

話し言葉音声認識における話者間の認識率変動要因の解析

篠崎 隆宏[†] 古井 貞熙[†]

[†] 東京工業大学 大学院情報理工学研究科
〒 152-8552 東京都目黒区大岡山 2-12-1

E-mail: †{staka,furui}@furui.cs.titech.ac.jp

あらまし 話し言葉音声認識の認識性能は十分ではなく、また話し言葉の音声認識における単語正解精度低下の要因はあまり分かっていない。そこで、日本語話し言葉コーパスの多数の話者を対象に音声認識実験を行い、話者間での単語正解精度変動の分布の解析を行った。個人差の主たる要因が発話速度、未知語率および言い直し頻度であることを明らかにした。また、教師無し話者適応化は単語正解精度の向上に効果的に働くが、適応化を行った後も発話速度の影響は減少しないことを示した。

キーワード 話し言葉音声認識、日本語話し言葉コーパス、個人差、教師無し話者適応

A statistical analysis of individual differences in spontaneous speech recognition performance

Takahiro SHINOZAKI[†] and Sadaoki FURUI[†]

[†] Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Ookayama 2-12-1, Meguro-ku, Tokyo, 152-8552 Japan

E-mail: †{staka,furui}@furui.cs.titech.ac.jp

Abstract This paper reports results of various investigations on recognizing spontaneous presentation speech. Individual differences in the speech recognition performances are analyzed. A restricted set of the speaker attributes comprising the speaking rate, the out of vocabulary rate and the repair rate is found to be most significant to yield individual differences in the word accuracy. It is shown that unsupervised MLLR speaker adaptation works well for improving the word accuracy but does not compensate for the effect of the speaking rate.

Key words spontaneous speech recognition, Corpus of Spontaneous Japanese, individual differences, unsupervised speaker adaptation

1. はじめに

話し言葉のためのモデルと技術の構築を目指し、開放的融合研究推進制度によるプロジェクト「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」が、1999年から5年間の計画で行われている[1]。プロジェクトでは話し言葉を集め、大規模コーパス「日本語話し言葉コーパス(CSJ: Corpus of Spontaneous Japanese)」の構築を行っている。これまでの研究でCSJを用いて学習した音響モデル・言語モデルを使用した認識システムは、講演音声の認識に対して従来のシステムと比較して極めて優れていることが示されている[2]。しかし、単語正解精度は音響モデルの教師無し話者適応を行った場合で70%程度であり、いまだ不十分である。

話し言葉音声が書き言葉の読み上げ音声に比較して音声認識が難しい理由としては様々な要因が考えられる。これまでに音声認識の観点から話し言葉と書き言葉を比較した研究としては、模擬対話音声におけるフィラーや言い直しの調査[3]、模擬対話音声および読み上げ音声から学習した音響モデルの比較[4]、講義音声および読み上げ音声に基づくシステムの講義音声認識における単語正解精度の比較[5]などが挙げられる。また話し言葉の特徴である発話速度の変化への対処法の研究として、音素モデルおよび発音辞書の発話速度に応じた多重化[6]、シラブルベースでの認識[7]などが挙げられる。しかし話し言葉において、発話の様々な個人差を考慮したとき、それらが単語正解精度にどのように影響しているのかはあまり研究されていない。

そこで、発話スタイルの種々の音響的・言語的な個人差が話し言葉音声の音声認識性能にどう影響を与えるか、CSJ中の多数の話者を対象とした認識実験を基に要因の分析を行う。また、音響モデルには不特定話者モデルおよび教師無し話者適応モデルを使用し、話者適応化が個人差の分布に与える影響も併せて分析する。

2. 日本語話し言葉コーパス(CSJ)

CSJには約7M語700時間の話し言葉が収録される予定である。主な収録対象は学会講演や模擬講演、ニュース解説などのモノローグである。模擬講演はコーパスの収録のために行われた、一般的な話者による10分程度の日常的な話題のスピーチである。録音された音声に対し、人手により書き起こしが作成される。書き起こしには時間情報やフィラーや言い直

しなど、発声の属性を示す付加情報が含まれる。CSJにおいて「言い直し」とされる発声は、言い直された単語の断片および助詞、助動詞である。書き起こしには、かな漢字混じりの「基本形」と、発話に忠実なカタカナ表記の「発音形」がある。

3. 音声認識システム

3.1 認識タスク

話者間の個人差の解析のため、男性51名の話者による講演音声をテストセットとした。テストセット中の話者は全て異なる話者であり、また学習セットには含まれていない。分析には各講演の始めの10分間を用いた。表1にテストセットの概要を示す。

3.2 言語モデル

CSJにおいて既に書き起こしが得られている部分を試験的に使用した。言語モデルの作成に利用したのは「基本形」書き起こしテキストである。学習セットとして使用したのは、610講演であり、内訳は多い順に模擬講演が336、音響学会が139、言語処理学会が63講演、その他72講演となっている。形態素数にすると約1.5Mのサイズがある。

形態素とその発音のペアをモデル化の単位として用い、2-gramと逆向き3-gramを作成した。語彙サイズは30kである。語彙は学習データ中の出現頻度が上位のものとした。3-gramのカットオフは1とした。

フィラーは単に通常の単語としてモデル化した。言い直しはN-gramで有効にモデル化することが難しいことから、学習テキストから取り除きモデル化は行わなかった。

3.3 音響モデル

音響モデルは不特定話者モデル“SI”と教師無し話者適応化モデル“SA”を用いる。どちらも状態共有 triphone HMMで、状態数は2000、混合数は16である。使用した音素は43種類である。

“SI”は言語モデルの学習に使用した講演のうちで男性話者による338講演、約59時間から学習したモデルである。“SA”はSIをもとに教師無し話者適応化を行ったモデルである。教師無し話者適応化は音響モデルにSIを用いて認識した結果をもとに、MLLRを用いてSIを各話者に適応化することにより行った。適応化はHMMの正規分布の平均のみに対して行つ

表1 テストセット(51名)の概要

Presentation	No. presentations
人工知能学会	32
日本音響学会	12
その他	7

た。MLLR ではモデル中の全正規分布を、予め 64 の葉を持つ 2 分木の葉に対応させることで分類しておき、学習時のデータ量に応じて使用するクラスを決定する方法を用いた。正規分布の分類は centroid-splitting により行った。

3.4 実験条件

音声は 16kHz で標本化、16 ビットで量子化を行った。音響パラメータは MFCC12 次元、 Δ ケプストラム 12 次元、対数パワーの 1 次差分の計 25 次元で、切り出した疑似的な文単位ごとに、平均ケプストラムによる正規化 (CMS) を行った。音響モデルの学習と話者適応には HTK2.2 [8] を使用した。

形態素解析には、NTT で開発された形態素解析ツール JTAG を使用した。言語モデルは、CMU SLM Tool Kit v2.05 [9] を使用して作成した。デコーダは Julius3.1 [10] を使用した。

認識実験の際、言語重み、挿入ペナルティは音響モデルと言語モデルの組合せ毎に最適化し、テストセット中では共通の値を用いた。

単語正解精度の計算は形態素を単位とし、フライヤーを含めて行った。その際、フライヤーのうち「あー」と「あ」などは同一のものとした。これは、これらの区別があまり重要ではないと考えられること、書き起こしの際、どちらの表記とするかはつきりしない場合も多いことによる。「えー」と「えーっと」などは別のものとして区別した。

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 以上の無音に対応する。

4. 話し言葉における話者間単語正解精度分布の構造

4.1 話者属性

音声認識に係わる話者(講演)の属性として以下の 7 種類を分析の対象とした。

Acc 講演の単語正解精度(%)。

AL 講演のフレーム平均音響ゆう度。

SR 講演の平均発話速度。(1 秒当たりの音素数)。

PP テストセットパープレキシティ。

OR 未知語率(%)。

FR フライヤー頻度。(正解文の単語数に対するフライヤーの%値)。

RR 言い直し頻度。(正解文の単語数に対する言い直しの%値)。

発話速度と音響ゆう度は正解音素列の強制アライ

メントの結果から、無音にマッチした部分を除いて求めた。パープレキシティは言語モデルに 3-gram を用いて求めた。未知語の予測は含めていない。単語がフライヤーや言い直しかどうかの判定には、CSJ の書き起こしに含まれているタグ情報を用いた。未知語率、パープレキシティの計算では、正解文として言い直しを除いたものを用いた。発話速度の計算では音響モデルに SI を用いた。

4.2 平均及び分散

テストセット 51 名の話者における各属性の平均と標準偏差を表 2 に示す。単語正解精度の平均は音響モデルに SI を用いた場合で 64.1%、SA を用いた場合で 68.6% であった。標準偏差は SI を用いた場合で 7.4%、SA を用いた場合で 7.5% であり、単語正解精度が話者により大きくばらつくことが分かる。

4.3 相関分析

各属性間の相関を示すため、表 3 に相関行列を示す。表において下三角行列は相関係数、上三角行列は有意確率を示す。太字で表記された相関係数は、5% 水準において有意であることを示している。

4.3.1 音響ゆう度と発話速度

音響モデルに SI を用いた場合、音響ゆう度と発話速度の相関係数は -0.54 である。図 1 に音響ゆう度と発話速度の散布図を示す。図では 2 乗誤差を最小にするようにフィットさせた直線を重ねて示してある。発話速度の速い話者で音響ゆう度が低下する傾向が観察される。他方、発話速度が非常に遅い場合であっても、音響ゆう度の低下は見られない。音響ゆう度を予測する発話速度に関して 1 次と 2 次の回帰モデルを赤池情報量基準 (AIC) [11] を用いて比較した場合、1 次のモデルの方が優れたモデルとなつた。講演を単位としてみた場合、発話速度が増加すると音響ゆう度が減少する直線的な関係があると言える。発話速度の増加とともにゆう度が下がる原因としては、調音結合の増加による音響的特徴の不鮮明化などが考えられる。教師無し話者適応を行った場合、音響ゆう度は全体的に上がるものの、発話速度との負の相関関係は残ることが図 1 より分かる。

4.3.2 パープレキシティと言語的属性

パープレキシティと未知語率の相関係数は 0.52 である。図 2 にパープレキシティと未知語率の散布図を示す。未知語率の高い講演ではパープレキシティも高い傾向があることが分かる。

フライヤー頻度とパープレキシティの相関係数は -0.18 であり、ほぼ相関は無いと言える。言い直し頻度とパープレキシティの相関係数は 0.06 である。パープ

表 2 各属性の平均と標準偏差

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Mean	64.1	68.6	-55.5	-53.1	15.0	227	2.09	8.60	1.54
Standard deviation	7.4	7.5	2.3	2.2	1.3	63	1.18	3.67	0.73

表 3 相関係数行列; 下三角行列は相関係数、上三角行列は有意確率を示す。5%水準において有意となる相関係数は、太字で表記した

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR	
Acc(SI)			-	4.7%	-	0.2%	0.4%	0.0%	0.6%	3.4%
Acc(SA)		-	-	-	2.2%	0.1%	1.9%	0.0%	0.5%	2.5%
AL(SI)	0.28	-	-	-	-	0.0%	55.8%	12.0%	6.8%	53.9%
AL(SA)	-	0.32	-	-	0.0%	56.3%	8.3%	6.7%	33.4%	-
SR	-0.42	-0.47	-0.54	-0.62	-	92.0%	1.7%	0.0%	20.2%	-
PP	-0.40	-0.33	-0.08	-0.08	-0.01	-	0.0%	20.0%	69.4%	-
OR	-0.54	-0.51	-0.22	-0.25	0.33	0.52	-	0.3%	66.5%	-
FR	0.38	0.38	0.26	0.26	-0.50	-0.18	-0.41	-	33.8%	-
RR	-0.30	-0.31	-0.09	-0.14	0.18	0.06	-0.06	0.14	-	-

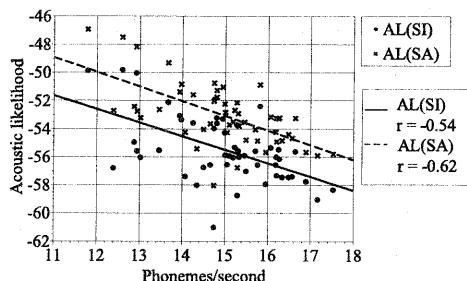


図 1 音響ゆう度と発話速度

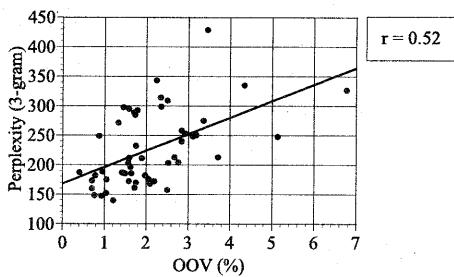


図 2 パーペルキシティと未知語率

レキシティは言い直しを除いた正解文を用いて計算していることから、この結果は言い直しを除いた部分の言語的な難しさはもともとの言い直し頻度とは関係ないことを示している。

4.3.3 単語正解精度と諸属性

発話速度と単語正解精度 (SI) の相関係数は-0.42である。図 3 に発話速度と単語正解精度の散布図を示す。発話速度と単語正解精度の関係は単調であり、図 1 における発話速度と音響ゆう度の場合と同様に発話速度が非常に遅い場合でも単語正解精度が下がらない様子が観察される。AICにおいても単語正解

精度を予測する 1 次と 2 次の発話速度を用いた回帰式では、1 次のモデルが選択された。発話速度と単語正解精度 (SA) の相関係数は-0.47であり、教師無し適応化を行っても相関係数は減少しないことが分かる。

音響モデルに SI を用いた場合の音響ゆう度と単語正解精度との相関係数は 0.28 であり 5% 水準で有意であるが、発話速度を制御した場合の単語正解精度と音響ゆう度の偏相関係数は 0.07 と、小さな値となった。音響ゆう度を制御した単語正解精度と発話速度の偏相関係数は-0.33 であり、5% 水準で有意である。また、単語正解精度を制御した音響ゆう度と発話速度の偏相関係数は-0.48 であり、1% 水準で有意である。このことは、音響ゆう度と単語正解精度の相関は発話速度を介したみかけの相関であることを示している。発話速度の増加は、音響ゆう度と単語正解精度をそれぞれ独立に低下させていると言える。音響モデルに SA を用いた場合の音響ゆう度と単語正解精度との相関係数は 0.32 である。偏相関係数を用いた分析に関しても SI の場合と同様の結果となつた。

言い直し頻度と単語正解精度 (SI) の相関係数は-0.30 である。図 4 に言い直し頻度と単語正解精度 (SI) の散布図を示す。

フィラー頻度と単語正解精度 (SI) の間には相関係数 0.38 の正の相関が見られる。しかし、発話速度を制御したフィラー頻度と単語正解精度 (SI) の偏相関係数は 0.22 であることから、この相関はみかけの相関であることが分かる。フィラー頻度を制御した発話速度と単語正解精度の偏相関係数は-0.29 で、5% 水準で有意である。単語正解精度を制御したフィラー

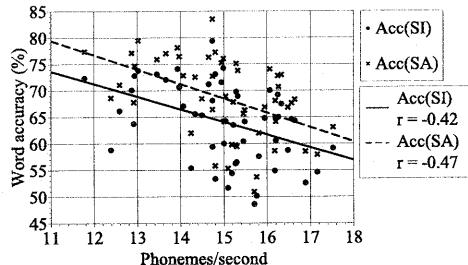


図 3 発話速度と単語正解精度

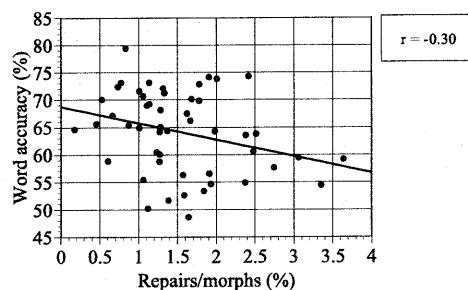


図 4 言い直し頻度と単語正解精度 (SI)

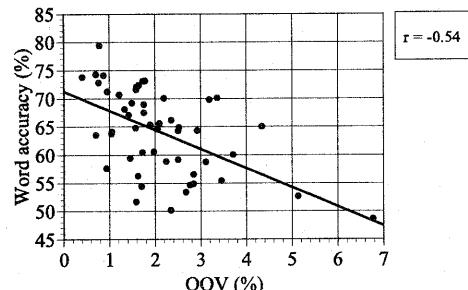


図 5 未知語率と単語正解精度

頻度と発話速度の偏相関係数は-0.40であり、1%水準で有意である。

図 5 に未知語率と単語正解精度 (SI) の散布図を示す。未知語率と単語正解精度の相関係数は-0.54である。

パープレキシティと単語正解精度の間には-0.40の相関係数があるが、これもみかけの相関である。未知語率を制御したパープレキシティと単語正解精度の偏相関係数は-0.16である。パープレキシティを制御した未知語率と単語正解精度の偏相関係数は-0.43、単語正解精度を制御したパープレキシティと未知語率の偏相関係数は0.39である。

以上の相関関係、みかけの相関関係を図 6 に示す。

4.4 重回帰分析

音響モデルに SI および SA を用いた場合の単語正

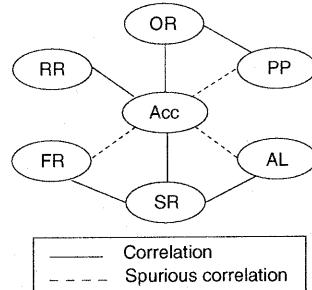


図 6 相関関係とみかけの相関関係

解精度を諸属性から予測する重回帰式を式 (1) より (2) に示す。

$$\begin{aligned} Acc_{SI} &= 0.12AL_{SI} - 0.88SR \\ &\quad - 0.020PP - 2.20OR + 0.32FR \\ &\quad - 3.0RR + 95 \end{aligned} \quad (1)$$

$$\begin{aligned} Acc_{SA} &= 0.024AL_{SA} - 1.3SR \\ &\quad - 0.014PP - 2.10OR + 0.32FR \\ &\quad - 3.2RR + 99 \end{aligned} \quad (2)$$

式 (1)において言い直し頻度の係数は-3.0である。このことは1%の言い直し頻度の増加が3.0%の単語正解精度の低下に相当することを示している。同様に、未知語率の係数は-2.2であり、1%の未知語率の増加は2.2%の単語正解精度の低下に相当する。これらは言い直しや未知語による1つの認識誤りが、言語的な繋がりにより2次的な誤りをひき起こすためと考えられる。

回帰式の決定係数は式 (1) の場合で0.48、式 (2) の場合で0.47である。このことは話者間の単語正解精度分布の分散の約半分が、回帰式により説明されることを示している。回帰式 (1) と (2) を比較すると、音響モデルの教師無し話者適応を行う前と後で、定数項が増加していること、発話速度の係数の大きさが減少しないこと等が分かる。

説明変量の影響力の大きさを見るため、各変量を平均と分散で正規化した後、回帰分析を行った、標準偏回帰係数、有意確率および95%信頼区間を表4に示す。係数の値が比較的小さくなった音響ゆう度、パープレキシティおよび未知語率は4.3節において偏相関係数を用いた分析で示したように、単語正解精度との直接の相関が小さい変量である。発話速度に関しては適応化を行った音響モデル SA を用いた場合の方が係数が大きくなつた。

表 4 単語正解精度に対する標準偏回帰分析. 標準偏回帰係数 (Coeff), 有意確率 (P), 95%信頼区間 (95% CI) を示す

	Coeff(SI)	P	95% CI		Coeff(SA)	P	95% CI
AL(SI)	0.04	76.7%	(-0.22, 0.30)	AL(SA)	0.01	96.1%	(-0.28, 0.29)
SR(SI)	-0.16	32.8%	(-0.47, 0.16)	SR(SA)	-0.23	19.0%	(-0.57, 0.12)
PP	-0.17	20.1%	(-0.44, 0.10)	PP	-0.11	40.4%	(-0.39, 0.16)
OR	-0.34	2.2%	(-0.63,-0.05)	OR	-0.32	3.2%	(-0.62,-0.03)
FR	0.16	24.9%	(-0.11, 0.43)	FR	0.16	26.0%	(-0.12, 0.44)
RR	-0.30	1.4%	(-0.53,-0.06)	RR	-0.31	1.3%	(-0.54,-0.07)

4.5 主要属性の分析

重回帰モデルにおいて単語正解精度を予測する上で重要な説明変量を特定するため、変数減少法を用いた分析を行った。変数減少法ではまず全ての説明変量を含む回帰式を求める。回帰式中で一番大きな有意確率を持つ説明変量を1つ取り除き、残った説明変量を用いて回帰式を再計算する。変数を取り除く作業を回帰式中の全ての説明変量の有意確率が0.10以下になるまで繰り返す。この操作において回帰式に残った説明変量は、音響モデルにSIを用いた場合もSAを用いた場合も同じであり、発話速度、未知語率、言い直し頻度であった。これらの変数はいずれも相関分析の結果として図6に示した単語正解精度と直接の相関を有する属性であり、音響モデルにSIを用いた場合の発話速度を除けば、表4において比較的大きな係数を持つ変数に一致している。

同定された3属性のみを用いた重回帰式の決定係数は、音響モデルにSIおよびSAを用いた場合とも0.44であり、6種類全ての説明変量を用いた場合とほぼ同じであった。単語正解精度の個人差の主たる要因は発話速度、未知語率および言い直し頻度であるといえる。

5. まとめ

本論文では、自由発話された講演の音声認識において、発話スタイルの個人差の様々な要素が単語正解精度にどのように影響を与えていているか、51人の話者を用いた認識実験を基に解析を行った。種々の話者(講演)の属性のうちで、発話速度、未知語率および言い直し頻度が単語正解精度に与える影響が大きいことを示した。逆に、正解文の音響ゆう度やテストセットパープレキシティと、単語正解精度との直接の相関は小さいことを示した。MLLRを用いた教師無し話者適応は単語正解精度の向上に効果的に働くものの、適応化を行っても発話速度の影響は減少しないことを示した。発話速度に対応するためには別の対処法が必要である。種々の話者属性を考慮した重回帰式により、単語正解精度の分散の約半分が説

明された。また、単語正解精度の説明において効果の大きい変量として同定された3属性のみを用いた重回帰式においても、ほぼ同じ説明力が得られた。

今後の課題としては更に広い範囲の発話を分析すること、人による講演の主観評価とデコーダーによる認識性能の関係へと分析対象を広げること、本論文において同定された単語正解精度に対する影響の大きい属性への対処法を開発することが挙げられる。

文献

- [1] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira. Toward the realization of spontaneous speech recognition. In *Proc. ICSLP*, Vol. 3, pp. 518-521, Oct 2000.
- [2] 篠崎隆宏, 細川貴生, 古井貞熙. 話し言葉コーパスを用いた音声認識の検討. 2001春季音学講論集, pp. 31-32, 2001.
- [3] 村上仁一, 嶋峨山茂樹. 自由発話音声における音響的な特徴の検討. 信学論, Vol. J78-D-II, No. 12, pp. 1741-1749, 12 1995.
- [4] 山本一公, 中川聖一. 発話スタイルによる話速・音韻間距離・ゆう度の違いと音声認識性能の関係. 信学論, Vol. J83-D-II, No. 11, pp. 2438-2447, 11 2000.
- [5] 西村雅史, 伊東伸泰. 講義コーパスを用いた自由発話の大語彙連続音声認識. 信学論, Vol. J83-D-II, No. 11, pp. 2473-2480, 11 2000.
- [6] S. Zheng, H. Franco, F. Weng, A. Sankar, and H. Bratt. Word-level rate of speech modeling using rate-specific phones and pronunciations. In *Proc. ICASSP*, Vol. 3, pp. 1775-1778, June 2000.
- [7] H. Nanjo, K. Kato, and T. Kawahara. Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition. In *Proc. Eurospeech*, Vol. 4, pp. 2531-2534, Sept 2000.
- [8] Entropic Ltd. *The HTK Book (for HTK Version 2.2)*, 1999.
- [9] P. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proc. Eurospeech*, Vol. 5, pp. 2707-2710, Sept 1997.
- [10] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proc. ICSLP*, Vol. 4, pp. 476-479, Oct 2000.
- [11] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proc. ISIT*, pp. 267-281, 1973.