社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ＨＩ－122 (19)
2007－SLP－ 65 (19)
2007／2／10

公共案内システムにおける幼児音声認識部の開発と評価

進藤 泉† ツィンツァレク・トビアス† 戸田 智基† 猿渡 洋†
鹿野 清宏†

† 奈良先端科学技術大学院大学 情報科学研究科 〒 630–0101 奈良県生駒市高山町 8916–5
E-mail: †{izumi-s,cincar-t,tomoki,sawatari,shikano}@is.naist.jp

あらまし 近年，子供音声認識の需要が日々高まっているが，実際の認識性能は低迷している.「日本語ディクテーション基本ソフトフェア」として規定されている小児音声モデルを用いた幼児音声の認識率は 21.4%であり，実用には至らない．幼児の発話は不安定なものであり，ひとつの単語を発話する際にも様々なバリエーションが現れる．そして，幼児特有の拗音や言い誤りなどが頻繁に現れる事も認識精度低下の原因の一部であると考えられる．本稿では幼児発話の特徴を実験データから調査し，音声の変化を考慮した認識手法を提案する．さらに，幼児発話認識に特化した音響モデル・言語モデルの構築を行い，提案法とあわせて評価を行う．実験データとして，公共施設に常設されている音声情報案内システム「たけまるくん」に対して自由に話しかけられた不特定多数の子供発話 (2〜15 歳) を使用する．実験データを用いたモデル構築と提案法の併用により，幼児音声の認識率は 54.8%にまで向上した.

キーワード 幼児，子供音声認識，実環境，選択学習，発音パターン

Development and Evaluation of Preschool Children Speech Recognition Module of a Public Guidance System

Izumi SHINDO†, Tobias CINCAREK†, Tomoki TODA†, Hiroshi SARUWATARI†, and Kiyohiro SHIKANO†

† Graduate School of Information Science, Nara Institute of Science and Technology
Takayama-cho 8916–5, Ikoma-shi, Nara, 630–0101 Japan
E-mail: †{izumi-s,cincar-t,tomoki,sawatari,shikano}@is.naist.jp

Abstract In recent years, there is an increasing demand for speech recognition of children. However, the recognition of children's speech, especially preschool children (2 to 5 years of age), is very difficult. For example, recognition accuracy using a children's acoustic model provided by the Japanese Dictation Toolkit is only 21.4%. Many different variations of children speech with palatal sounds and pronunciation error decrease recognition performance. This paper proposes an approach to recognition preschool children speech. The characteristics of preschool children's speech using experimental data are investigated in order to consider phonetic changes. An acoustic model, a language model for preschool children speech recognition and a pronunciation dictionary are built. For experiments, a large database of spontaneous child speech (2 to 15 years of age) collected with the speech-oriented public guidance system, "Takemaru-kun," is employed. Recognition performance increases to 49.2% by acoustic model adaptation using preschool children's speech. By combination of using proposal method and new models build experimental data, further improvement to 54.8% is achieved.

Key words preschool children, speech recognition of children, real environment, selective training, pronunciation pattern

1. はじめに

近年，子供用携帯端末・ゲーム機を筆頭とする商品が市場に普及し，子供音声認識の需要が日々高まっている [1]. しかし，実際の認識性能は低迷している．子供音声の中でも，幼児音声に関しては特に厳しい認識状況である．「日本語ディクテーション基本ソフトフェア」[2] として規定されている小児音声モデルを用いた幼児音声の認識率は21.4%であり，技術を実用化するには時間を要する．

幼児音声認識が困難な理由として，幼児の発話は不安定なものであり，ひとつの単語を発話する際にも様々なバリエーションが現れる事が上げられる [3]．また，幼児特有の拗音や言い誤りなどが頻繁に現れる事も認識精度低下の原因の一部であると考えられる．本稿では幼児発話の特徴を実環境で収録した実験データから調査し，音声の変化を考慮した認識手法を提案する．また，幼児音声認識に特化した音響モデル・言語モデルを実験データから構築することにより認識率向上を目指す．

本稿の構成として，2. 章で使用する実環境収録データ，3. 章で提案する幼児音声認識システムの構成，4. 章で実験的評価，5. 章で結論について述べる．

2. 使用データ

2.1 音声データの収集

子供音声認識技術が向上しない原因の一つに，音声データを収集する事が困難である事が挙げられる．近年の子供音声認識の研究では，子供音声を収集するための専用システムを用い，小学生データをはじめとした様々なデータベース構築がなされている [4] が，自由発話のデータベースは存在しない．本稿は，公共施設に常設されている音声情報案内システム「たけまるくん」[5] に対して自由に話しかけられた不特定多数の子供発話 (2 歳〜15 歳) を使用する．

2.2 データ内訳

本実験で使用するデータの内訳を図 1 に示す．図の外円は子供 (幼児・小学生・中学生)，成人 (高校生以上の大人・高齢者) のカテゴリで収集データの内訳を示している．内円はそれぞれのカテゴリ内の詳細なデータ内訳を示す．公共案内システムで収集できる子供発話データの割合は，子供発話が約 8 割を占め，子供音声認識の需要の高さを示している．子供発話のデータは幼児 (2〜5 歳)・小学生 (6〜11 歳)・中学生 (12〜15 歳) と分類されている．本稿では主に幼児音声の認識率向上を目的とする．本稿で扱う幼児音声部のデータ量は全データのうち約 12%を占めている．

3. 提案する幼児音声認識システムの構築

3.1 正解判定基準の変更

「たけまるくん」で収集した音声データの書き起こし文は，言い誤りや幼児特有の発話を含み，実際に聞こえる音に忠実に表記されている．現在幼児音声の認識率計算に使用している正解ラベルファイルは，この書き起こし文を用いて作成されている．正解ラベルファイルを詳細に調査すると，幼児特有の表現の混入による形態素解析の失敗が多数生じ，単語のつながりが不安定である事が確認できる．さらに，幼児の発話は発達段階であるため，個人差による発声できない音が多数存在し，ある一つの単語に対して複数の言い回しが存在する．音声対話システムでの適応を考えると，音声認識部は必ずしも実際の発声にそった認識結果を出力することはない．後段の処理を考えると，成人が発話するような標準発話を出力をする方が好ましい．本稿では，正解ラベルファイルを作成する際に用いる書き起こし文を標準発話に変更し，単語の表現をそろえる事で認識率がどのように変化するかを調査する．

3.2 発音パターンの追加

正解ラベルファイルを標準発話に修正したことにより，認識時に入力される幼児発話を標準発話に補正するための手段が必要になる．本稿では，デコード時に使用する発音辞書に幼児の発音を追加し，幼児発話が入力された際に発話辞書を通して標準発話表現で出力する事で上記の問題を解決する．発音の追加手順を以下に示す．

（1） 幼児音声・学習データ書き起こし文（重複をなくし，出現頻度順にソートしたもの，約 660 文）を標準発話に変換し，

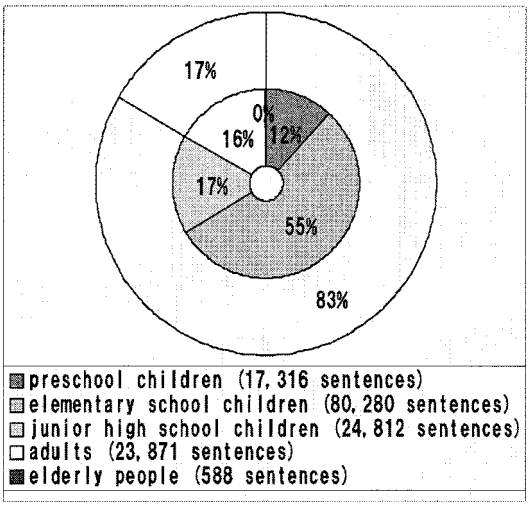

図 1 たけまるデータ内訳

表 1　コンテキスト非依存・発音パターン表 (一部)

標準発話	幼児表現	出現頻度	相対頻度
ザ	ジャ	6	0.0845
ス	シュ	187	0.0566
サ	シャ	43	0.0336

表 2　コンテキスト依存・発音パターン表 (一部)

標準発話			幼児表現			出現頻度
デ	シュ	カ	デ	ス	カ	139
デ	チュ	カ	デ	ス	カ	100
デ	チュ	.	デ	ス	.	41

- 幼児発音と標準発音の変化の対応
- 発音パターンの出現頻度
- 発音パターンの相対頻度

を記述した発音パターン表 (表 1) を作成する．発音パターン表は 2 種類用意する．一方は標準発話・幼児表現間の変化する音のみを記録したコンテキスト非依存の表 (表 1)，もう一方は変化する音の前後の音もあわせて記録したコンテキスト依存の表 (表 2) であり，本稿ではそれぞれの発音パターンを使用した際の認識結果を比較する．相対頻度とは，ある幼児発音パターンの出現頻度を，全学習データに出現する標準発音パターンの総数で除算したものである．

（2） 発音パターン表を用いて，自動的に発音辞書に単語を追加する．一つの単語には一つの発音パターンのみ適応する．発音パターンの追加方法を以下に示す．

- 出現頻度順にパターン追加
- 相対頻度順にパターン追加

上記の手順で発音を追加して認識結果で標準発話に直す事で，幼児発話の特徴を補正する．

本稿では，使用する発音パターンを増やすごとに新たに発話辞書を作成し，上記のパターン追加方法によって認識精度にどのような変化があるか調査する．

3.3　音響モデル作成

現在幼児音声認識で使用している音響モデルは，成人が発話した新聞記事読み上げコーパス JNAS を用いて作成した音響モデルに幼児音声を用いて適応を行ったものである (JNAS-BASE 音響モデル)．よって幼児音声認識で使用するには音素体系が異なる可能性がある．本稿では，新たに幼児音声のみで音響モデルを構築し，認識性能を比較する．

本稿では，以下の 3 種類の方法で音響モデルを構築する．

- 全幼児データでの音響モデル構築 (flat-start)
- JNAS-BASE 音響モデルを初期モデルとし，尤度基準による選択学習により選択された音声データと幼児データを用いて学習
- 尤度基準による選択学習により選択された音声データと幼児データを用いた音響モデル構築 (flat-start)

ここで，尤度基準による選択学習とは，音響モデルを作成したいタスクのデータ (目的タスクデータ) を用いて，既存の別タスクデータベース (学習データプール) から目的タスクデータに特徴の似たデータを選別する手法である [6]．本稿では，目的タスクデータを幼児音声，学習データプールを幼児音声を除く全ての子供音声として選択学習を行い，音響モデル構築に使用可能な音声データを補うことで認識精度にどのように影響があるかを調査する．

3.4　言語モデル作成

現在，幼児音声認識で使用されている言語モデルは，以下のデータを用いて作成されている [7]．

（1） Web 検索を用いて収集した生駒市関連及び生駒市ホームページ内の Web ページテキスト:1,080,272[文]

（2） 人手で収集した本システムを想定した質問文テキスト:6,488[文]

（3） 収集発話の書き起こしテキスト:24,498[文]

幼児の発話体系は，成人の発話体系と比較すると柔軟であり，成人発話との規則の差異によって音声認識が困難になっている事が予想される．本稿では，音響モデルと同様，幼児に特化した言語モデルの構築を行う．以下の 5 種類のデータを用いて言語モデルを構築し，性能を比較する．

- 幼児データ使用 (11,327[文])
- 子供データ使用 (67,458[文])
- 幼児＋子供データ使用 (78,785[文])
- 子供＋大人データ使用 (82,357[文])
- 幼児＋子供＋大人データ使用 (93,724[文])

言語モデルの作成にあたり，SRILM [8], [9] を使用する．

4.　実験的評価

前節で述べた処理の効果を明らかにするために，実験的評価を行う．

4.1　実験条件

実験条件を表 3 に示す．本稿では，音響モデルとして PTM [10] モデルを用いる．言語モデルとしては，幼児用言語モデルが存在しないため，たけまる用子供用言語モデルを用いる．

4.3 節では，3.2 節末で述べたとおり，発音パターンを加えるごとに作成される発話辞書を用いる．実験に使用した発音パターンの総数は 50 パターンである．なお，使用する発話辞書の初期単語数は 43,502 単語である．これらの単語すべてに発音パターンを適応すると，語彙数が大量に追加され，認識に悪影響を及ぼす可能性がある．そ

表 3 実験条件

使用ツール	音響モデル作成	HTK ver.3.2
	言語モデル作成	SRILM
	認識デコーダ	Julius ver.3.5.1
音響モデル	ベースライン	既存幼児用 (JNAS-BASE, EM × 5, 1,965 状態, 64 混合)
	提案法	1) 幼児用 PTM (flat-start, EM × 10, 328 状態, 32 混合) 2) 選択学習を用いて作成した PTM
	特徴量	12MFCC+12Δ MFCC+Δ E
言語モデル	ベースライン	たけまる子供用
	提案法	各データを用いて作成 (3.4 章参照)
発音辞書	ベースライン	たけまる子供用:43,502[単語]
	提案法	変換パターンを使用し単語を増やした発音辞書
学習データ	幼児音声, 11,327[文]	
評価データ	幼児音声 (学習データに含まれない), 1,000[文]	

表 4 正解ラベルファイル変更の効果

条件	認識率 [%]
変更前	45.5
変更後	49.2

表 5 追加された単語数

初期単語数		43,502
コンテキスト非依存	単語頻度	47,678 (+4,176)
	相対頻度	48,849 (+5,347)
コンテキスト依存	単語頻度	44,379 (+877)

こで，本稿では，発音パターンを適応する単語の制約として，小学生が発話した単語（3,123 単語）に限定する．

4.2 正解判定基準の変更

実験結果を表 4 に示す．現在の幼児音声認識精度は 45.5%である．正解ラベルファイルの変更により，幼児音声認識精度は 49.2%にまで向上した．以降の実験では，正解ラベルファイル変更後の実験結果をベースラインとする．

4.3 発音パターンの追加

実験結果を図 2 に示す．始めに，出現頻度順に発音パターンを利用した際の認識結果について考察する．発音パターンを一つ適応した時点でベースラインの認識精度 (49.2%) を 0.8%上回った．その後，発音パターンを追加するごとに認識精度が向上し，最も認識精度が向上したのは発音パターンを 14 パターン追加した時で，認識精度は 52.2%であった．14 パターン以上発音パターンを追加すると徐々に認識精度は悪化し，最終的にすべての発音パターンを使用した場合の認識精度はベースラインを 0.2%下回った．以上の結果より，発音パターンを発音辞書に追加することは有効な手法であるが，むやみに単語を追加することは認識時に悪影響を及ぼす事が分かる．最も認識精度が向上した際の発話辞書の単語数は 47,678 単語であり，始めの単語数に比べて約 4,000 語ほど増加した．

相対頻度順に発音パターンを利用した際には，出現頻度順に発音パターンを利用した際と比較し，認識率増減の傾向がより安定している事が確認できる．最も認識精度が向上したのは発音パターンを 13,16,18,19 パターン追加した時で，認識精度は 51.6%であり，発話辞書の単語数は 48,849 単語 (19 パターン追加時) であった．

表 6 認識率が向上した際に使用した発音パターン

標準発話	幼児表現	出現頻度	相対頻度
ス	シュ	187	0.0566
ス	チュ	80	0.0242
サ	シャ	43	0.0336
デ	レ	29	0.0097
サ	チャ	28	0.0219
バ	ハ	24	0.0126
コ	ト	23	0.0072
イ	エ	22	0.0041
ハ	ア	20	0.0039
ス	シ	17	0.0051
ナ	マ	15	0.0034
サ	タ	13	0.0102
ス	ツ	13	0.0039
チ	ヂ	10	0.0043

コンテキスト依存発話パターンを使用した際の実験結果を図 3 に示す．最高認識精度はコンテキスト非依存の認識結果を同等であったが，追加される単語数が約 5 分の 1 に削減できる事が確認できた．

認識率が向上した際に使用した発音パターンを表 6 に示す．発音パターンの傾向を考察すると，

- 直音が拗音になる（例：です→でちゅ）
- 子音成分の置換

（例：どこですか→どこれすか）

- 母音成分の置換（例：おなまえ→おなまい）

の 3 つの規則に分類される事がわかる．この変換特徴は，文献 [3] に記載されている "幼児が正しく発音できない不

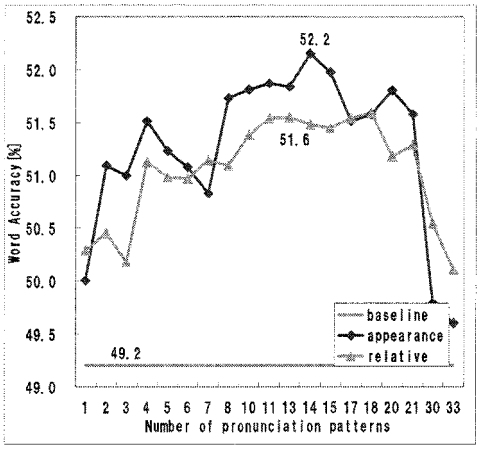

図 2　発音パターン追加による認識率の変化 (コンテキスト非依存発音パターン表使用)

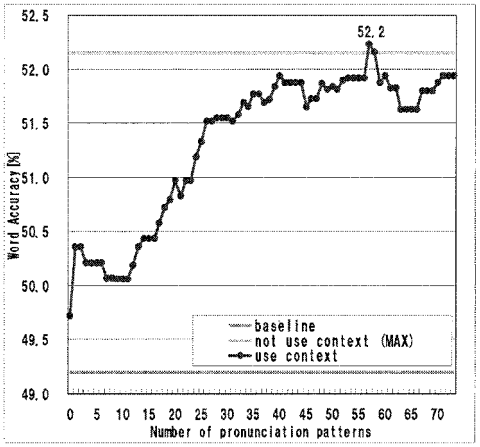

図 3　発音パターン追加による認識率の変化 (コンテキスト依存発音パターン表使用)

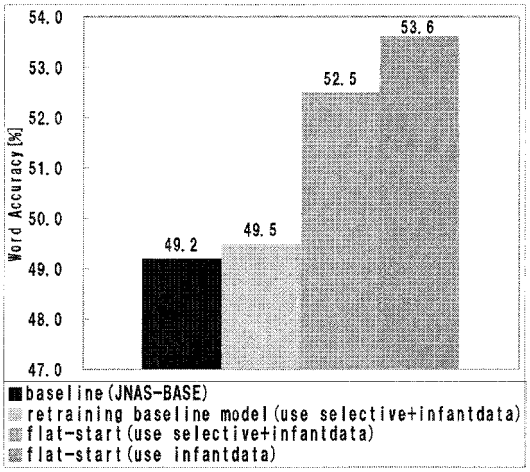

図 4　音響モデル変更による効果

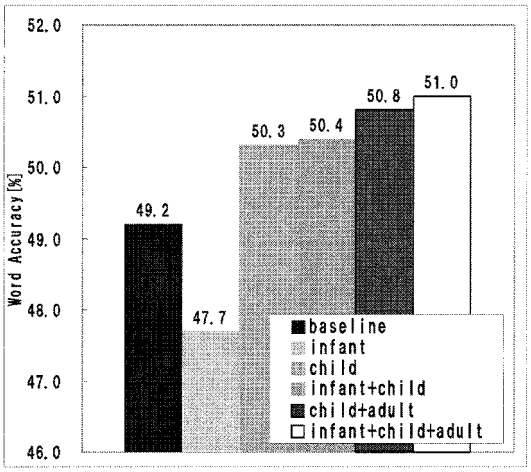

図 5　言語モデル変更による効果 (使用音響モデル:ベースライン)

完全な構音状態の例 " に一致し,「たけまるくん」に対して入力される自由発話内にはこれらの規則が頻繁に起こる事が確認できる.

4.4　音響モデル変更

実験結果を図 4 に示す. 新聞記事読み上げ文で作成した音響モデルに幼児データで適応行った音響モデルより, 一から幼児データのみを用いて作成した音響モデルの性能が勝り, 4%の認識率向上が確認できた. 選択学習を行う際には, 既存音響モデルの学習を行うより, 選択されたデータを用いて一から音響モデルを構築するほうが良い認識率を得られたが, 幼児データのみを用いた音響モデルの性能には至らない事が確認できた. 理由として, 選択学習によって追加されたデータ数は 23,767[文] であり, 幼児データの 2 倍ものデータを追加しているため, 幼児以外の子供の影響を大きく受けることで, 幼児の特徴との差が生じた事が考えられる.

表 7　各言語モデルのパープレキシティ (PP), 総単語数, Out-Of-Vocabulary(OOV)

	PP	総単語数	OOV	OOV 率
infant	15.9	3,101	169	4.2
child	16.6	8,517	109	2.7
infant+child	15.9	9,294	76	1.9
child+adult	16.7	9,257	100	2.5
infant+child+adult	15.8	9,995	73	1.8

4.5　言語モデル変更

実験結果を図 5, パープレキシティ・OOV 率等を表 7

表 8　音響・言語モデル及び発音パターン追加の組み合わせ

実験条件	発音追加前 [%]	発音追加後 [%]
ベースライン	**49.2**	52.2
AM 更新 幼児用 PTM (flat-start)	53.6	**54.8**
LM 更新 幼児+子供+大人データ使用	51.0	51.9
AM・LM 更新 (上記 AM・LM を使用)	54.0	53.3

に示す．最高認識精度が得られたのは幼児＋子供＋大人データを用いて作成した言語モデルであり，ベースラインと比較すると約 2%の認識率向上が観測できた．言語モデル作成に使用するデータが増加するにつれて，OOV 率は減少するが，パープレキシティに関しては大きな変化は観測できなかった．幼児データのみで作成した言語モデルがたけまる全データで作成した言語モデルの性能より劣った考察として，言語モデルを作成するにはデータ量が少ないこと，使用する正解ラベルファイルが標準発話のものであることが想定できる．また，既存の言語モデルには Web 上から収集したテキストと質問想定文を使用していたため，それらのデータが幼児の発話内容と異なっていたことなどが想定できる．

4.6　音響・言語モデル及び発音パターン追加の組み合わせ評価

本稿の 3.3 節，3.4 節で最も良い認識精度を示した音響モデル・言語モデルを用い，各モデルを併用する際の認識率の変化について評価する．実際に使用する各モデルの詳細・実験結果を表 8 に示す．結果として，新たに幼児用に作成した音響モデル・言語モデルの併用が最も良い認識精度を示し，ベースラインに比べて約 5%の認識率向上が観測できた．言語モデル変更よりも，音響モデル変更の効果が高く，幼児音声認識に対する音響的特徴の定義の重要性が確認できる．

発音パターンを追加した辞書を用いた際の実験結果を表 8 にあわせて示す．音響モデル変更と発話パターンを追加した既存発話辞書を併用する時，最高 5.6%の認識率向上が確認できた．言語モデルの変更と発話パターン手法を併用する際は，言語モデルを構築する時に作成される発音辞書に対して発話パターンを適応している．既存発話辞書と新たに作成した発話辞書の中に含まれる単語の違いと，それらの単語に発話パターンを使用することによって生成される単語によって，認識率が変動することが確認できる．また，使用する発話辞書により追加に適した発話パターンが異なることが考察できる．

5. ま と め

本稿では幼児発話の特徴を実験データから調査し，音声の変化を考慮した認識手法を提案した．また，幼児に特化した音響モデル・言語モデルを構築することによりさらなる認識率向上に努め，提案法を用いて実験的評価を行った．提案法を用いることにより，従来の幼児音声認識精度から 5.6%向上した．認識率向上に伴い，実環境音声対話システムでどの程度性能向上が観測できるかを調査する必要がある．音声対話システムの性能評価尺度として，応答正解率があげられている．今後の課題として，応答正解率と認識率の関係を調査する必要がある．

a）謝　　辞

本稿は，文部科学省のリーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われたものである．

文　　献

[1] Shrikanth Narayanan et al., "Creating Conversational Interfaces for Children", IEEE Transactions on Speech and Audio Processing, vol.10, no.2, 2002.
[2] http://www.lang.astem.or.jp/CSRC/
[3] 野田 雅子, 岩村 由美子, 内藤 啓子, 飛鳥井 きよみ, "幼児の構音能力の発達に関する研究", 日本総合愛育研究所紀要 第 4 集, pp.153-170, 1968.
[4] 小川 厚徳, 山口 義和, 松永 昭一, "小学生音声データベースを用いた子供音声認識の検討", 信学論, Vol.J87-DII, No.8, pp.1572-1580, 2004.
[5] 西村 竜一, 西原 洋平, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏, "実環境研究プラットホームとしての音声情報案内システムの運用", 信学論, Vol.J87-DII, No.3, pp.789-798, 2004.
[6] Tobias Cincarek, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Utterance-based Selective Training for the Automatic Creation of Task-Dependent Acoustic Models," IEICE Trans. Information and Systems, Vol.E89-D, No.3, pp.962-969, 2006.
[7] Ryuichi Nisimura, Kumiko Komatsu, Yuka Kuroda, Kentaro Nagatomo, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano: "Automatic N-gram Language Model Creation from Web Resources", Eurospeech2001 (7th European Conference on Speech Communication and Technology), pp.2127-2130, Aalborg, Denmark, 2001.
[8] http://www.speech.sri.com/projects/srilm/
[9] A. Stolcke, " SRILM - An Extensible Language Modeling Toolkit, " in Proc. ICSLP, pp.901-904, 2002.
[10] 李 晃伸, 河原 達也, 武田 一哉, 鹿野 清宏, "Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識", 信学論, Vol.J83-DII, No.12, pp.2517-2525, 2000.