情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-SLP-85 No.3
2011/2/4

音響モデル学習のための相対エントロピーを用いた学習文選択

村上博子[†1]　篠田浩一[†1]　古井貞熙[†1]

大語彙連続音声認識器の音響モデル学習には大規模な音声データが必要となるが，その構築にかかるコストは大きい．本稿では，学習文を選択することにより，従来より少ない学習文数で同等程度の認識性能をもつ音響モデルを学習する手法を提案する．まず，少量の発話データを用いて学習した音響モデルで音素認識を行い，認識単位の誤認識個数の分布を得る．そして，その分布と文内に出現する認識単位の累積頻度分布が近い文集合を文候補から選択する．分布間距離として相対エントロピーを用いる．そして，選択済みの学習文を用いて再度音響モデルを学習し，認識単位を切り替えて再度選択を行う．相対エントロピーの計算において近似を用いることで，計算時間を削減する．提案手法を，教師付き学習と半教師付き学習の両方の条件で，日本語話し言葉コーパスの 152 時間の音声データを用いて評価した．教師付き学習では，ランダムな学習文選択より顕著に良い結果を得た．提案手法は，全学習データを用いたときの単語正解精度 74.7%に，その半分の学習データで到達した．半教師付き学習では高い効果を得られなかった．

A relative entropy based data selection approach for acoustic model training

Hiroko Murakami ,[†1] Koichi Shinoda [†1]
and Sadaoki Furui[†1]

We propose a training data selection method for large vocabulary continuous speech recognition. First, we prepare a large text corpus as a sentence set for training, and obtain phone occurence distribution for each sentence. Second, we calculate phone error distribution from phone recognition result using an initial acoustic model. Then we select sentences whose accumulated phone occurrence distribution is close to the phone error distribution. Our method was evaluated by using 152-hour speech data in the Corpus of Spontaneous Japanese. It was evaluated in situations of supervised training and semi-supervised training. In supervised training, it proved to be significantly better than random selection. It required only 76h of speech data to achieve word accuracy of 74.7%, while standard training (i.e., random selection) required 152h of data to achieve the same rate. It was not significantly effective in semi-supervised training.

1. はじめに

近年の音声認識システムにおいて，高い認識性能をもつ統計的音声認識器を構築するためには，大量の音声データとその書き起こしテキストが必要となる．このような大規模音声データベースの構築には，大きく分けて次の 2 つの形態が考えられる．1 つは，録音した音声データを人手で書き起こすものである．書き起こしなしの音声データの収集は比較的容易であるが，その書き起こしには多くのコストがかかる．もう 1 つは，予めテキストを用意し，それを多数の人が読み上げるものである．書き起こしを必要としないが，音声データの収集にコストがかかる．また，自然発話の収集は難しい．

本稿では，まず最初に後者について検討する．音響モデルの認識性能の向上に対しより効果の高い文をデザインすることで，発声に用いる文数を削減できると期待できる．従来手法としては，予め多様な音素が含まれる音素バランス文セットを作成する手法が主流である[1)2)3)]．音素バランス文には学習に必要な音素が満遍なく含まれているため，様々な音素をバランスよく学習できる．しかし，音響モデルによる認識率が低いことが予想される音素を，認識率が充分高いことが予想される音素よりも多く出現させた方が，より認識性能の高い音響モデルを構築できる可能性がある．また，音素バランス文は，diphone，triphone などの環境依存の認識単位のバランスは考慮されていない場合が多く，これらの認識単位を用いた認識において効果があるとは限らない．そこで，認識率が低い認識単位が多く含まれる文を学習文候補からより多く選択することで，従来より少ない学習文数で同等の認識性能を持つ音響モデルを学習する手法を提案する．

音声認識のための文選択の研究は数多くあり[4)5)6)]，その多くが，選択の基準となる分布を定義し，その分布との分布間距離が小さい文を候補から選択する手法をとっている．分布間距離として相対エントロピーを用いるのが一般的である．提案手法も同様に，相対エントロピーを用いて，認識単位の誤認識個数の分布と，選択した文の認識単位の出現頻度の分布の類似度が最も高くなる文を選択する．相対エントロピーの計算において，近似を用いるこ

†1 東京工業大学大学院 情報理工学研究科 計算工学専攻
Department of Computer Science, Tokyo Institute of Technology

とで，計算時間を削減した．

前述したように，提案手法はそのままでは自然発話の音声データの収集には用いることができない．この問題を解決するために，書き起こしありの音声データと書き起こしなしの音声データを用いて学習を行う，半教師付き学習における学習文選択の検討も行う．書き起こしありのデータを用いて学習した音響モデルを用いて，書き起こしなしのデータを認識する．そして，その認識結果をもとに書き起こす文を選択することで，書き起こしコストの削減を目指す．

2. アルゴリズム

2.1 概　　要

提案手法では，データの増加に伴い，認識単位を monophone，diphone，triphone と切り替えて学習文選択を行う．認識は triphone の音響モデルを用いて行うため，triphone のみを用いる方が良い可能性もある．しかし，triphone は数が多いため，最初に用意される初期発声データだけでは，正確に音響モデルの認識率を予測することは難しい．そこで，上述のように，認識単位を変更する方式をとる．

提案手法では，最初に少量の発声データを用いて学習した音響モデルを用いて音素認識を行い，その認識結果から認識単位の誤認識個数の分布である誤り分布を求める．そして，学習文候補から，文内の認識単位の出現数の分布である頻度分布が誤り分布と最も近くなる文セットを選択する．誤り分布，及び頻度分布の作成を，最初は monophone 単位で行うが，diphone，triphone と認識単位を切り替えて文選択を続ける．

2.2 提案手法の流れ

提案する学習アルゴリズムの概略を図 1 に示す．まず，準備段階として，選択に用いる学習候補文各々において，monophone，diphone，triphone の 3 つの認識単位について各認識単位の出現数の分布である頻度分布を求めておく．また，初期発声データを用意し，その半分を選択済み学習文セット，残りを認識文セットとして用いる．1 つの認識単位に対する文選択は以下の 6 ステップで実行される．

(1) 選択済み学習文セットを用いて monophone の音響モデルを構築し，認識文セットを音素認識する．
(2) 1. の認識結果から，誤り分布を求める．
(3) 相対エントロピーを求めるため，学習文候補それぞれについて，その文を加えた選択済みの文集合の頻度分布と 2. で求めた誤り分布との間の分布間距離を計算する．分布間距離として，カルバック・ライブラー距離 (Kullback-Leibler divergence, KLD)[7] を用いる．候補の中で最も距離が小さい文を選択し，その値を D_{select} とする．
(4) D_{select} と，選択直前の KLD 値 D を比較する．D_{select} の方が小さければ，$D = D_{\text{select}}$ とし，3. に戻る．そうでなければ選択を打ち切る．
(5) 話者に選択した文の発声を促す．
(6) 選択した文を学習文候補から取り除き，選択済み学習文セットに追加し，1. に戻る

(3) で用いる分布間距離の詳細については 4 章で述べる．

提案手法は，上記の (1) から (6) までのステップを，最初は monophone を用いて行い，残りの候補文に対し，2 回目は diphone，3 回目は triphone を用いて行う．最後に，選択された全てのデータを用いて triphone で音響モデルを作成し，単語単位の認識を行う．

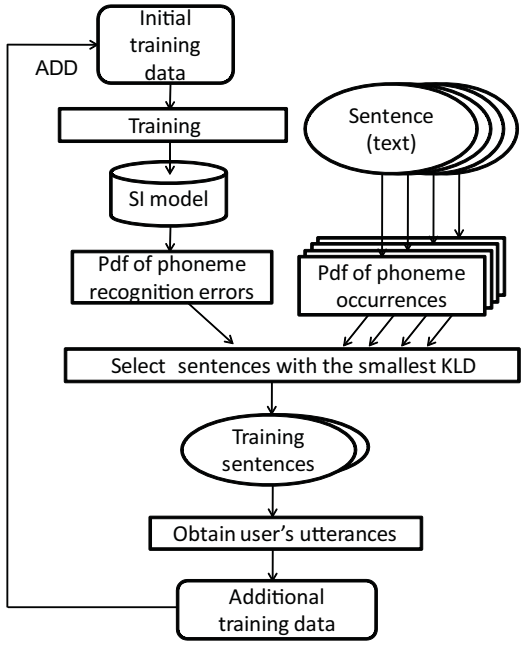

図 1 提案手法の流れ．
Fig. 1 Flow of the proposed method.

Vol.2011-SLP-85 No.3
2011/2/4

3. 誤り分布と頻度分布

本章では，誤り分布と頻度分布の導出について述べる．まず，誤り分布を定義する．2.2 節のステップ 1 で得られる音素認識結果から，各認識単位の誤認識個数を得る．別の単位に誤って認識した場合だけでなく，別の単位をその単位と誤って認識した場合も誤認識個数として数える．分布作成に用いる認識単位は，全ての認識単位から，出現数が多い認識単位を選ぶ．分布を求める際に利用する認識単位の集合を U とし，認識単位 u の誤認識個数を $r(u)$ とすると，全認識単位にわたる誤り分布 $p(u)$ は以下のようになる．

$$p(u) = \frac{r(u)}{\sum_{u \in U} r(u)} \tag{1}$$

次に，頻度分布を定義する．$s(u)$ をある候補文に含まれる認識単位 u の出現回数とすると，その文の頻度分布 $q(u)$ は以下のようになる．

$$q(u) = \frac{s(u)}{\sum_{u \in U} s(u)} \tag{2}$$

4. 分布間距離

提案手法では，初期発声データを含む選択済みの文の集合に追加すると，誤り分布と頻度分布の間の KLD が減少する文を候補から選択する．選択済みの文までの頻度分布を $q'(u)$，選択済みの文の総数を N として，誤り分布と文集合の頻度分布の間の KLD 値 D を以下のように定義する．

$$D = \sum_{u \in U} p(u) \log \frac{p(u)}{q'(u)} \tag{3}$$

そこに，頻度分布 $q(u)$ をもつ文が文集合に加えられたときの KLD 値 D^+ は以下のようになる．

$$D^+ = \sum_{u \in U} p(u) \log \frac{p(u)}{(Nq'(u) + q(u))/(N+1)} \tag{4}$$

diphone，triphone では分布の単位数が多いため，新たな文が選択される度に KLD 値を直接計算すると，計算量が多くなる．ここで，直前の KLD 値との差分のみを計算することで，計算量を削減できる．差分は以下の式で表すことができる．

$$\Delta = D^+ - D = \log\left(1 + \frac{1}{N}\right) - \sum_{u \in U} p(u) \log\left(1 + \frac{q(u)}{Nq'(u)}\right) \tag{5}$$

$\Delta \geq 0$ となるとき，選択を終了する．

各文の頻度分布 $q(u)$ の各認識単位の値のほとんどが 0 となることに着目する．差分のみを計算することで，$q(u)$ の各認識単位の値が 0 となるとき，log の計算を省略することができる．さらに，式 (5) において，テーラー展開式を用いて近似を行うことで，さらに計算量を削減することができる．テーラー展開の第 1 項，及び第 2 項まで用いて近似した差分の式 Δ^1，Δ^2 は以下のようになる．

$$\Delta^1 \approx \frac{1}{N} \sum_{u \in U} p(u)\left(1 - \frac{q(u)}{q'(u)}\right) \tag{6}$$

$$\Delta^2 \approx \Delta^1 - \frac{1}{2N^2} \sum_{u \in U} p(u)\left(1 - \left(\frac{q(u)}{q'(u)}\right)^2\right) \tag{7}$$

$q(u)$ に対して $Nq'(u)$ が大きいほど近似の精度が上がる．

5. 半教師付き学習

本章では，少量の書き起こしありの音声データと，大量の書き起こしなしの音声データを用いて，書き起こす文の選択を行う，半教師付き学習による文選択手法のアルゴリズムを述べる．基本的には，2.2 節で述べた教師付きの手法と同じアルゴリズムで選択を行う．異なる点は，学習文候補として書き起こしなしの音声データを用いるため，候補文を認識して，仮の書き起こしを得る必要があるという点である．仮の書き起こしの精度はできるだけ高いことが望ましい．そのため，認識精度の低くなる音素認識結果ではなく，triphone の音響モデルを用いた連続音声認識 (単語単位) の結果の音素列を仮の書き起こしとして用いる．

まず，書き起こしありの音声データ全てを用いて，triphone で音響モデルを学習する．そのモデルを用いて，書き起こしなしの音声データを連続音声認識 (単語単位) し，その認識結果を仮の書き起こしとして用いる．後は，教師付きの文選択手法と同じアルゴリズムで選択を行い，選択した音声データを実際に書き起こし，選択済み学習文セットとして用いる．

6. 実　　験

6.1 実験条件

データベースとして，日本語話し言葉コーパス (CSJ)[8] における男性話者による学会講演音声を用いた．全データのうち，198,807 発話 (666 話者，152 時間) を学習データとし，2,328 発話 (10 話者，1.95 時間) をテストセットとした．

音声認識に使う特徴量は MFCC12 次元とパワー，及び，その一次微分と二次微分の計 39 次元を用いた．分析周期は 10ms，分析窓幅は 25ms とし，発話単位ごとに CMS を行った．音響モデルは 16 混合 3000 状態 triphone HMM を用いた．認識は 2 パスサーチを行い，言語モデルは 1 パス目に 2 gram，2 パス目に 4 gram を用いた．実験には HTK[9] を用いた．

全学習データからランダムに 13,028 発話 (10 時間) を選択し，半分を選択済み学習文セット，残りを認識文セットとして用いる．残りの学習データ 185,779 発話 (142 時間) は学習文選択の候補文として用いる．比較実験として，候補文から学習文をランダムに選択するランダム選択と比較した．初期音響モデル，及び各回の選択終了後に分布の更新のために構築する音響モデルは monophone とした．diphone は，前の音素からの遷移を考慮した左 diphone を用いた．言語モデルは全学習データを用いて学習したものを用いた．誤り分布，頻度分布の作成に用いる認識単位は，全ての書き起こしテキストの中に 10000 個以上出現するものを用いた．テストデータにおける，利用した各認識単位の占有率は，monophone で 99.9%，diphone で 90.8%，triphone で 71.3%である．

4 章で述べた近似は，認識単位が diphone，及び triphone のときに行った．近似を行わず log の計算をした場合 (Log) と，テーラー展開の 1 項目 (近似 1), 及び 2 項目 (近似 2) まで近似した場合を比較した．

6.2 実験結果

以下，6.2.1〜6.2.3 では教師付き選択手法の実験結果について述べ，6.2.4 では半教師付き選択手法の実験結果について述べる．

6.2.1 ランダム選択との結果比較

図 2 に提案手法とランダム選択の認識結果を示す．ランダム選択では，提案手法の各認識単位において選択された文と同じ時間分の学習文を候補からランダムに選択する．ランダム選択は 3 回行った．提案手法はランダム選択と比べて良い結果となった．候補文全ての 152 時間を学習に用いて到達できる単語正解精度 74.7%を，提案手法は 50%の 76 時間の学習で達成できた．また，monophone による選択終了時 (10h)，diphone による選択終了時 (37h) においてもランダム選択より良い結果を得ることができた．

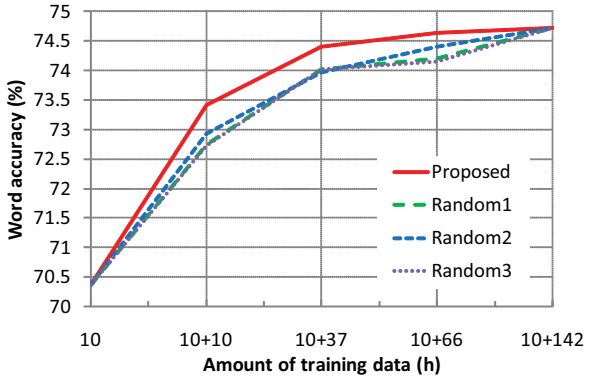

図 2 提案手法とランダム選択の比較実験．Random1，Random2，Random3 は異なる 3 回のランダム選択の結果である．

Fig. 2 Comparison of the proposed method with random selection. Random1, Random2, Random3 are results obtained by three different random selections of training sentences.

6.2.2 KLD 値の変化

図 3 に，選択文数の増加に伴う，誤り分布と文集合の頻度分布の間の KLD 値を示す．認識単位をより詳細な (種類数の多い) 認識単位に変更すると，KLD 値の減少の割合は小さくなり，選択される文数が多くなる．これは，選択回数を重ねると，選択済みの文数が多くなり，式 (4) における $Nq'(u)$ が大きくなるためである．KLD 値の最小値が認識単位を変更するごとに大きくなっていくのは，認識単位数が多くなると頻度分布を誤り分布に正確に近づけるのが難しくなるためである．

6.2.3 近似による結果比較

表 1 に近似を行ったときの提案手法の実験結果を示す．認識精度にばらつきはあるが，いずれもランダム選択よりも高い値となっている．計算時間は，近似 1 は diphone で 56.0%，triphone で 39.3%，近似 2 は diphone で 48.9%，triphone で 27.9%の削減になっている．選択された文数は，近似 2 では diphone，triphone ともに，近似を行わない場合と同じ数の文が選択されている．近似 1 は少し多い数の文が選択される．また，近似では，選択される順番に違いはあるが，近似を行わない場合とほとんど同じ文が選択されている．各認識単位における選択終了後の，誤り分布と文集合の頻度分布の間の KLD 値は，近似の有無，近似方法の違いによる大きな違いはなかった．

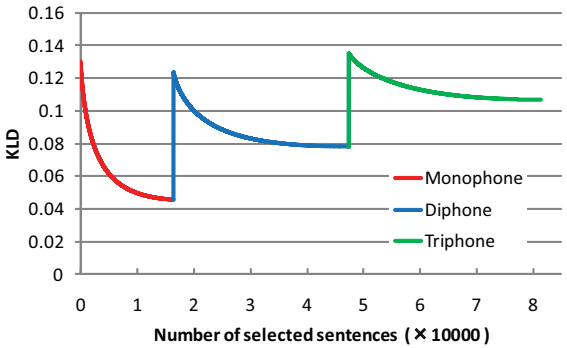

図 3 選択文数の増加に伴う KLD 値の変化.
Fig. 3 Change of KLD values according to the number of selected sentences.

表 1 近似による比較実験. Log は近似を行わない場合の結果であり，Approx1(近似 1)，Approx2(近似 2) はテーラー展開の 1 項目，及び 2 項目まで用いた近似の結果である．認識精度，選択終了までにかかる時間を示す.
Table 1 Comparison of the proposed methods and random selection. Log indicates results without approximation. Approx1 and Approx2 indicate results using approximation by Taylor expansion. The table shows recognition accuracy, and time needed for sentence selection.

	Diphone			Triphone		
	Log	Approx1	Approx2	Log	Approx1	Approx2
認識精度 (%)	74.3	74.2	74.4	74.6	74.7	74.5
計算時間 (h)	4.0	1.8	2.0	5.3	3.2	3.8

6.2.4 半教師付き選択の実験結果

図 4 に半教師付きの文選択手法の認識結果を示す．提案手法は，triphone における選択終了時に，ランダム選択の平均と比べ 0.1 ポイント良い結果となった．残念ながら，教師付きの文選択手法と比べると，ランダム選択との違いがほとんどなくなった．これは，誤りが含まれる認識結果文を文選択の際のラベルとして用いたため，誤り分布と最も近い文を選択できなかったことが原因と考えられる．

7. おわりに

音響モデル学習のための，相対エントロピーを用いた学習文選択手法を提案した．まず，少量の発話データを用いて学習した音響モデルを音素認識し，monophone の誤り分布を得る．そして，誤り分布と文内に出現する音素の頻度分布が近い文を文候補から選択する．分布間距離として相対エントロピーを用いた．そして，選択済みの学習文を用いて再度音響モデルを学習し，分布単位を diphone，triphone と切り替えて文選択を繰り返す．提案手法を日本語話し言葉コーパスの 152 時間の音声データを使い評価し，ランダムな学習文選択より良い結果を得た．候補文全てを学習に用いて達成される 74.7%の単語正解精度に，提案手法は半分の時間の学習で到達することができた．また，相対エントロピーの計算において，近似を用いることにより，計算時間を diphone で 56.0%，triphone で 39.3%削減した．半教師付きの文選択の実験も行ったが，教師付きの文選択ほど良い結果は得られなかった．

今後の課題として，まず，今回の実験では高い効果を得られなかった半教師付きの文選択を改良する必要がある．今回は，学習候補文の認識を選択前にしか行なわなかった．しかし，各認識単位で文セットを選択する度に，音響モデルを構築しなおし，それを用いて次の候補文の認識を行うことで，最初より精度の高い仮の書き起こしを得られることが期待できる．また，学習データの選択単位を文単位から変更することによるさらなる計算時間の削減が挙げられる．選択単位が小さいほど，目的とする分布に近いデータを選びやすいが，オーバーフィッティングする可能性がある．本手法では，1 文ごとに選択を行ったが，選択単位を文の集合ごと，一定数の単語ごと，章ごとなどに変更することで，結果にどのような違いが出るか確認したい．選択単位が大きければ，少ない選択回数で充分な量のデータを得ることができるため，計算時間の削減も期待できる．さらに，音響モデルの誤り傾向の分

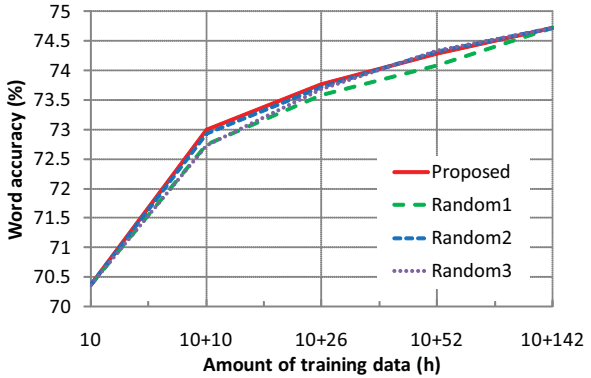

図 4 半教師付き選択手法とランダム選択の比較実験
Fig. 4 Comparison of the semi-supervised selection method with random selection.

Vol.2011-SLP-85 No.3
2011/2/4

析を今回と違う基準で行うことにより，より優れた誤り分布を考案したい．最終的な評価はtriphoneの音響モデルによる単語単位の認識実験において行うため，単語単位の認識結果から音響モデルの認識性能の誤り傾向を分析する方法も考えられる．

参 考 文 献

1) A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis”, Speech Communication, vol. 9, pp. 357-363, 1990.
2) K. Maekawa, and K. Hanae, “Design of a spontaneous speech corpus for Japanese,”, Proc. the International Symposium: Toward the Realization of Spontaneous Speech Engineering, pp. 70-77, 2000.
3) 磯健一, 渡辺隆夫, 桑原尚夫, “音声データベース用文セットの設計,” 日本音響学会講演論文集, 2-2-19, pp.89-90, 1988.
4) Q. Huo and W. Li, “An active approach to speaker and task adaptation based on automatic analysis of vocabulary confusability,” Proc. Interspeech2007, pp. 1569-1572, 2007.
5) A. Sethy, P. G. Georgiou, B. Ramabhadran, and S. Narayanan, “An iterative relative entropy minimization-based data selection approach for n-gram model adaptation,” IEEE Trans. Audio, Speech and Language Processing, vol. 17, no. 1, pp. 13-23, 2009.
6) K. Shinoda, H. Murakami, S. Furui, “Speaker adaptation based on two-step active learning,” Proc. Interspeech2009, pp. 576-579, 2009.
7) S. Kullback, and R. A. Leibler, J. B. MacQueen, “On information and sufficiency,” Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79-86, 1951.
8) K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of Japanese,” Proc. LREC, vol. 2, pp.947-952, 2000.
9) Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/