音声言語情報処理 44－31
(2002. 12. 20)

言語モデルのバッチ型教師なし適応化法

横山忠介† 篠崎隆宏† 岩野公司† 古井貞煕†

† 東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152–8552 東京都目黒区大岡山 2–12–1
E-mail: †{tadasuke,staka,iwano,furui}@furui.cs.titech.ac.jp

あらまし 本稿では話し言葉音声認識の性能向上を目的とした，クラスモデルを用いた言語モデルのバッチ型教師なし適応化法を提案する．対象としているタスクは日本語講演音声認識である．提案手法では，複数の講演から構築される話題非依存の単語 n-gram を用いて一つの講演音声を全て認識し，その認識仮説から講演ごとの話題依存クラス言語モデルを学習する．得られた話題依存クラス言語モデルを話題非依存言語モデルと線形補間する事により講演ごとに言語モデル適応を行い，その適応モデルを用いて講演音声を再認識する．提案する手法を用いた評価実験を行った結果，評価セット中の全ての講演について適応による単語正解精度の向上を確認した．適応化における最適なクラス数は 100 程度であり，そのときの単語正解精度の改善は絶対値で 2.3% であった．さらに，音響モデルの教師なし適応を併用した場合についても言語モデルの適応化の効果を評価する実験を行ったところ，同様の認識性能の改善が得られ，最終的な講演音声認識性能は，単語正解精度で約 71.8% に達した．
キーワード 教師なし言語モデル適応，バッチ型適応，クラス言語モデル

Unsupervised batch-type adaptation method for language models

Tadasuke YOKOYAMA†, Takahiro SHINOZAKI†, Koji IWANO†, and Sadaoki FURUI†

† Department of Computer Science, Tokyo Institute of Technology
2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
E-mail: †{tadasuke,staka,iwano,furui}@furui.cs.titech.ac.jp

Abstract This paper proposes an unsupervised, batch-type, class-based language model adaptation method for spontaneous speech recognition. The word classes are automatically determined by maximizing the bigram likelihood using a training set. A class-based language model is built based on recognition hypotheses obtained using a general word-based language model, and linearly interpolated with the general language model. All the input utterances are re-recognized using the adapted language model. The proposed method was applied to the recognition of spontaneous presentations and was found to be effective in improving the recognition accuracy for all the presentations. The best condition was found to be using 100 word classes, and in this condition 2.3% of the absolute value improvement in the word accuracy averaged over all the speakers was achieved, using speaker independent acoustic models. It was also found that effectiveness of the proposed method is additive to that of the acoustic model adaptation. Consequently, 71.8% word recognition accuracy was achieved for spontaneous presentations after adapting both acoustic and language models.
Key words Unsupervised language model adaptation, batch-type adaptation, class-based language model

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音声認識精度は非常に高く，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精度は，未だ非常に低い．たとえば,「日本語話し言葉コーパス（CSJ）[1]」を用いた講演音声の認識では，音響モデルの教師なし話者適応を行った状態で，単語正解精度 70% 程度の認識精度である [2]．このように十分な認識率が得られない主な要因としては，話し言葉の多様な言い回しや発話内容，発話様式に十分対応できるだけの学習デー

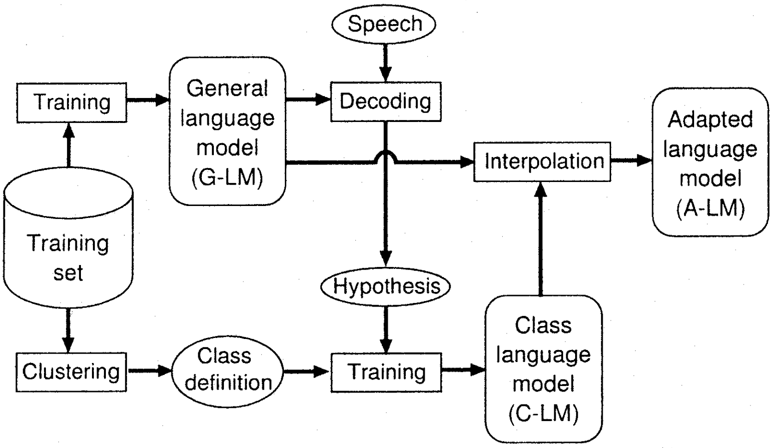

図 1　教師なしクラス言語モデル適応法の概要.
Fig. 1　An overview of the unsupervised class-based language model adaptation method.

タが十分に存在しないため，音響・言語モデルと認識対象との間にミスマッチが生じていることが考えられる．講演音声は，その話者や話題に応じて，音響的・言語的に大きく変動しているため，音声認識精度を向上させるためには，講演ごとに音響・言語モデルの適応化が不可欠である．

適応化手法は，教師なし適応と，教師あり適応に大別される．このうち，教師なし適応は，認識結果をそのまま適応用データとして利用することが可能であり，教師あり適応に比べて応用範囲が広く，有用性が高い．しかし，適応データとなる認識結果中に認識誤りが含まれているため，教師有り適応に比べ性能の向上が得られにくい．特に，話し言葉音声の認識結果には相当数の誤りが含まれているため，教師無し適応は困難な課題となる．

講演音声認識に音響モデルの教師なし適応を行った例としては，講演音声認識をオフラインで行うことを前提とした，バッチ型の教師なし適応に関する報告がある．この手法では不特定話者音響モデルを用いて各講演における全ての発話を認識し，認識結果を用いて音響モデルの適応化を行い，全ての発話を再認識する．適応化を繰り返し用いることによって単語正解精度で約 5% の改善が得られている [2]．一方，これまで言語モデルの教師なし適応化手法も提案されているが，話し言葉の音声認識において，認識率の大きな改善が得られたケースは少ない [3]．これは，適応元の言語モデル空間が認識タスクに対して非常にスパースであり，かつ，認識仮説中には多くの認識誤りが含まれていることから，信頼できる少量の適応データで効率的に言語モデルを適応化することが困難であるためである．このような言語モデルのスパースネスを解決する方法として，クラス言語モデルを用いた様々な適応化手法が提案されている [4], [5]．これらの多くは，教師あり適応手法であり，また，話し言葉音声認識への適用例は少ない．

そこで本稿では，クラスモデルを用いた言語モデルのバッチ型教師なし適応化手法を提案し，CSJ を用いた話し言葉音声認識実験において本手法の有効性を示す．

2. バッチ型教師なし言語モデル適応

図 1 に本手法の概略を示す．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 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，後述するクラスタリング手法によって作成する．

提案する適応手法は次の 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．

(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．

$$P_A(w|h) = (1-\lambda)P_G(w|h) + \lambda P_C(w|h) \tag{1}$$

単語クラスタリングは，一つの単語が複数のクラスに属さないという条件のもとで，学習セットに対するバイグラム尤度を最大化するように行われる．

クラスの集合 $\pi = \{c_i\}$ が与えられたときの，クラスバイグラムモデルによる学習データ w の尤度関数 $L(\pi)$ は以下のように定義される．なお，C は学習データにおける出現回数を表す．

$$L(\pi) = \sum_{c_1,c_2} C(c_1,c_2)\log\frac{C(c_1,c_2)}{C(c_1)C(c_2)} + \sum_{w} C(w)\log C(w) \tag{2}$$

クラスに依存するのは右辺第一項のみなので，尤度関数を最大化するためには右辺第一項のみを最大化するクラス集合 π を求めればよい．その際，クラスタリングの総数は有限なので，総あたりで行えば最適解を得ることが理論的には可能であるが，現実的なリソースで計算を行うため "incremental greedy merging algorithm [6]" を用いて近似解を得る．

3. 実験条件

3.1 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，およびパープレキシティを表 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．

なお音声認識には、エネルギーを用いて切り出した単位をもとに、単語の途中などで切れないように人手により修正した音声単位を用いた。認識単位の切れ目はおよそ 500ms 以上の無音に相当する。

3.2 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 [7] 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 [8] 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．

3.3 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．

4. 実験結果

適応言語モデル (A-LM) を用いて各講演のパープレキシティの計測，音声認識実験を行った．

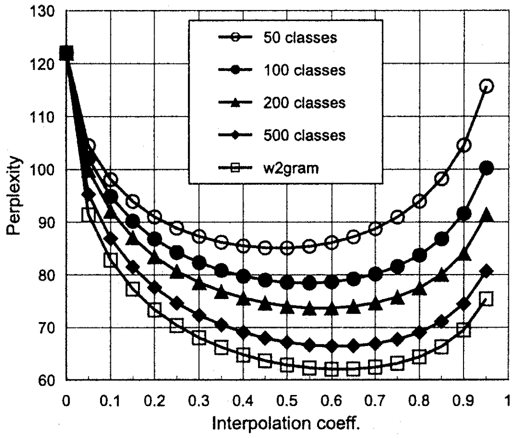

図 2　言語モデル適応によるパープレキシティの変化．

Fig. 2　Test-set word perplexity as a function of the interpolation coefficient λ.

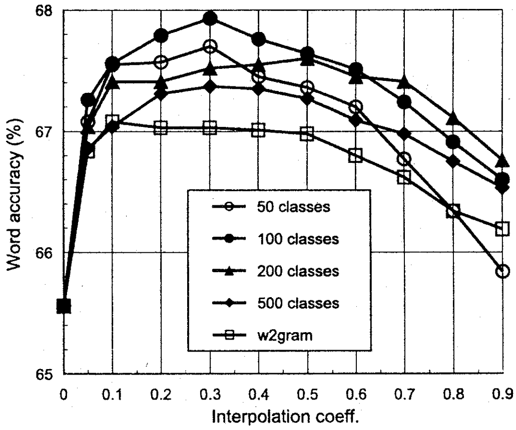

図 3　言語モデル適応による単語正解精度の変化．

Fig. 3　Word accuracy as a function of the interpolation coefficient λ.

図 2 に各講演の正解文に対する単語パープレキシティの平均と適応モデルの線形補間係数の関係を示す．パープレキシティは trigram を用いて計測した．話題依存クラス言語モデル (C-LM) におけるクラス数を 50，100，200，500 とした時の結果と，話題依存モデルをクラス化を行わず単語 bigram で作成した時の結果 (w2gram) をあわせて示す．線形補間係数 $\lambda = 0$ となる点がベースラインの話題非依存モデル (G-LM) を使用した結果であり，言語モデルの適応化によるパープレキシティの大きな減少が示されている．クラス数が十分多い場合においては最適な線形補間係数において約半分まで減少している．また，クラス数の増加にともない，パープレキシティが単調に減少していることもわかる．

評価セットの話者 10 名の単語正解精度の平均と線形補間係数，およびクラス数の関係を図 3 に示す．認識の際に用いる言語重み，挿入ペナルティは話題非依存言語モデルと不特定話者音響モデルを用いた認識実験において，平均で最も単語正解精

表 1　評価セットの一覧.
Table 1　List of the test set data.

ID	Conference name	Number of words	Word accuracy (%)	Perplexity
A01M0007	日本音響学会	4,610	73.11	82.97
A01M0035	日本音響学会	6,151	59.11	127.01
A01M0074	日本音響学会	2,479	75.71	99.48
A02M0076	国語学会	5,045	70.19	138.86
A02M0098	国語学会	3,817	64.52	148.17
A02M0117	国語学会	9,887	67.20	140.70
A03M0100	言語処理学会	2,735	66.49	87.58
A03M0111	言語処理学会	3,376	57.24	138.10
A05M0031	音声学会	5,288	66.36	151.14
A06M0134	社会言語学会	4,585	58.23	106.42

度が高かった値を共通に用いた．なお，最適な言語重み，挿入ペナルティはファーストパス 10，-6，セカンドパス 8，-6 であった．適応化を行ったモデルを用いることで話題非依存モデル ($\lambda = 0$) に比べ，単語正解精度が向上していることがわかる．最も高い認識率が得られたのは線形補間係数が 0.3，クラス数 100 で適応したときであり，話題非依存モデルと比較した単語正解精度の向上は絶対値で約 2.3% であった．

図 2, 3 より，パープレキシティの減少および単語正解精度の向上はクラス数，線形補間係数に大きく依存することがわかる．一般に，ある文に対するパープレキシティが減少すれば，その文に対する単語正解精度は向上する．しかし，本実験では，クラス数が 100 以上となったとき，クラス数の増加に対して，パープレキシティは単調に減少しているが，単語正解精度が低下する傾向にある．これは，クラス数の増加にともない適応モデルが精密化することから，誤りを含んでいる仮説文に対して過度に適応が進んでしまい，正解文に対する適応モデルの尤度向上の効果が得られにくくなることに起因しているものと考えられる．

図 4, 5 に，講演ごとの線形補間係数とパープレキシティおよび単語正解精度の関係を示す．この実験では，平均で最も高い認識率が得られたクラス数 100 の適応モデルを使用している．

図 4 からパープレキシティが全ての講演で改善していることが分かる．改善の大きさ，最適な線形補間係数は講演ごとに異なり，改善が大きい講演ほど最適な補間係数が大きくなる傾向が見られる．図 5 からは，単語正解精度においても全ての講演で改善していることが分かる．線形補間係数によっては，ベースラインを下回る話者もいるが，補間係数を 0.4 以下にとれば 1 講演を除く全ての講演で改善が得られている．最も認識性能が向上した講演で絶対値で約 6%，逆に向上がみられなかった講演でも約 0.5% の単語正解精度の改善が見られた．

適応化による講演ごとの（正解文に対する）パープレキシティの改善と単語正解精度の改善をプロットしたものを図 6 に示す．ここでは，クラス数は 100，線形補間係数は 0.3 とした．パープレキシティの減少が大きい講演ほど，単語正解精度の改善が大きいという傾向が見られる．相関係数は -0.79 で 1% 水準で有意であり，両者に強い相関がみられる．

図 7 には話題非依存言語モデル (G-LM) で計測した各講演の

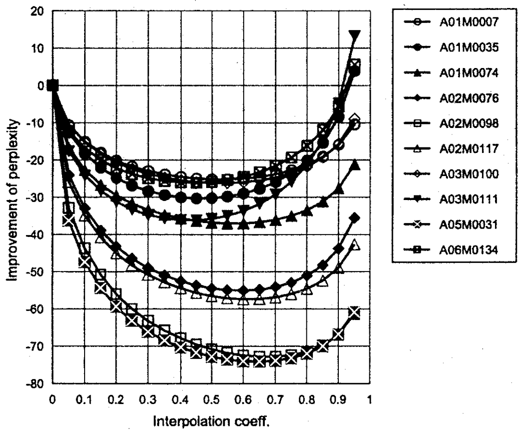

図 4　言語モデル適応による講演ごとのパープレキシティの改善.
Fig. 4　Improvement of the perplexity as a function of the interpolation coefficient λ for each presentation.

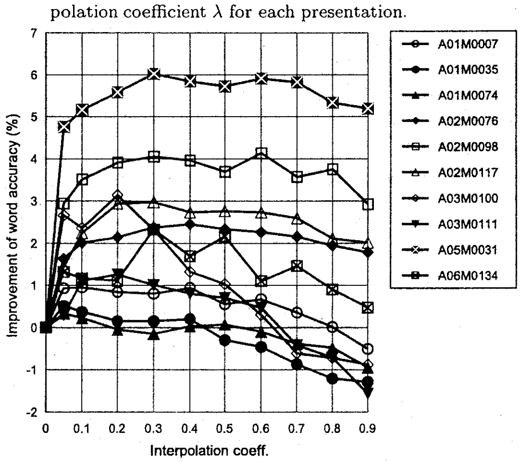

図 5　言語モデル適応による講演ごとの単語正解精度の改善.
Fig. 5　Improvement of the word accuracy as a function of the interpolation coefficient λ for each presentation.

正解文のパープレキシティから認識仮説文のパープレキシティを引いた値と，単語正解精度の改善の関係を示してある．なお考察に際し，より確かな関係を見いだすために，追加で 10 講

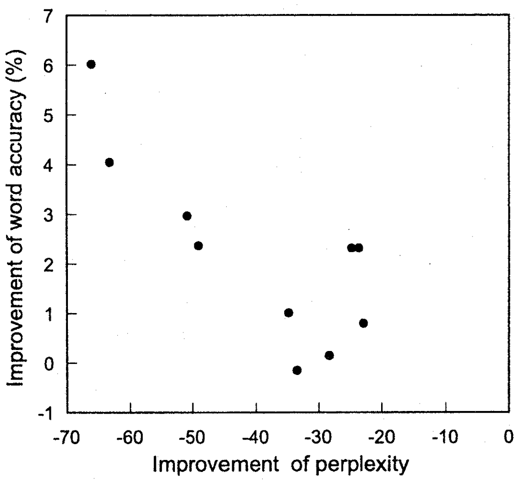

図 6　言語モデル適応による正解文に対するパープレキシティの改善と単語正解精度の改善との関係.

Fig. 6　Relationship between the improvement of perplexity and word accuracy.

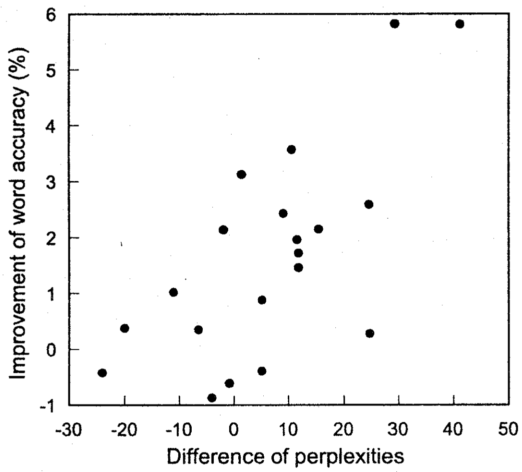

図 7　話題非依存モデル (G-LM) の正解文に対するパープレキシティと認識仮説文に対するパープレキシティの差と単語正解精度の改善との関係.

Fig. 7　Relationship between the difference of the perplexities between recognition hypothesis and correct transcription calculated using G-LM and the improvement of word accuracy.

演の認識実験を行い，計 20 講演のグラフとした．ここでもクラス数 100，線形補間係数 0.3 で適応化を行っている．この結果から，パープレキシティの差分が単語正解精度の向上に対応していることが分かる．グラフの相関係数は 0.70 で，0.1% 水準で優位であるという結果が得られた．パープレキシティの差分は話題非依存モデルにおける講演の不適応度の大きさと見なすことができるため，話題非依存モデルで十分に対応できない講演に関して適応化の効果が大きいということができる．

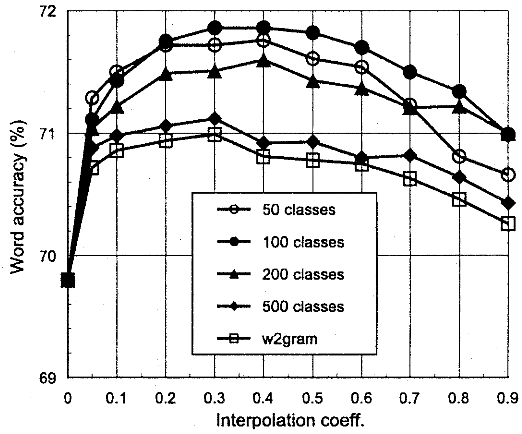

図 8　音響モデル適応を併用したときの言語モデル適応による単語正解精度の変化.

Fig. 8　Word accuracy using the speaker adapted acoustic models as a function of the interpolation coefficient λ.

5.　音響モデル適応の効果

提案する言語モデル適応手法とあわせて，音響モデル適応を行った場合の講演音声認識の性能について，評価実験を行う．

手順としては，まず，話題非依存言語モデルと不特定話者音響モデルを用いて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．得られた適応音響モデルを用いて再度認識仮説を求め，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を構築する．

このように音響・言語モデルとも適応した場合の，単語正解精度の線形補間係数，クラス数との関係を図 8 に示す．

音響モデル適応を用いていないときの結果（図 3）と比較すると，線形補間係数，単語クラス数による単語正解精度の変化の傾向は酷似しており，音響モデル適応を行ったときでも，クラス数 100，線形補間係数 0.3 で最適な認識性能を得た．線形補間係数 $\lambda = 0$ のときの結果を比較すると，音響モデルの適応化により，約 4% の認識性能が得られていることがわかる．また，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，さらに認識性能が最高で約 2% 向上しており，音響モデル適応を行った場合でも，提案する言語モデル適応手法が有効に作用することが確認できる．その結果，音響・言語モデルの適応化によって合計 6% の認識性能の改善が得られ，最終的な講演音声の単語正解精度は約 71.8% に達した．

6.　ま と め

本稿では，話し言葉音声認識を対象として，言語モデルの適応化手法を提案した．提案手法では，まず，学習コーパスに対してバイグラム尤度が準最大化するようにクラスを自動獲得する．そして，話題非依存言語モデルを用いた認識仮説文から講演ごとの話題依存クラス言語モデルを学習し，そのクラスモデルと話題非依存言語モデルを線形補間することで適応化を行う．CSJ の講演話者 10 人を対象とした認識実験において，クラス

数 100 の適応化モデルを用いることにより，単語正解精度の平均が絶対値で 2.3% 向上したことから，本手法の有効性が確認された．また，音響モデルを教師なし適応化することにより，さらに約 4% の性能向上がみられ，最終的な講演音声認識性能は，単語正解精度で約 71.8% に達した．

今後は，適応化モデルの最適なクラス数，線形補間係数を教師なしで推定する予定である．また，本手法では話題非依存な言語モデルを用いたときの認識における仮説文に出現しない単語に対して誤りを改善することはできない．そこで，教師なしで学習したクラスモデルの語彙に，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．

文　献

[1] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous speech corpus of Japanese," *Proc. LREC2000*, Athens, Greece, vol.2, pp.947–952, 2000.

[2] T. Shinozaki, C. Hori and S. Furui, "Towards automatic transcription of spontaneous presentation," *Proc. Eurospeech2001*, Aalborg, Denmark, vol.1, pp.491–494, 2001.

[3] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," *Proc. ICSLP2002*, Denver, vol2, pp.1413–1416, 2002.

[4] G. Moore and S. Young, "Class-based language model adaptation using mixtures of word-class weights," *Proc. ICSLP2000*, Beijing, China, vol.4, pp.512–515, 2000.

[5] 山本 博司, 勾坂 芳典, "話題と文型の違いを同時に考慮した言語モデルの適応", 電子情報通信学会論誌, vol. J85-D-II, No.8, pp.1284-1290, 2002.

[6] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol.18, no.4, pp. 467–479, 1992.

[7] S. M. Katz, "Estimation of probabilities from sparse data for language model component of a speech recognizer," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.35, no.3, pp.400–401, 1987.

[8] A. Stolcke, "SRILM - an extensible language modeling toolkit," *Proc. ICSLP2002*, Denver, vol2, pp.901–904, 2002. http://www.speech.sri.com/projects/srilm/