論文 / 著書情報
Article / Book Information

論題(和文)	能動的な適応文選択に基づく話者適応化
Title(English)	
著者(和文)	村上 博子, 篠田 浩一, 古井貞熙
Authors(English)	Hiroko Murakami, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2009年春季講演論文集, Vol. , No. , pp. 191-194
Citation(English)	, Vol. , No. , pp. 191-194
発行日 / Pub. date	2009, 3

能動的な適応文選択に基づく話者適応化*

◎村上博子, 篠田浩一, 古井貞熙 (東工大)

1 はじめに

話者適応化技術は, 音声認識システムにおいて新しい話者の声質に合わせて音響モデルのパラメータを変更し, 認識性能を改善させる技術である. 話者への負担を減らすためにも, できる限り少量の発声による認識性能向上が望まれる.

話者適応化において, 適応データを選択することにより, その性能改善が期待できる. 関連研究として, 例えば, 話者適応化における学習語彙依存性についての研究 [1] やタスクが異なる場合の適応語彙の能動的選択の研究 [2] がある.

話者ごとに話し方に特徴があり, 認識精度の低い音素は話者によって異なるため, 同じ文を発声し適応しても, 同じように認識結果が改善されるとは限らない. すなわち, 話者ごとに適応の効果が高い適応文は異なる可能性がある. 同量の文を適応に用いるならば, 例えばその話者にとって認識精度の低い音素がより多く含まれる文を用いた方がより高い効果が得られると期待できる.

本稿では音声認識システムが話者の少量の発声から性能向上に役立つ情報を能動的に引き出すことによって, より効率的に話者適応を行う手法を提案する.

2 能動的な適応文選択

2.1 概要

教師あり適応の初期段階において, 少量の適応発声から話者の認識精度の低い音素を推定する. そして, 予め用意した適応文の候補からそれらの音素が多く含まれる文を選択し, 話者にその文の発声を促し適応を行う. Fig. 1 に提案手法の全体的な流れを示す.

2.2 音素頻度分布と音素誤り分布の計算

まず, 話者の少量の発声 (初期適応データ) を用いて音素誤り分布を求める. 不特定話者モデル (SI モデル) を用いて初期適応データに対し音素認識を行い, 各音素の認識精度を得る. 音素 u_i の誤認識精度を $r(u_i)$ とすると, 全音素にわたる

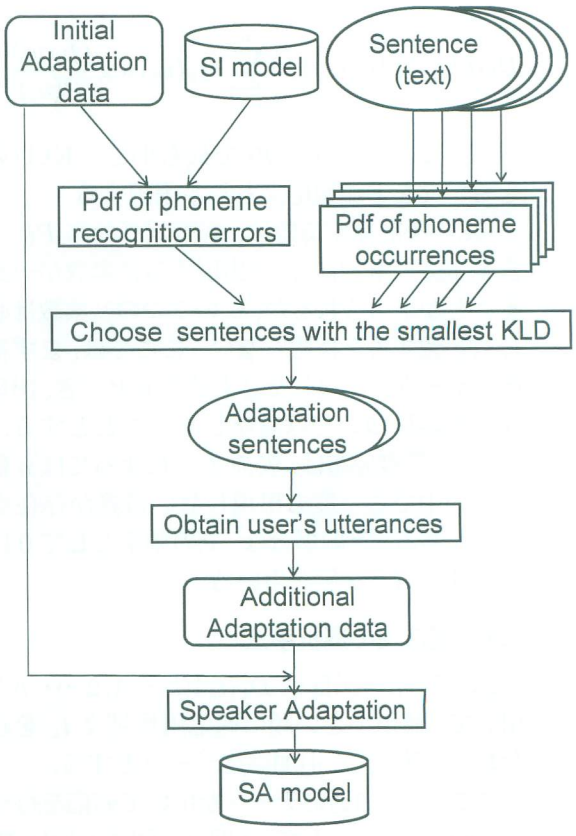

Fig. 1 提案手法の全体的な流れ

音素誤り分布 $P(u_i)$ は以下のようになる.

$$P(u_i) = \frac{r(u_i)}{\sum_{j=1}^{n} r(u_j)}, \quad i = 1, \cdots, n \qquad (1)$$

ここで n は音素数である.

次に, 適応に使用する文の候補である適応用候補文セットを予め用意し, セット中の各文 t_k $(k = 1, 2, \cdots, N)$ の音素頻度分布を求める. $s_k(u_i)$ を候補文 t_k に含まれる音素 u_i の出現回数とすると, 候補文 t_k の音素頻度分布は以下のようになる.

$$Q_k(u_i) = \frac{s_k(u_i)}{\sum_{j=1}^{n} s_k(u_j)}, \quad i = 1, \cdots, n \qquad (2)$$

2.3 カルバック・ライブラー情報量

候補文セットからの文選択の基準として, 音素誤り分布と音素頻度分布の間のカルバック・ライブ

*An active approach to speaker adaptation based on adaptation data selection By Hiroko Murakami, Koichi Shinoda and Sadaoki Furui (Tokyo Institute of Technology)

ラー情報量 (Kullback-Leibler divergence, KLD) を使用する. 2.1 で求めた音素誤り分布 $P(u_i)$ と音素頻度分布 $Q_k(u_i)$ の間の KLD は以下のように定義される.

$$D_k(Q_k(u_i)||P(u_i)) = \sum_{i=1}^{n} Q_k(u_i) \log \frac{Q_k(u_i)}{P(u_i)} \quad (3)$$

候補文セット内の文の中で最も小さい KLD の値をもつ文を追加適応文として選択する.

KLD を計算する際に, 音素誤り分布 $P(u_i)$ と音素頻度分布 $Q_k(u_i)$ で出現する音素数が一致することが求められるが, この 2 つの音素数は必ずしも一致するとは限らない. ここでは, まず初期適応データで出現した音素数を n とおき, 出現しない音素については考慮しないこととする. また, 音素頻度分布は, 候補文 t_k によっては n 個の音素の中でも一度も出現しない音素が存在することがある. その場合は, その確率として 0 に近いクリッピング値を用いる.

2.4 適応文の選択手法

2.3 で得た KLD 値 $D_k(t_k)(k = 1, 2, \cdots, N)$ を用いて, 値が小さい順に適応用候補文 t_k を必要な数だけ選択し, 追加適応データとする.

ここで, 追加適応データを用いて適応を行う際に, 適応の方法として, 初期モデルを固定して適応データの適応処理を一度に行うバッチ適応と, 1 文ずつ適応データの適応処理を繰り返す逐次適応が考えられる.

バッチ適応では, まず SI モデルを用いて初期適応データを認識し, 音素誤り分布を求める. 次に候補文から追加適応文を 1 度に必要な数だけ選択し, 話者の発声から追加適応データを得る. 最後に, SI モデルを初期モデルとして, 初期適応データと追加適応データを併せたデータを用いて適応を行う.

逐次適応では, まず SI モデルを初期モデルとして初期適応データを用いて適応を行い, 適応後のモデルを用いて, 初期適応データを認識し, 音素誤り分布を求める. 次に候補文から追加適応文を 1 文選択し, 話者に発声を促し, そのデータを用いて適応を行う. そして 1 文発声が追加されるごとに適応モデルで音素誤り分布を求めなおし, 追加適応データの総数が必要な数に達するまで適応を繰り返す.

今回は比較的処理が簡単なバッチ適応を用いた.

3 MLLR 法

最ゆう回帰 (Maximum Likelihood Linear Regression, MLLR) 法 [3] は, 音響特徴量空間における話者間の線形写像を用いる手法である.

この手法では, 以下に示す変換により, HMM のガウス分布の平均ベクトル $\mu = (\mu_1, \cdots, \mu_n)'$ が更新される. n は特徴ベクトルの次元数である.

$$\hat{\mu} = A\mu + b \quad (4)$$

A は $n \times n$ の行列, b は次元数 n のベクトルである.

行列 A, ベクトル b は, EM アルゴリズムによる最ゆう推定により求められる.

4 認識実験

4.1 実験条件

データベースとして, 新聞記事読み上げ音声コーパス (JNAS)[4], 高齢者の音声認識用大規模データベース (S-JNAS)[5] を用いた. JNAS は各話者, 新聞記事を約 100 文の他に音素バランス文 50 文の計約 150 文を読み上げている. S-JNAS は各話者, 新聞記事を約 100 文の他に音素バランス文 100 文の計約 200 文を読み上げている. 実験には, JNAS の 222 話者 (男女各 111 話者), S-JNAS の 300 話者 (男女各 150 話者) の計 522 話者を学習データとして用い, JNAS の 44 話者 (男女各 22 話者) を評価データとして用いた.

学習データを用い, monophone(音素数 43), 16 混合, 3 状態の SI モデルを作成した. このとき各分布は, MFCC(12 次元)+Δ MFCC(12 次元)+Δ パワー (1 次元) の計 25 次元のパラメータをもつ.

適応・認識には JNAS の 44 話者のデータを用い, 各話者 60 文を適応用, 40 文を評価用に用いた. 適応用の 60 文の中から, 初期適応データをランダムに 5 文選択し, 残りの 55 文を適応用候補文セットとして使用する. 2.3 のの選択手法を用いて, 適応用候補文セットから予め定められた文数の追加適応文を選択し, 初期適応データの 5 文と併せたデータを, SI モデルを初期モデルとしてバッチ適応する. 初期適応データは 3 通り用意し, それぞれ Set1, Set2, Set3 とする. データセットの使い方を示す図を Fig. 2 に示す。適応には MLLR 法を使用し, その際のクラスタ数は 32 とした.

2.1 で述べた音素誤り分布と音素頻度分布について, 出現率が低い音素は適応に効果が少ない

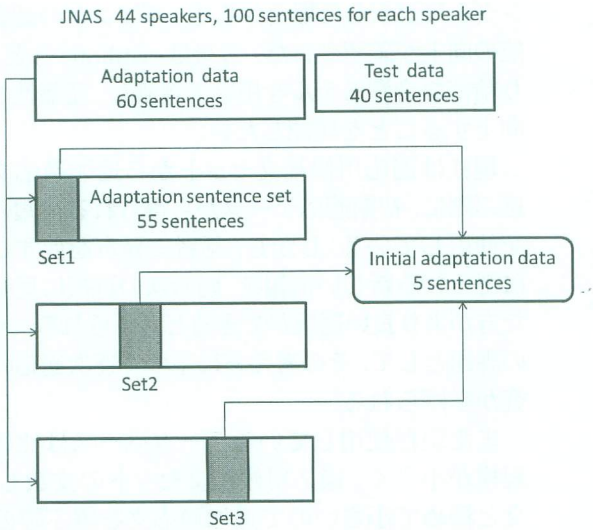

Fig. 2　データセットの使い方

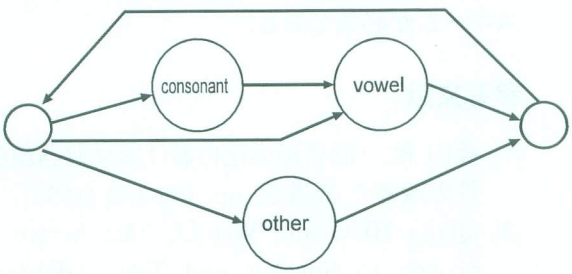

Fig. 3　音素認識の文法

と考え, 考慮しないことにした. ここで考慮する音素は, 出現率の高い $a, i, u, e, o, u:, o:, N, w, y, j, ky, t, k, ts, ch, b, d, g, z, m, n, s, sh, h, r, q$ の計 27 音素である. また, 文ごとの音素頻度分布を求める際に, 出現しない音素が存在する場合のクリッピング値として 1.0×10^{-25} を使用した.

評価には連続音素認識を行った. Fig. 3 に示す単純な音素認識用の文法を用いた. ここで other には促音「っ」(/q/), 撥音「ん」(/N/), 無音区間 (/sp/) が含まれる. 評価基準としては音素正解精度 (phoneme accuracy) を用いた.

4.2　実験結果

4.2.1　追加適応データ数の違いによる比較実験

ここでは Set1 について, 上記のように適応用候補文セット 55 文から必要な文数の追加適応文を選択し, 初期適応データ 5 文とあわせて適応を行った. 提案手法とランダム選択法とで比較実験を行う. ここでランダム選択法とは, 適応用候補文セット 55 文からランダムに追加適応文を選択

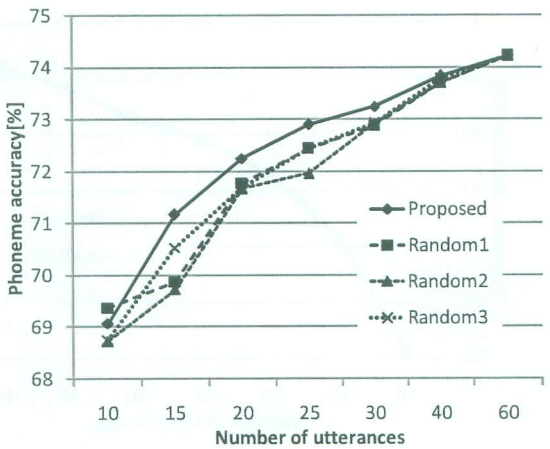

Fig. 4　提案手法とランダム選択法との比較実験

する手法である. ランダム選択法は, 3 通り実験した. 実験結果を Fig. 4 に示す. 横軸が適応に用いた発声数, 縦軸が音素正解精度となっている.

提案手法はランダム選択法と比べて, 平均して全ての発声数において効果があった. しかし, 初期適応データと追加適応データの発声数が等しい 10 文適応の場合は, ランダム選択法の方が提案手法よりも認識制度が高くなるという結果も得られた. 追加適応データ数が少ないため, 追加適応データの違いによる効果が小さくなったことが原因と考えられる. 追加適応データ数が多くなるにつれて, 提案手法とランダム選択法で選択する語彙に違いがなくなり, その差は減少していく.

提案手法による認識性能の改善が最も大きかったのは, 適応数が 15 文の時であり, ランダム選択法で得られた結果の平均から 1.1 ポイントの改善が見られた.

4.2.2　初期適応データの違いによる比較実験

初期適応データを変更すると音素誤り分布が変わるので適応用候補文セットから選択する追加適応文も異なるものになる. そこで, 初期適応データの違いが適応の性能に与える影響を調べるために 3 通りの初期適応データについて実験を行った. ここで使用する初期適応データは Set1, Set2, Set3 である. 結果を Fig. 5 に示す. 初期適応データを変更しても結果はほとんど変わらなかった. この実験条件下では, 初期適応データによる影響はほとんどないと考えられる.

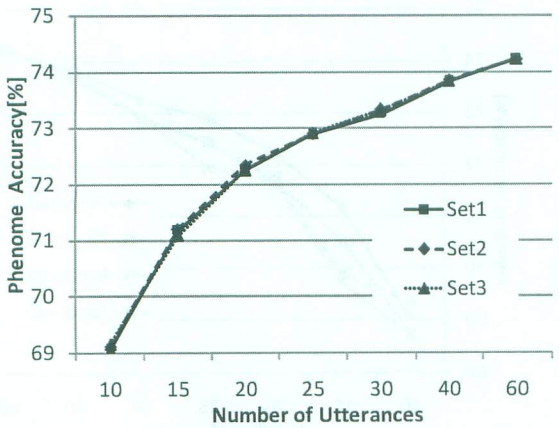

Fig. 5　初期適応データを変更した比較実験

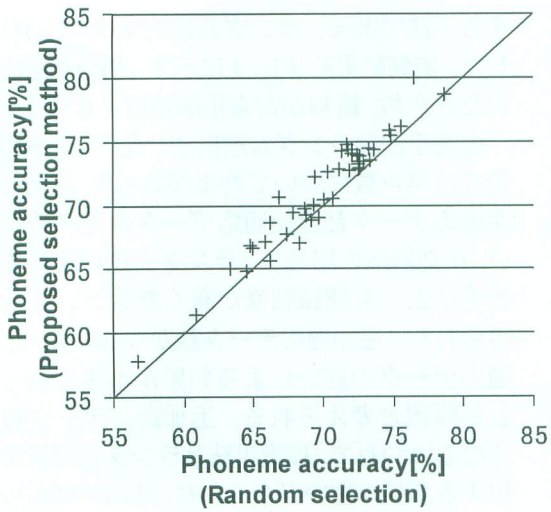

Fig. 6　15 文適応の提案手法の効果 (＋が各話者に対応している)

4.2.3　話者ごとの提案手法の効果

3.1 の実験で効果が大きかった適応数が 15 文の条件での話者ごとの認識結果を示す散布図を Fig. 6 に示す. 横軸がランダム選択法の認識精度の平均, 縦軸が提案手法の認識精度となっている. ほとんどの話者において提案手法の方が高い効果が得られている.

5　おわりに

本稿では, 個々の話者に対して適応の効果が期待できる文を適応文の候補から能動的に選択し, 話者にその文の発声を促し適応することにより, より少量の文での認識精度改善を目指す話者適応化手法を提案した. 比較実験から, 候補からランダムに選択した場合から 1.1 ポイントの認識性能の向上が確認された. 今後は triphone など, より精密な音響モデルを用いて評価し, 認識性能が向上することを確認したい.

現在は適応用候補文セットから追加適応文を選ぶ際に, 初期適応データから得られる情報のみを使用している. しかし, 話者に発声を促すごとに得られる新しい情報を, 随時次の選択に反映した方がより良い選択ができると考えられる. 今後の課題として, その考え方に基づく逐次適応の研究が挙げられる.

また現在使用しているデータベースは比較的規模が小さく, 適応用候補文セットの文数も 55 文と極めて小さいので追加適応文を選ぶ際の選択の幅が限られている. 従って, さらなる性能の改善のためには, 大量の文候補を予め用意し, オンラインで被験者の適応用発声を収集する評価スキームが必要である.

参考文献

[1] 篠田 他, "話者適応化における学習語彙依存性の改善," 音講誌, pp. 132-133 (1992).

[2] Qiang HUO and Wei LI, "An Active Approach to Speaker and Task Adaptation based on Automatic Analusis of Vocabulary Confusability" (2007).

[3] C.J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden markov models," Computer Speech and Language, vol.9, pp.171-185, (1995).

[4] JNAS(新聞記事読み上げ音声コーパス), http://www.milab.is.tsukuba.ac.jp/jnas/.

[5] S-JNAS(高齢者の音声認識用大規模データベース), http://db.ciair.coe.nagoya-u.ac.jp/dbciair/koureisha_files/.