論文 / 著書情報
Article / Book Information

論題(和文)	言語モデルの教師なしバッチ型話題適応
Title(English)	
著者(和文)	横山忠介, 篠崎隆宏, 岩野公司, 古井 貞熙
Authors(English)	Takahiro Shinozaki, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2003年春季講演論文集, Vol. , No. 3-4-1, pp. 129-130
Citation(English)	, Vol. , No. 3-4-1, pp. 129-130
発行日 / Pub. date	2003, 3

言語モデルの教師なしバッチ型話題適応 *

◎横山 忠介　篠崎 隆宏　岩野 公司　古井 貞熙　(東工大)

1　はじめに

現在, テキストの読み上げ音声等, 書き言葉を対象とした大語彙連続音声認識は認識精度が高く, 単語正解精度で 90%を上回る. しかしながら話し言葉の音声認識では認識精度が低い. たとえば, 日本語話し言葉コーパス (CSJ) を用いた講演音声の認識においては, 音響モデルを話者適応化した状態で単語正解精度が 70% 程度である. この様に話し言葉音声認識において十分な認識精度が得られない理由として, 話し言葉特有の多様な発話様式, 発話内容等に対応できるだけの十分な学習セットが得られないことに起因した, 認識対象とモデルのミスマッチがあげられる. 話し言葉音声は話者, 話題に応じて音響的, 言語的にも大きく変動するため, 認識率を向上させるためには, モデルの適応化が不可欠である.

以前, 我々は認識をオフラインで行うことを前提とした言語モデルのバッチ型教師なし適応手法を提案した [1]. 文献 [1] では CSJ 講演音声認識をタスクとした評価実験において, 言語モデルのみの適応により, 絶対値で 2.3%単語正解精度が改善したことを報告している. 本稿では, 提案した言語モデル適応手法を音響モデル教師無し適応と併用した場合の認識性能について報告する.

2　バッチ型教師なし言語モデル適応

文献 [1] で提案した手法の概略を図 1 に示す. まず, 適応元の言語モデルとして, 複数の講演から成る学習セットから話題に非依存の単語 n-gram (G-LM) を作成する. また, 適応に先立ち, クラス言語モデルの作成に必要な単語クラスを, 学習セットに対するクラス bigram の尤度が準最大化するように獲得する. なおクラスタリングには incremental greedy merging [2] を用いる.

言語モデル適応手法は次の 3 つのステップからなる.

(1) 話題非依存言語モデル (G-LM) を用いて, 一つの講演音声, 全ての発話を認識する.

(2) (1) で得られた認識仮説文, および単語クラスの定義を用いて, 講演ごとの話題に依存したクラス言語モデル (C-LM) を学習する.

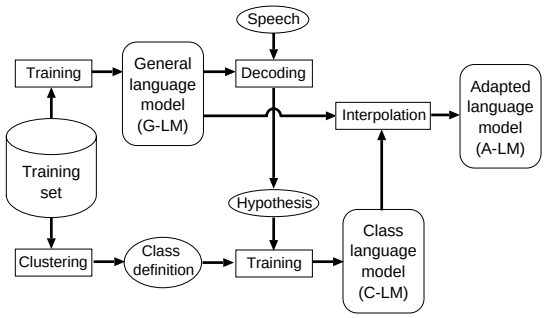

図 1. バッチ型教師なし言語モデル適応法の概要.

(3) 学習した話題非依存言語モデル (G-LM) と話題依存クラス言語モデル (C-LM) を線形補間することにより, 適応モデル (A-LM) を生成する. すなわち, 単語列 h に続く単語 w の G-LM における生成確率を $P_G(w|h)$, C-LM における生成確率を $P_C(w|h)$ としたとき, A-LM における生成確率 $P_A(w|h)$ は線形補間係数 λ を用いて次式で表される.

$$P_A(w|h) = (1-\lambda)P_G(w|h) + \lambda P_C(w|h) \quad (1)$$

3　音響・言語モデル適応

2 節で示した言語モデル適応と, 音響モデル適応を併せて行う. 適応化手順を図 2 に示す. まず, (1) 話題非依存言語モデル (G-LM) と不特定話者音響モデル (G-AM) を用いて認識を行う. (2) 得られた認識仮説を利用して MLLR による教師なし音響モデル適応を行う. (3) 適応音響モデル (A-AM) を用いて再度認識仮説を求め, (4) そこから話題依存クラス言語モデルを作成し, 補間によって適応言語モデル (A-LM) を構築する. (5) 最終的に得られた適応言語・音響モデルを用いて認識を行い評価する.

4　実験条件

4.1　使用コーパス

学習セットとして使用したのは, CSJ の学会講演・模擬講演である. 形態素数にして約 3M, 1289 講演を学習セットとして用いた. なおコーパスは形態素解析ツール JTAG を用いて形態素解析を行った.

評価セットは CSJ の学会講演のうち, 学習セットに含まれていない男性話者 10 名の講演とした. 話者の種別, 形態素数, 話題非依存言語モデル・不特定話者音響モデルを用いた時の単語正解精度を表 1 に示す. 評価セットの形態素数は 48k, 単語正解精度の平均は 65.6% である.

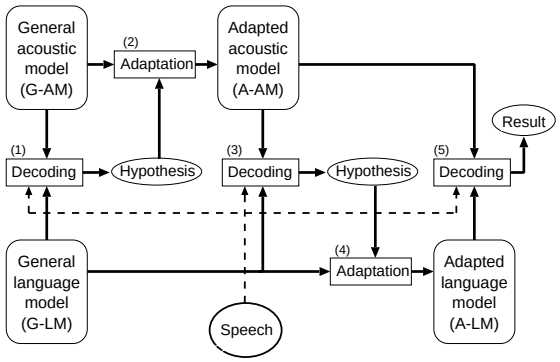

図 2. 音響モデル, 言語モデル適応手順.

∗ Unsupervised batch-type topic adaptation for language models

By Tadasuke Yokoyama, Takahiro Shinozaki, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

表 1. 評価セットの一覧.

ID	Conference name	Number of words	Word accuracy(%)
A01M0007	日本音響学会	4,610	73.11
A01M0035	日本音響学会	6,151	59.11
A01M0074	日本音響学会	2,479	75.71
A02M0076	国語学会	5,045	70.19
A02M0098	国語学会	3,817	64.52
A02M0117	国語学会	9,887	67.20
A03M0100	言語処理学会	2,735	66.49
A03M0111	言語処理学会	3,376	57.24
A05M0031	音声学会	5,288	66.36
A06M0134	社会言語学会	4,585	58.23

4.2 言語モデル

話題非依存言語モデル (G-LM) としては, 順向き単語 bigram と逆向き単語 trigram を使用する. 出現しない n-gram 確率は Katz のバックオフ・スムージング によって推定する. G-LM の作成には, 学習セット中に 2 回以上出現した語彙を使用しており, 語彙サイズは約 35k である.

話題依存クラス言語モデル (C-LM) は, 順向きおよび逆向き bigram を利用する. クラスの連鎖確率, クラスにおける単語の占有確率は認識仮説文から学習する. したがって, C-LM は仮説文に出現した単語のみで構成される.

適応モデル (A-LM) は, 順向き単語 bigram と逆向き単語 trigram である.

なお, 全ての言語モデルの作成には SRILM[3] を用いた. 認識には 2 パスデコーダである Julius を利用し, ファーストパスで順向き bigram, セカンドパスで逆向き trigram を使用する. 認識の際の言語重み, 挿入ペナルティは話題非依存モデルを用いた認識において 10 話者平均の単語正解精度が最も高かった値を全ての認識で共通に用いた.

4.3 音響モデル

16kHz で標本化, 16 ビットで量子化された講演音声から, 音響特徴量として MFCC 12 次元・Δ MFCC 12 次元・Δ 対数パワーの計 25 次元を抽出した. なお, 入力音声ごとに CMS を行っている. 言語モデルの学習に使用した講演に含まれる, 455 講演, 約 94 時間の男性話者による講演音声を用いて, 不特定話者音響モデル (G-AM) を作成する. 総状態数は 3000, 混合数 16 の状態共有 triphone である. 音響モデルの学習には HTK v2.2 を用いた.

5 実験結果

適応モデルを用いて各講演の音声認識実験を行った. 図 3 に言語モデル適応による単語正解精度の変化を示す. グラフ中の 「with A-AM」 は音響モデル適応と併用した結果, 「with G-AM」 は 言語モデル適応のみを行った結果を表す. グラフの横軸には線形補間係数 λ を示し, C-LM のクラス数ごとに評価セット 10 講演の平均の単語正解精度を示した. また w-2gram は C-LM として認識仮説文から学習した単語 bigram を使用した結果である. まず, 線形補間係数 $\lambda = 0$ のときの結果から, 音響モデル適応により 4.3% の単語正解精度の改善が得られていることがわかる. 音響モデルの適応に加え, 言語モデルの適応を行うことにより, さらに最大 2.0%の改善が得られた. 線形補間係数, 単語クラス数に依存する単語正解精度の改善は言語モデルのみを適応した結果と酷似しており, 音響モデル適応の有無に関わらずクラス数 100, 線形重み 0.3 のモデルで最適な認識性能を得た. その結果, 単語正解精度で合計 6.3%の改善が得られた.

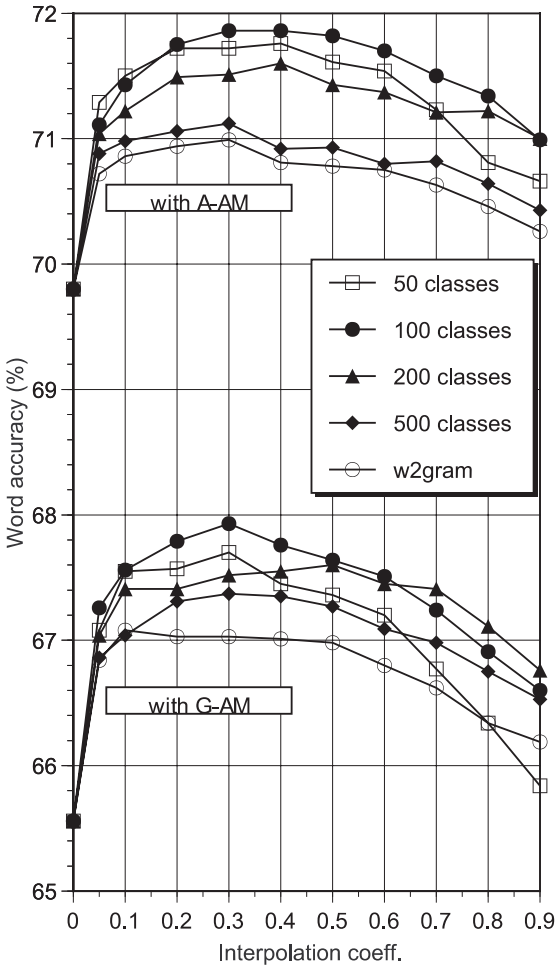

図 3. 言語モデル適応による単語正解精度の変化.

6 まとめ

提案した言語モデル適応手法が, 音響モデル適応と併用した場合においてもほぼ加算的に作用することを確認した. 最も高い単語正解精度が得られたのは言語モデルのみを適応した場合と同様のクラス数 100, 線形重み 0.3 の適応モデルであり, 絶対値 2.0%の改善を得た.

今後は, 講演ごとのクラス数, 線形補間係数の自動決定について検討を行う予定である. また, 話題依存クラスモデルに類似した講演の語彙を加えることによる改善の効果について実験を行う予定である.

参考文献

[1] 横山 忠介, 篠崎 隆宏, 古井 貞熙, “講演音声認識を対象とした言語モデルの話者適応化,” 日本音響学会 2001 年秋季講演論文集, 3-9-6, pp.141–142(2002).

[2] Brown, P.F., Della Pietra, P.V., Lai, J.C., and Mercer, R.L., “Class-Based ngram Models of Natural Language,” Computational Linguistics, vol.18, no.4, pp.467–479(1992).

[3] http://www.speech.sri.com/projects/srilm/