Vol.2010-SLP-82 No.6
2010/7/23

会議音声認識におけるBICに基づく高速な話者正規化と話者適応

三村正人[†1]　河原達也[†1]

本報告では、高精度かつ高速な会議音声の認識を指向して、音声の区分化と音響特徴量の声道長正規化および音響モデルの話者適応を、BIC(ベイズ情報量基準) に基づいて統一的に行う手法について提案する。提案手法では、音響モデルの学習コーパスに含まれる各発話区間と自動区分化した入力発話区間を ΔBIC により比較し、事前に推定済みのワープ係数や MLLR 変換行列を用いて、認識時に最尤推定を行うことなしに高速に声道長正規化 (VTLN) および話者適応を実現する。国会審議音声を用いた評価実験により、VTLN に関しては、従来の初回認識を行う教師なし最尤推定と同等の認識精度となることを確認した。MLLR 話者適応に関しても、適応を行わない場合よりも有意に精度が向上することを確認した。

Fast Speaker Normalization and Adaptation based on BIC for Meeting Speech Recognition

Masato Mimura[†1] and Tatsuya Kawahara[†1]

This report proposes a unified method for speech segmentation, speaker normalization of spectral features and speaker adaptation of acoustic models based on BIC for efficient and accurate meeting speech recognition. In our method, an input speech segment is compared against each speech segment in the training corpus of the acoustic model based on ΔBIC, and fast VTLN and MLLR adaptation are performed using pre-estimated warping factors and MLLR linear transformations of the corresponding speakers, respectively. Experimental evaluations in Congressional speech transcription demonstrated that the proposed method achieves comparable ASR accuracy to the baseline unsupervised ML estimation for both VTLN and MLLR adaptation.

1. はじめに

会議録等の作成支援を目的として、多数の話者からなる会議音声を書き起こすための音声認識システムが望まれている。不特定話者を対象とした音声認識では、音響特徴量の声道長正規化 (VTLN: Vocal Tract Length Normarization)[1)2)] や音響モデルの話者適応が効果的であるが、その適用の際に音声データを単一の話者のみからなる区間へ区分化する必要がある。さらに通常、話者正規化や話者適応のためのパラメータは、最初に一度音声認識 (初回認識) を行い、その認識結果に対する最尤推定により求める必要があり、計算コストが高い。特に会議のように多数の話者が入れ替わり発話する状況においては、発話交代 (ターン) 毎にこの推定を行うのはリアルタイムを指向した音声認識では現実的でない。

本報告では、高精度かつ高速な会議音声の認識を指向して、音声の区分化と音響特徴量の話者正規化および音響モデルの話者適応を BIC(ベイズ情報量基準) に基づいて統一的に行う手法について提案する。提案手法では、音響モデルの学習コーパスに含まれる各発話区間 (ターン) に対して、事前に VTLN のためのパラメータ (ワープ係数 α) や話者適応のための MLLR 線形変換行列を (人手の) 書き起こしを用いて推定し、BIC の計算に用いる各種統計量とともにデータベース化しておく。認識時には、(区分化の際に計算した統計量を用いて) 入力発話区間とデータベース中の各発話区間の ΔBIC を計算し、値が最小となる発話とその話者を特定し、当該発話区間のワープ係数や当該話者の変換行列をそのまま用いる。これにより、計算コストの高い初回認識による書き起こしの作成や最尤推定を行うことなしに高速に VTLN および MLLR 適応を実現する。

これらの手法を国会審議音声の自動書き起こしのタスクにおいて評価を行った。VTLN に関しては、提案手法は従来の初回認識を行う教師なし推定と同等の認識精度となることを確認した。また、MLLR 適応に関しても、複数の話者の線形変換行列を選択し、その平均を用いることで、適応を行わない場合よりも有意に精度が向上することを確認した。さらに、これらの結果は、発話の区分化を人手で行う場合と BIC により自動で行う場合のいずれの条件でも確認できた。

†1 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

Vol.2010-SLP-82 No.6
2010/7/23

2. ΔBIC に基づく発話の区分化

BIC はモデル選択のための基準であり、尤度と自由パラメータ数に応じたペナルティ項により定義される。二つの区間 S_1(長さ N_{S_1})、S_2(長さ N_{S_1}) に対して、単一のモデル $M_{S_1+S_2}$ で表現した方がよいか、異なるモデル M_{S_1}、M_{S_2} で表現した方がよいか、$\Delta BIC = BIC(M) - BIC(M_1) - BIC(M_2)$ により評価する。特に, モデルとしてガウス分布 (d 次元の全共分散 Σ) を用いる場合、ΔBIC は次式のようになる。

$$\begin{aligned}\Delta BIC = &\frac{1}{2}(N_{S_1} + N_{S_2}) \log |\Sigma_{S_1+S_2}| \\ &- N_{S_1} \log |\Sigma_{S_1}| - N_{S_2} \log |\Sigma_{S_2}|) \\ &- \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1)) \log(N_{S_1} + N_{S_2}) \end{aligned} \quad (1)$$

連続する区間 S_1 および S_2 に対し、この ΔBIC が 0 を上回れば、これらは異なる分布で表現した方がよいといえるため、これらの区間の境界に話者などの音響的条件が変化する点が存在すると判定できる。なお λ は分割重みであり、本稿では $\lambda = 2.0$ とした。

ΔBIC に基づく音声の区分化は、以下のような手続きで行うことができる[3]。

(1) 前回の分割点 (最初は入力の先頭) を始端とする探索窓を設定
(2) 探索窓の中で境界候補点を動かし、$\Delta BIC > 0$ となる点を探索
(3) 窓の中に $\Delta BIC > 0$ となる点がなければ、窓幅を大きくして始端から探索を再開
(4) $\Delta BIC > 0$ となる点があれば、その中で最大となる点を分割点として抽出し、1. へ

区分化された各音声区間は同一の話者からなると考えられるため、この区間を単位として音響特徴量の正規化や音響モデルの話者適応を行うことができる。

3. ΔBIC に基づく高速な声道長正規化および話者適応

区分化に限らず、離れた発話区間 S_i、S_j について ΔBIC(式 (1)) が 0 を下回れば、区間 S_i と S_j は同一のクラスタ (話者) の発話と推定することができる。同一クラスタの区間では、声道長正規化の際に同一のパラメータを用いることができるため、パラメータの最尤推定の回数を削減することができる[4]。

さらに、区分化された現在の入力発話区間と音響モデルの学習コーパス中の各発話区間の間でも ΔBIC に基づいて同一クラスタに属するかどうかを評価することができる。学習コーパス中の各発話区間に対しては、事前に人手の書き起こしを用いて声道長正規化のパラメータ (ワープ係数 α) や MLLR 話者適応のための線形変換行列を高精度に推定しておくことが可能であるため、ΔBIC を用いて学習コーパス中から入力発話と類似したクラスタの発話区間を見つけることができれば、認識時にこれらのパラメータを初回認識結果に基づきオンライン的に最尤推定することが不要となるため、高速な正規化や適応が実現できる。

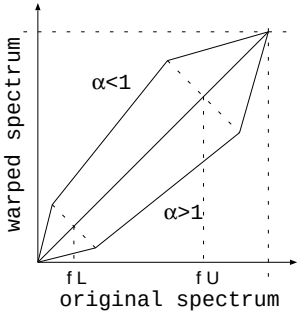

図 1 区分的線形関数による周波数軸伸縮

3.1 声道長正規化 (VTLN)

特徴量の話者正規化として声道長正規化 (VTLN)[1][2] を用いる。VTLN は周波数軸の伸縮処理を行うことで声道長の差に起因するスペクトルの分布を正規化する。伸縮は傾きが α^{-1} の区分的線形関数 (図 1) により行う。

通常、認識時における入力発話区間毎の α の推定は、種々の α の値 ($0.8 \leq \alpha \leq 1.2$) で作成した音響特徴量に対して、別途生成した初期音声認識結果による強制アライメントを行い、最も尤度の高くなる α を探索することで行う。

高速に推定を行う手法としては、GMM を用いるものが一般的である[5]。この手法では、種々の α の値毎に GMM を構築し、入力発話の特徴量に対して尤度が最大となる GMM の α を推定結果とする。また江森らは、α の複数の候補について尤度計算を行うことなく、単一の音響モデルの尤度に基づいて (未知話者のホルマント位置が学習話者からかけ離れた値にならないという仮定のもとで)、解析的・近似的に求める手法[6] を提案している。

3.2 音響モデルの MLLR 話者適応

音響モデルの話者適応として MLLR を用いる[7]。MLLR 適応では、音響モデル自体のパラメータを推定するのではなく、各ガウス分布に対する線形変換行列を最尤基準で推定する。MAP 適応に比べてパラメータ数が少ないため、比較的少量の適応データでも頑健に動

作する利点がある。

しかし、MLLR 適応でも線形変換行列の最尤推定を行うためには適応データに対する音素列の書き起こしが必要となるため、通常初回音声認識により別途書き起こしを生成する。

3.3 ΔBIC に基づく認識対象区間と学習コーパス中の発話区間の高速照合

式 (1) において、S_1 を現在の入力発話区間、S_2 を音響モデルの学習コーパス中の発話区間とすれば、第 2 項 $\log(|\Sigma_{S_1}|)$ は区分化の過程で計算済であり、また第 3 項 $\log(|\Sigma_{S_2}|)$ は事前に計算してデータベースに登録しておくことができる。また、第 1 項には区間 S_1 と S_2 の全サンプルを用いた共分散行列 $\Sigma_{S_1+S_2}$ が現れるが、各区間の十分統計量 (一次統計量、二次統計量、サンプル数) を保持しておけば、個々のサンプルがなくても計算可能である。このようにして、入力発話区間と学習コーパス中の発話区間との ΔBIC の値は高速に計算可能である。

3.4 ワープ係数 α および MLLR 線形変換行列の高速推定

音響モデルの学習コーパスは、通常、人手により単一の話者からなる発話区間 (ターン) に分割済みであり、話者の ID も付与されているものと考えられる。このターンを単位として、以下の項目からなるデータベースを事前に構築しておくことが可能である。

- 発話区間 (ターン)ID
- VTLN のワープ係数 α
- 話者 ID(MLLR 線形変換行列 ID)
- ΔBIC の計算に用いる各種統計量 (一次統計量、二次統計量、サンプル数、$|\Sigma_S|$)

認識時には (自動区分化された) ターンを単位として処理を行うため、データベースもターンを単位として構築するのが望ましい。ただし、事前に何らかの手法でクラスタリングを行えば、照合すべき単位数を削減することも可能である。

ターン毎のワープ係数 α は、通常音響モデルの学習時に人手の書き起こしを用いて推定済みである。

MLLR 適応のための線形変換行列は、同一話者の複数のターンからなる十分長い音声データと人手の書き起こしを用いて、学習済みの音響モデルに対し推定する。

これらの各種統計量は、ターン全体の特徴量自体に比べて十分小さいサイズになる (24 次元の特徴量を用いた場合、300 次元程度)。また、前節で述べたように、これらの統計量を利用すれば、個々のフレームのサンプルがなくても二つの区間の Δ BIC を (簡単な行列演算により) 高速に計算できる。

認識時には、入力発話区間とデータベース中の各ターンとの ΔBIC を計算して値が最小となるターンを選択する。当該ターンのワープ係数 α をそのまま利用して、音響特徴量の声道長正規化を行う。また、当該話者の MLLR 線形変換行列を用いて音響モデルの話者適応を行うが、VTLN に比べて MLLR のためのパラメータ数は多く、1 名の話者では信頼性の高い推定が行えない可能性が高い。HMM の十分統計量を用いた高速教師なし適応の先行研究[8] では、GMM による尤度で上位 N 人の統計量を用いることが効果的であるとされていることから、本研究でも頑健性を考慮して (ΔBIC が小さい) 複数のターンを選択し、それらの話者の線形変換行列を合成して用いる。

4. 音声認識実験

提案手法を衆議院審議音声認識タスクにより評価した。

評価セットには、2010 年 2 月に行われた 3 つの会議 (予算委員会 7 時間、法務委員会 2.5 時間、文部科学委員会 2.5 時間、計 12 時間) を用いた。話者数は 54 名である。また、音声認識実験は、人手による分割と 2 節で述べた ΔBIC による自動分割の両方の条件について行った。分割されたセグメント数は、人手分割では 974、自動分割では 786 となった。自動分割を行う場合は、分割の過程で得られる統計量をデータベースの照合にそのまま利用することができるため効率的である。ΔBIC による自動分割やデータベースの照合に用いる特徴量は MFCC12 次元+ΔMFCC 計 24 次元とした。一方、音声認識に用いる特徴量は MFCC12 次元、ΔMFCC、ΔΔMFCC、Δ パワー、ΔΔ パワーの計 38 次元とした。

音声認識に用いる音響モデルは、2001 年、2004 年、2006 年、2007 年の衆議院審議音声を用いて学習した。データ量は 225 時間である。トライフォン HMM の状態数は 3000、状態あたりの混合分布数は 16 である。特徴量には CMN、CVN、VTLN を適用し、MPE 学習[9] により構築した。また、認識時に VTLN を行わない場合のために、CMN と CVN のみを適用したモデルも用意した。

言語モデルは、1999 年から 2009 年までの会議録に言語モデルの統計的話し言葉変換[10] を適用して作成したものを用いた。語彙サイズは 64k である。

デコーダは、Julius-4.1.2 を用いた。

照合用のデータベースは、音響モデルの学習に用いたコーパスのうち、30 秒以上の長さのターン (9645 ターン、1189 話者) を用いて構築した。VTLN のためのワープ係数 α はターン毎に推定し、MLLR 適応のための線形変換行列は話者毎に推定した。ただし、会議室毎

表 1 音声認識精度 (文字正解精度 %)

推定手法	人手分割	自動分割
(VTLN なし)	83.80	83.58
最尤推定	85.79	85.66
GMM	85.22	85.06
提案手法	85.63	85.54

の音響条件の違いを考慮して、同一人物であっても異なる会議では異なる話者として区別した。

4.1 高速 VTLN の評価

提案手法による高速 VTLN の評価を行った。比較として、3.1 節で述べた従来の教師なし最尤推定による手法、および α の高速推定手法として通常用いられる GMM を用いた手法についても実験した。最尤推定のための初回認識および強制アライメントに基づく尤度の計算には、音声認識で用いる音響モデルと同一のデータで学習したモノフォンモデル (16 混合) を用いた。GMM による手法では、0.8 から 1.2 までの 41 種類の α の値 (0.01 刻み) に対応する GMM を、学習コーパス中の対応する α の値の音声区間を用いて学習した。各 GMM の混合数は 16 とした。

提案手法では、入力発話区間に対する ΔBIC の値が最小となるデータベース中のターンを一つ選択し、そのターンの α を用いて特徴量を正規化した。提案手法は、個々のフレームのサンプルについて多数のガウス分布による尤度を計算しなければならない GMM による手法よりも高速である。

音声認識精度を表 1 に示す。まず人手分割の結果に着目すると、最尤推定による VTLN により CMN、CVN のみを行う場合より認識精度が 2.0% 向上した。GMM による手法ではそれよりも 0.6% 低下した。提案手法では、最尤推定とほぼ同等の結果となった。

自動分割でも、全体的に人手分割とほぼ変わない認識精度になった。また、種々の推定手法の比較についても人手分割とほぼ同一の傾向になった。

4.2 高速 MLLR 適応の評価

提案手法による高速 MLLR 適応についても音声認識実験により評価した。

MLLR 適応では VTLN よりも推定しなければならないパラメータが多いため、データベース中から一つではなく複数のターンを選択し、それらの話者の線形変換行列を合成して入力発話区間のための線形変換行列を生成した。合成の方法にはサンプル数や BIC による重みつき合成等、種々の手法が考えられるが、今回は単純に線形変換行列の各パラメータ毎に相加平均を取ることで生成した。平均は回帰木のクラス毎に計算した。音響モデルの適応は、この線形変換行列をベースライン音響モデルに適用することで行う。選択するデータベース中のターン数は ΔBIC の値の小さいものから順に 20 とした。

表 2 音声認識精度 (文字正解精度 %)

	人手分割	自動分割
提案手法 VTLN	85.63	85.54
+ 提案手法 MLLR	86.06	86.00

音声認識精度を表 2 に示す。提案手法による高速な MLLR 適応により、人手分割、自動分割のいずれの条件でも、提案手法による VTLN を適用した場合よりさらに 0.4% 程度文字認識精度が向上した。これらは、有意水準 1% で有意な結果である。

5. おわりに

BIC を用いて音声の自動区分化と特徴量の声道長正規化および音響モデルの MLLR 適応を統一的に行う手法について提案した。提案手法により、多数の話者からなる会議音声に対して高速かつ高精度な音声認識が実現できることを確認した。

VTLN では、従来の初回認識結果を用いる最尤推定と同等の認識精度を実現した。また、GMM による簡易推定手法よりも高い認識精度を実現した。さらに、MLLR 適応の枠組みを用いて、適応なしの場合よりも有意に高い認識精度を高速に実現した。

提案手法はデータベース化した学習コーパス中のターンと BIC を用いて高速に比較を行い、類似したターンのパラメータをそのまま入力発話に対して適用するという単純な手法であるため、CMLLR などの他の適応手法にも容易に適用可能である。

謝辞 本研究は JST CREST 及び科学研究費補助金によって行われた。

参 考 文 献

1) L.Lee and R.C.Rose. Speaker normalization using efficient frequency warping procedures. In *ICASSP*, pp. 353–356, 1996.
2) S.Wegmann, D.McAllaster, J.Orloff, and B.Peskin. Speaker normalization on conversational telephone speech. In *ICASSP*, pp. 339–342, 1996.
3) S.Chen and P.Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998.

Vol.2010-SLP-82 No.6
2010/7/23

4) 三村正人, 河原達也. 会議音声認識における発話の区分化と話者正規化の高速化. 音講論 (2010 春), pp. 263–264, 2010.
5) L.Welling, S.Kanthak, and H.Ney. Improved methods for vocal tract normalization. In *ICASSP*, pp. 761–764, 1999.
6) 江森正, 篠田浩一. 音声認識のための高速最ゆう推定を用いた声道長正規化. 信学論, Vol. J83-DII, No.11, pp. 2108–2117, 2000.
7) C.J.Leggetter and P.C.Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. In *Computer Speech and Language*, Vol.9, pp. 171–185, 1995.
8) R.Gomez, T.Toda, H.Saruwatari, and K.Shikano. Improving rapid unsupervised speaker adaptation based on hmm sufficient statistics. In *ICASSP*, Vol.1, pp. 1001–1004, 2006.
9) D.Povey and P.C.Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *ICASSP*, pp. 105–108, 2002.
10) Y.Akita and T.Kawahara. Topic-independent speaking-style transformation of language model for spontaneous speech recognition. In *ICASSP*, Vol.4, pp. 33–36, 2007.