.01	0.7	0.1	0.1	...
3	0.01	0.5	0	0.4	...
...	...	...	...	...	...
1232	0.9	0	0.01	0	...

$$\mathrm{GOP} = \frac{0.8 + 0.7 + 0.4 + \cdots + 0.9}{1232} = 0.63$$

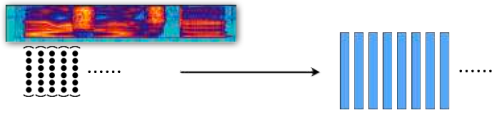
DNN-GOP

（モデル音声そのものは使っていない）

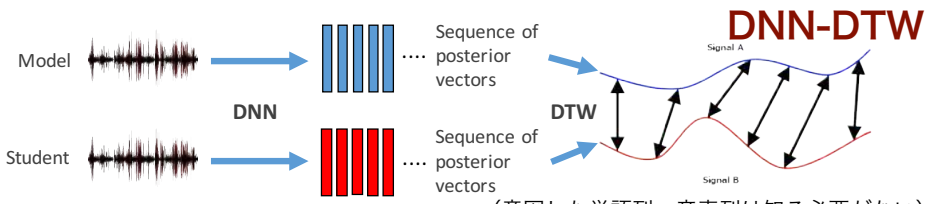
もう一つのスコア計算法

学習者音声・モデル音声を比較する

- 2つの長さの異なる時系列の対応をとりながら比較 ＝ DTW
  - Dynamic Time Warping 法
- 両方の音声をまず事後確率ベクトル系列にする



- DTW法で両者を比較する
  - 特徴ベクトル系列で比較すると，性別・年齢の影響が出る。



（意図した単語列・音素列は知る必要がない）

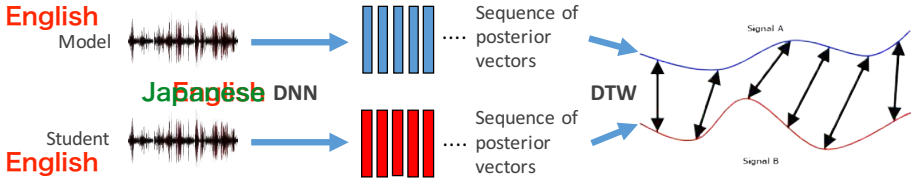
二手法の比較

何が必要で何が必要でないのか？

	学習者音声	モデル音声	音素列
DNN-GOP	必要	**不要**	必要
DNN-DTW	必要	必要	**不要**

音素事後確率 → 音(素)クラス事後確率

- beat it ⇄ いえいえ・・・
- 日本語の音(素)クラスの中に英語も含まれる？



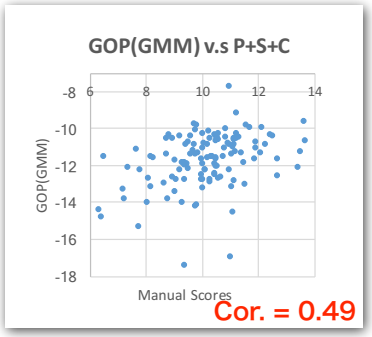
- 学習対象言語と事後確率化の言語は揃える必要はない!?
- 山内科研「多言語に対応できるシャドーイング・・・・」

DNN-GOPと手動スコア

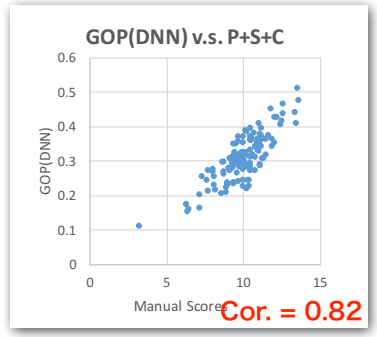
話者単位スコア

- 27節手動スコアの平均値 → 話者単位での手動スコア
  - 3尺度の合計値，3～15
- 27節自動スコアの平均値 → 話者単位での自動スコア

HMMによるGOP計算



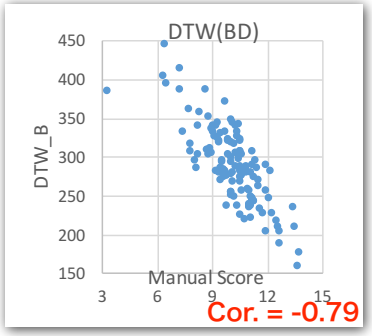
DNNによるGOP計算

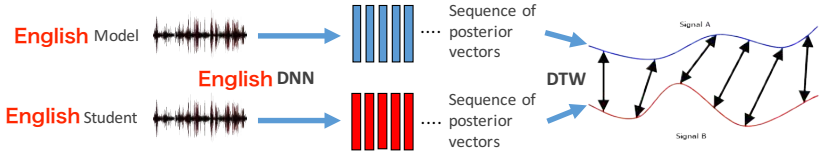


DNN-DTWと手動スコア

話者単位スコア

- 学習者・モデル発話のDTW距離平均 → 話者単位での自動スコア
  - クラス数：3384

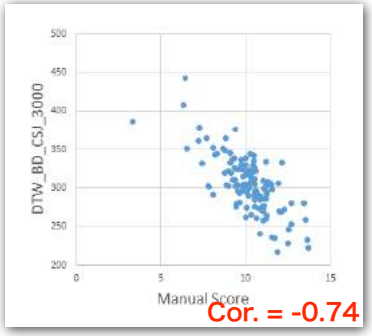


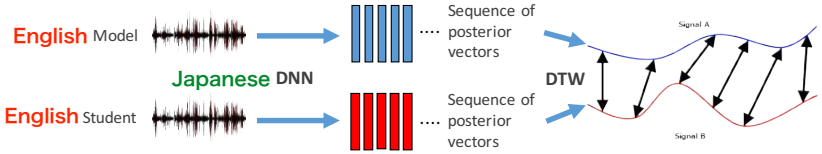


DNN-DTWと手動スコア

話者単位スコア

- 学習者・モデル発話のDTW距離平均 → 話者単位での自動スコア
  - クラス数：2856





本発表の流れ

- はじめに
- シャドーイングコーパスの構築
- シャドーイング音声の手動評価
- Deep Learning を用いた自動評価
  - 学習者音声と（モデル話者が読み上げた）テキストとの比較
- Deep Learning + DTW を用いた自動評価
  - 学習者音声とモデル音声との比較
- まとめと今後の課題

まとめと今後の課題

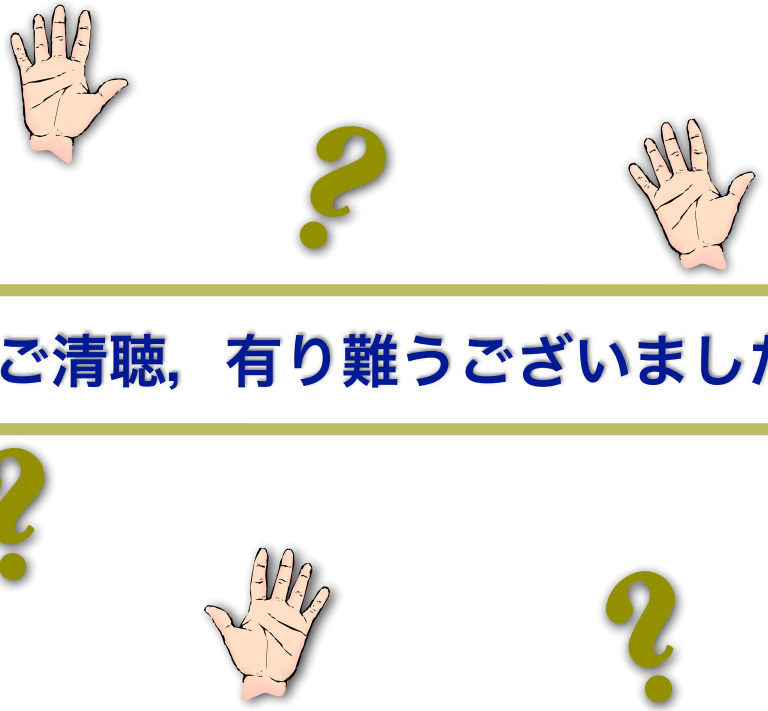
- シャドーイングコーパスの構築
- シャドーイング音声の手動評価
- DNN-GOPとDNN-DTWによる自動評価
  - 学習言語と事後確率化言語は違っても良い!?
- 今後の課題
  - クラス数の最適化
  - 10文／50文で評価（事前選択）　評価に適した10文の自動選出
  - 読み上げ音声 → 感情のこもった音声へ
    - DNN-GOP：モデル話者の意図した音素列と比較
    - DNN-DTW：モデル音声そのものと比較
    - **モデル音声が読み上げでも，感情音声でも，前者は比較対象が同じ**
  - 評価単位をより細かく（人 → 文 → 単語 → 音素）
    - 誤り音声区間の自動検出タスクへ

二つ告知させてください

- 峯松・齋藤研オープンラボ
  - 毎年五月祭の時に，オープンラボを行なっています。
  - 今年は**5月20日, 21日(土, 日)**です。是非お越しください。
- 人文系大学院生向け「音響音声学１・２」
  - 通年・**毎週水曜2限（10:25〜12:10）**
  - 授業 web：goo.gl/Ly9ZpH　シラバスは下記
  - 毎年，外部の人が数名（無料で）受講しています
    - 語学教師，言語聴覚士，声優，ボイストレーナー，言語障害児の母親，保育士，など音声に興味はあるが物理を学んだことがない方々
    - 「共同研究上必要」と言えば，事務方もOKするはず・・・・

希望者はまずメールを!!
メアドはスライド1枚目

本授業では高校で物理を履修しなかった学生を対象に，音声の物理的・音響的側面について分かり易く解説する。音声は音，即ち，空気（酸素・窒素・二酸化炭素など）の振動現象でしかない。しかし，その振動現象を鼓膜が捉えると，言語メッセージ，意図，感情，更には話者の健康状態など，様々な情報を我々は知覚できる。一体，空気振動のどこにこれらの豊富な情報が隠れているのだろうか？
音響音声学（1）では，音の基礎物理から始め，音声を音響的に眺めるために必要な基礎知識を提供すると共に，音刺激に対するインタフェースである聴覚の処理についても学ぶ。



ご清聴，有り難うございました