音声言語情報処理 39－1
(2001. 12. 20)

音声認識における高精度な動的特徴量計算法の提案

小早川 健† 世木 寛之† 松井 淳† 尾上 和穂† 佐藤 庄衛†
今井 亨† 安藤 彰男†

† NHK 放送技術研究所
〒157-8510 東京都世田谷区砧 1-10-11

E-mail: †{kobayakawa.t-ko,segi.h-gs,matsui.a-hk,onoe.k-ec,satou.s-gu,imai.t-mq,andou.a-io}@nhk.or.jp

あらまし 本研究では音声認識における新たな動的特徴量の計算法を提案する．提案法では，1) 動的特徴量の推定に用いる静的特徴量の時間分解能を上げ，2) 動的特徴量の推定時間をフレーム間隔の整数倍に拘束されることなく任意に設定可能とした．提案法をニュース番組で発声される音声の認識実験によって評価したところ，特徴量を推定する時間を適切に選べば，認識率の改善が見られることが分かった．単語正解精度でみた誤認識改善率は，雑音を含むニュース文の評価セットで 24%と効果が大きかった．全体での誤認識改善率は 6.5%であった．

キーワード 音声認識，動的特徴量，回帰係数

A method for improving the accuracy of dynamic features calculation in speech recognition

Takeshi S. KOBAYAKAWA†, Hiroyuki SEGI†, Atsushi MATSUI†, Kazuo ONOE†, Shoe SATO†, Imai TORU†, and Akio ANDO†

† NHK Science & Technical Research Labs.
1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan

E-mail: †{kobayakawa.t-ko,segi.h-gs,matsui.a-hk,onoe.k-ec,satou.s-gu,imai.t-mq,andou.a-io}@nhk.or.jp

Abstract A new method for estimating dynamic features used in speech recognition is proposed. The proposed method involves: 1) increasing time resolution of static features used for dynamic feature estimation, 2) relaxing the restriction that estimation span must be an integer multiple of frame shift, and making the estimation span arbitrary. Evaluating the proposed method with a broadcast news testset, word accuracy improved by selecting appropriate estimation span. Error reduction rate in word accuracy was up to 24% relatively in noisy news speech. The overall error reduction rate was 6.5% relatively.

Key words Speech Recognition, Dynamic Features, Regression Coefficient

1. はじめに

NHKでは，音声認識による字幕自動付与の研究開発を行っている．音声認識による字幕付与は，ニュース番組においてアナウンサーによるスタジオ発声無雑音ニュース文では，実用化している．

一方，雑音が含まれる部分では，より一層の認識性能の改善が望まれている．雑音対策としては，雑音を推定して補償する方法と，耐雑音性のある特徴量を追求する方法に大別される．前者の雑音を推定する方法では，音響分析の段階に応じていくつかの方法がある．スペクトルの段階で処理を行うものとして，スペクトラルサブトラクション [1], [2] がある．また，ケプストラムの段階で処理を行うものとして，HMM分解合成法 [3] がある．後者の耐雑音性のある特徴量を追求する方法では，可変フレーム間隔による音響分析 [4], [5] などがある．

本研究報告では，特徴量そのものの高精度化の観点から特徴量の計算法を提案し，その有効性を検証する．特徴量抽出間隔を短くする方法は，雑音部分に有効であると思われる [6]．フレーム間隔の時間分解能を上げた場合に認識率が改善する要因として，

（1） 特徴量の (時間) 分解能が向上したこと

（2） 特徴量そのものが高精度化したこと

が考えられる．（1）の特徴量の (時間) 分解能を細かくする場合は，認識所用時間が増大してしまうのは止むを得ず，実時間処理の観点から課題を残している．（2）の特徴量そのものを高精度化する場合は，音響分析の処理時間が探索の時間と比較して十分短いことから，実時間処理の障害となるとは考えにくい．そこで，特徴量の高精度化を計りながら実時間処理を視野に入れた方法を提案する．

2. 提案法

音声認識では，静的特徴量とあわせて動的特徴量を用いる方法が，一般的になっている．本研究報告で提案する計算法は，動的特徴量の推定に関するものである．

動的特徴量の1つである静的特徴量の1次差分を例にとり，従来法を説明する．特徴量抽出間隔を Δt，時刻 t におけるフレームでの静的特徴量を $c(t)$ とすると，同時刻での差分 $d(t)$ は，前後 Θ フレームずつの静的特徴量 $c(t-\Theta\cdot\Delta t),\ldots,c(t),\ldots,c(t+\Theta\cdot\Delta t)$ を用いて，次式により推定する [7]．

$$d(t)=\frac{\sum_{\theta=1}^{\Theta}\{\theta\,(c(t+\theta\cdot\Delta t)-c(t-\theta\cdot\Delta t))\}}{2\sum_{\theta=1}^{\Theta}\theta^2} \quad (1)$$

動的特徴量の推定に用いる時間 $2\Theta\Delta t$ を動的特徴量の推定時間と呼ぶことにする [6]．

今回提案する方法では，デコードに用いる特徴量抽出間隔より短い間隔の静的特徴量を用いて，動的特徴量を推定する．デコードに用いる特徴量抽出間隔を Δt_1，動的特徴量の推定に用いる静的特徴量の時間間隔を Δt_2 とすると，

$$\Delta t_1 > \Delta t_2 \quad (2)$$

となるように選ぶ．このとき，$d(t)$ の推定式は，

$$d(t)=\frac{\sum_{\theta=1}^{\Theta_2}\{\theta\,(c(t+\theta\cdot\Delta t_2)-c(t-\theta\cdot\Delta t_2))\}}{2\sum_{\theta=1}^{\Theta_2}\theta^2} \quad (3)$$

となる．ここで，やはり前後 Θ_2 フレームずつの静的特徴量を用いて動的特徴量を推定する．

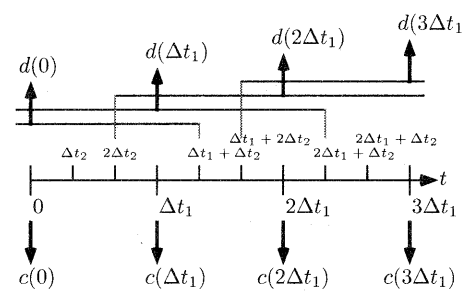

図 1 提案法の動的特徴量計算法

提案法を図 1 を用いて説明する．静的特徴量 $c(0), c(\Delta t_1), \ldots$ は，従来法で求める (図1下段)．動的特徴量の算出に用いる静的特徴量は，Δt_2 間隔で従来法で求める (図1上段)．ここで動的特徴量 $d(\Delta t_1)$ は，$c(\Delta t_1-\Theta_2\Delta t_2), \ldots, c(\Delta t_1), c(\Delta t_1+\Delta t_2), c(\Delta t_1+2\Delta t_2), \ldots, c(\Delta t_1+\Theta_2\Delta t_2)$ を用いて求める．このようにして，静的特徴量の系列 $c(0), c(\Delta t_1), c(2\Delta t_1), \ldots$ と，動的特徴量の系列 $d(0), d(\Delta t_1), d(2\Delta t_1), \ldots$ を求め，音響モデルの特徴量とする．

フレーム間隔を短くする場合には，Θ を固定したままでは推定時間が変化してしまい，認識性能を劣化させてしまう [6]．提案法で推定時間の変化による

劣化を防ぐためには，推定時間を一定に保ち，

$$2\Theta \cdot \Delta t \approx 2\Theta_2 \cdot \Delta t_2 \qquad (4)$$

となるように，パラメータを選ぶとよいと考えられる．本研究報告では，従来法の予備実験で良好な性能の得られた推定時間を中心として調べることにする．

3. 実　　験

2種類の認識実験を行った．実験1では，従来法の予備実験で良好な性能の得られた推定時間を採用し，Δt_2 を細かく設定した．実験2では，従来法で良好な結果が得られる推定時間よりも長い推定時間を採用した．推定時間は，従来法では Δt の整数倍にしか設定することができなかったが，提案法で Δt_1 の非整数倍に設定した．

音響分析して得られる特徴量としては，表1,2にあるものを用いた．Δt_1 は，従来法と同じく10ms

静的特徴量	ケプストラム (12次元 MFCC+対数パワー)
動的特徴量	Δ ケプストラム，ΔΔ ケプストラム

表1　実験に用いた特徴量

サンプリング周波数	16kHz
周波数分析窓	ハミング窓 25ms
プレエンファシス係数	0.97
フィルターバンク	24チャンネル
ケプストラムリフター	22

表2　ケプストラム算出の詳細

を採用した．

音響モデル，言語モデル，及びデコーダの諸元をそれぞれ，表3，4，5に示す．

音声認識タスクとしてニュースを中心とした放送音声を対象としている．認識実験に用いる評価セットは，2000年に放送されたもので，状況に応じて8つに分類されている．評価セットはすべて男性の発話で，詳細は表6に示す．評価セット E_1 から E_4 までは雑音を含まない音声である．評価セット E_5 から E_7 までは雑音を含む音声を含み，評価セット E_8 はS/N比が低い．評価セット E_5 から E_8 を雑音を含む評価セットとして扱う．なお，言語モデルは放送直前原稿を用いて適応化[9]されており，表6のパープレキシティ(pp.)と未知語率(OOV)は，語彙20kの場合の平均を示してある．

3.1 実　験　1

動的特徴量の推定時間を従来法の予備実験で良好な結果が得られた値と同じくし，動的特徴量の推定

種類	HMMによる音素 sub-word モデル
HMM	8混合ガウス分布 始・終状態を含む5状態 一方向性 jump なし
状態共有	tree-based clustering
環境音素	triphone (クロスワード無し)

表3　音響モデル

語彙	20k, 40k, 60k
種類	3-gram モデル
低頻度語の確率推定	Good-Turing による back-off スムージング

表4　言語モデル

種類	2パス方式
第1パス	音響モデル + 2-gram 言語モデル
第2パス	3-gram 言語モデル
状態内保存パス	4
第1パス出力文数	200
大域的ビーム幅	160
単語終端ビーム幅	110
言語重み	14
挿入ペナルティ	0
計算機	Compaq 21264(667 MHz) True64 UNIX(V5.1)

表5　デコーダ[8]

評価セット	文数 (単語数) pp. (OOV[%])	内容	雑音
E_1	339 (11,959) 5.0 (0.27)	アナウンサー発声 ニュース読み上げ	無
E_2	30 (667) 19 (3.9)	レポータ発声 ニュース文	無
E_3	139 (1,880) 21 (1.9)	スポーツ文	無
E_4	160 (3,316) 50 (5.0)	アナウンサー・ レポーターによる 自発発話	無
E_5	140 (4,244) 8.7 (0.39)	ニュース文	有
E_6	109 (3,428) 20 (1.1)	中継現場からの音声	有
E_7	254 (2,657) 51 (3.2)	スポーツ文	有
E_8	340 (3,453) 48 (1.2)	天気文	有

表6　評価に用いるニュース文1,511文の分類

に用いる Δt_2 を短くして実験を行った．実験で用いたパラメータ値は表7に示す．

音響モデル及び言語モデルの学習に用いた学習セットの概要を表8に示す．

認識実験の結果を表9に示す．雑音を含まない部分で改善する場合があったが有為な差ではなかった．

Δt_1	10ms
Δt_2	1.25ms
Θ_2	16
推定時間	40ms.

表 7　実験 1 に用いる提案法のパラメータ

音響モデル	男性アナウンサー発声の 雑音の無いニュース文 (20,892 文，54.6 h.)
言語モデル	'91.4.1 から放送直前までの ニュース原稿 1.87M 文 (76.2M 語)

表 8　実験 1 の各モデルの学習セット

他の部分では劣化する場合もあった．

評価セット	語彙	単語正解精度 (%) 従来法	提案法	誤認識改善率 (%)
E_1	(20k)	98.20	98.24	+2.22
	(40k)	98.29	98.31	+1.17
	(60k)	98.32	98.34	+1.19
E_2	(20k)	87.41	87.41	0
	(40k)	90.69	91.74	+11.28
	(60k)	90.99	92.19	+13.32
E_3	(20k)	87.61	86.76	-6.86
	(40k)	88.56	87.29	-11.10
	(60k)	89.15	87.56	-14.65
E_4	(20k)	76.14	76.82	+2.85
	(40k)	76.42	77.00	+2.46
	(60k)	76.40	77.05	+2.75
E_5	(20k)	95.10	94.77	-6.73
	(40k)	95.41	94.86	-11.98
	(60k)	95.38	94.79	-12.77
E_6	(20k)	88.83	88.54	-2.60
	(40k)	89.61	89.61	0
	(60k)	90.13	90.20	+0.71
E_7	(20k)	62.64	61.48	-3.10
	(40k)	63.60	62.22	-3.79
	(60k)	63.66	62.36	-3.58
E_8	(20k)	69.41	69.82	+1.34
	(40k)	69.73	70.36	+2.08
	(60k)	69.41	70.30	+2.91

表 9　実験 1 の単語正解精度

また，認識所要時間を表 10 に示す．ただし，ここでの認識所要時間は，音響分析を含まない．実験に使用した計算機は，表 5 に示してある．

3.2 実　験　2

実験 1 での推定時間付近で推定時間を変化させた．実験ではパラメータを連続的に変化させることはできないので，範囲を定めて離散的な推定時間値を 5 つ採用した．パラメータの具体的な値は，表 11 に示す．

音響モデル及び言語モデルの学習では，表 12 に示す学習セットを用いた．

以上の条件の下での認識率を図 2, 3 に示す．横軸に動的特徴量の推定時間をとり，縦軸に単語正解精度をパーセントで示してある．従来法より長い推定時間で認識率が改善する傾向を示している．Δt が 10ms. のときに (1) 式で実現可能な推定時間は 40ms., 60ms. であり，提案法はそれ以外の 56ms. 付近で改善を示した．

認識時間 56ms. の時の認識率の値を表 13 に示す．雑音環境下での改善効果が大きく，有為な改善を示す状況もあった．しかし，一部の状況では劣化も見られた．

また，認識所要時間を表 14 に示す．ここでも認識所要時間は，音響分析の所要時間をは含まれていない．

評価セット	語彙	認識所要時間 (× 実時間) 従来法	提案法
E_1	(20k)	0.46	0.47
	(40k)	0.45	0.46
	(60k)	0.49	0.49
E_2	(20k)	0.59	0.62
	(40k)	0.61	0.64
	(60k)	0.68	0.71
E_3	(20k)	0.52	0.51
	(40k)	0.53	0.56
	(60k)	0.60	0.64
E_4	(20k)	0.81	0.85
	(40k)	0.80	0.83
	(60k)	0.87	0.91
E_5	(20k)	0.64	0.61
	(40k)	0.68	0.70
	(60k)	0.75	0.78
E_6	(20k)	0.77	0.79
	(40k)	0.82	0.84
	(60k)	0.90	0.94
E_7	(20k)	1.03	0.99
	(40k)	1.05	1.11
	(60k)	1.16	1.26
E_8	(20k)	0.93	0.88
	(40k)	0.97	0.99
	(60k)	1.04	1.06

表 10　実験 1 の認識所要時間

Δt_1	10ms
Δt_2	1.00ms
Θ_2	20–30(20,23,25,28,30)
推定時間	40–60ms.(40,46,50,56,60ms.)

表 11　実験 2 に用いる提案法のパラメータ

音響モデル	男性 24 アナウンサー発声の 雑音の無いバランス文 (2,392 文，3.1 h.)
言語モデル	実験 1 に用いた言語モデルに同じ (表 8)

表 12　実験 2 の各モデルの学習セット

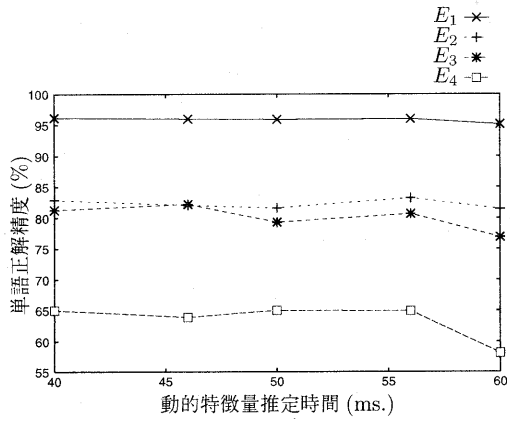

図 2　実験 2 の単語正解精度 (雑音の無い場合)

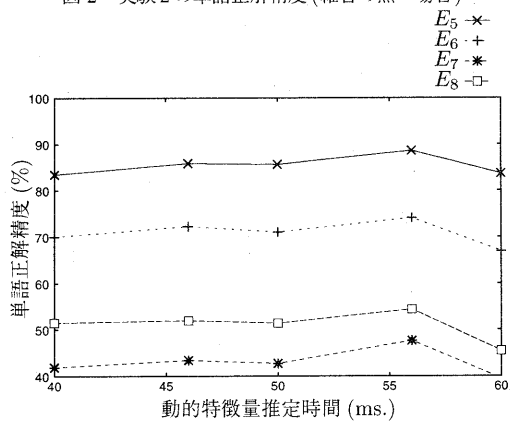

図 3　実験 2 の単語正解精度 (雑音を含む場合)

3.3 考　　察

従来法と提案法で，推定時間が同じ 40ms の場合に比較すると，雑音を含まない音声では提案法が若干性能が改善されたものの，他の場合では，従来法がよいという実験結果であった．この傾向は，他の認識実験でも同様であった．

一方，推定時間を 56ms. に設定すると，雑音を含む部分では，有為な認識率の改善が見られた．特に，評価セット E_5 単独では，24%の改善率であった．また，雑音を含まない部分では，有為な差でないものの劣化も見られた．しかし，評価セット E_1 から E_8 を含む全体での認識率は，表 15 に示すように 6.5%程度改善している．

本研究報告の実験 2 は，音響モデルの学習時間がバランス文のみで 3 時間分と短く，また，1 つのテストセットで開発用と評価用のセットを兼ねているため，より詳細な追実験が必要である．

参考のために，実験 2 に用いた音響モデルの状態数，triphone モデル数，最終 EM 学習時の平均出力確率を図 4 に示す．横軸に動的特徴量の推定時間，縦軸の左側に音響モデルの状態数と triphone モデル数，縦軸の右側に最終 EM 学習時の平均出力確率を示してある．学習時の tree-based clustering では，分割前後の音響尤度比でモデルの分割を行うかどうかを決定しており [10]，このパラメータは動的特徴量の推定時間に拘わらず一定とした．動的特徴量の推定時間を長くするにつれて，最終 EM 学習時の平均出力確率値はよくなる一方，状態数やモデル数は減少していく．動的特徴量の推定時間 60ms のところで状態数やモデル数が大幅に減少していることが，同条件での認識性能を劣化させている可能性もある．

評価セット	語彙	単語正解精度 (%) 従来法	提案法	誤認識改善率 (%)
E_1	(20k)	96.16	96.00	-4.17
	(40k)	96.19	96.06	-3.41
	(60k)	96.19	96.11	-2.10
E_2	(20k)	82.49	83.23	+4.23
	(40k)	85.16	86.06	+6.06
	(60k)	85.76	86.51	+5.27
E_3	(20k)	80.74	80.69	-0.26
	(40k)	80.96	82.03	+5.62
	(60k)	81.31	82.29	+5.24
E_4	(20k)	64.64	64.95	+0.88
	(40k)	64.85	65.36	+1.45
	(60k)	64.51	65.38	+2.45
E_5	(20k)	84.88	88.64	+24.87
	(40k)	84.81	88.31	+23.04
	(60k)	84.88	88.38	+23.15
E_6	(20k)	73.50	74.10	+2.26
	(40k)	74.09	74.74	+2.51
	(60k)	74.23	74.74	+1.98
E_7	(20k)	40.40	47.61	+12.10
	(40k)	40.61	47.65	+11.85
	(60k)	41.23	47.82	+11.21
E_8	(20k)	52.10	54.38	+4.76
	(40k)	52.86	54.47	+3.42
	(60k)	52.57	54.21	+3.46

表 13　実験 2 の単語正解精度

4. ま と め

動的特徴量の新しい計算法を提案し，その効果を放送音声で検証した．実験によると，雑音環境下で，認識性能の改善が見られた．その一方で，評価セットによっては認識性能がわずかに劣化する場合も見られた．全体としては，認識性能の改善が見られた．また，音響分析を除く認識所要時間は，従来法と同程度に押さえることもできた．

評価セット	語彙	認識所要時間 (× 実時間) 従来法	提案法
E_1	(20k)	0.24	0.23
	(40k)	0.28	0.29
	(60k)	0.35	0.32
E_2	(20k)	0.32	0.29
	(40k)	0.38	0.39
	(60k)	0.46	0.40
E_3	(20k)	0.31	0.30
	(40k)	0.38	0.39
	(60k)	0.45	0.43
E_4	(20k)	0.47	0.44
	(40k)	0.56	0.54
	(60k)	0.69	0.62
E_5	(20k)	0.43	0.38
	(40k)	0.51	0.47
	(60k)	0.61	0.54
E_6	(20k)	0.50	0.45
	(40k)	0.59	0.55
	(60k)	0.70	0.63
E_7	(20k)	0.61	0.56
	(40k)	0.73	0.71
	(60k)	0.87	0.79
E_8	(20k)	0.57	0.53
	(40k)	0.68	0.66
	(60k)	0.84	0.72

表 14　実験 2 の認識所要時間

評価セット	語彙	単語正解精度 (%) 従来法	提案法	誤認識 改善率 (%)
E_1	(20k)	78.17	79.58	+6.46
|	(40k)	78.43	79.83	+6.49
E_8	(60k)	78.47	79.87	+6.50

表 15　全評価セットでの単語正解精度 (実験 2)

4.1 おわりに

現段階での提案法の評価は，音響モデルの学習データが少ないので，より多くの学習データを用いて評価を行う必要がある．学習データとして，バランス文以外にいろいろな状況の放送音声を追加することを検討したい．

推定時間の影響もより詳細に調べる必要がある．現段階では，40–60ms. の中の 5 つの代表値のみを調べている．40ms. より短い場合の影響や，実験 2 でよい性能を示した 56ms. 付近の詳細な調査などは，今後の課題となる．また，認識タスクによって，推定時間が違うのかどうかも重要な問題である．

実用を考える場合，認識所要時間をより正確に見積もる必要も生じる．本研究報告での認識所要時間は，音響分析を除いた部分の実測値であり，提案法の音響分析が要する処理時間を考慮する必要が生じる．

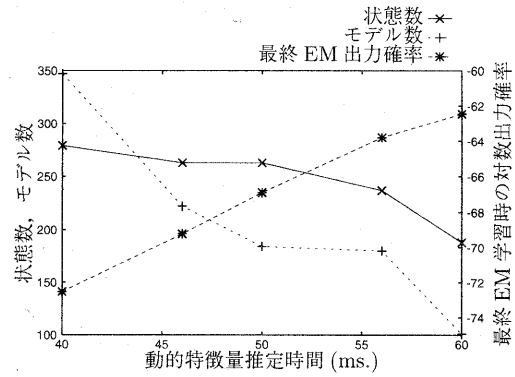

図 4　状態数，モデル数，最終 EM 学習時の対数出力確率

最後に，Δt_2 の選び方を考える必要がある．Δt_2 が小さい程よいのかは，興味深い問題である．もし，小さければ小さい程よいのであれば，Δt_2 が 0 のときの極限値を用いた特徴量を採用することも考えられる．具体的には，次のようにする．動的特徴量の推定時間 T_e とおき，T_e と Δt_2 を含む形で，動的特徴量を $d(t;\Delta t_2, T_e)$ と表す．このとき，新しく採用する動的特徴量 $d_l(t;T_e)$ は，$\lim_{\Delta t_2 \to 0} d(t;\Delta t_2, T_e)$ を用いる．この特徴量の可能性も今後検討していきたい．

文　献

[1] S. F. Boll: "Supression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, **27**, 2, p. 113 (1979).

[2] 尾上, 世木, 小早川, 佐藤, 今井, 安藤：“フィルタバンク・サブトラクションを用いたニュース番組現場リポート音声の認識”, 信学技報, SP2001-31 (2001).

[3] M. J. F. Gales, S. J. Young：“An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise” (1992).

[4] K. M. Ponting, S. M. Peeling：“The use of variable frame rate analysis in speech recognition” (1991).

[5] Q. Zhu, A. Alwan：“On the use of Variable Frame Rate Analysis in Speech Recognition” (2000).

[6] 小早川, 世木, 尾上, 小林, 今井, 安藤：“短いフレーム間隔による連続音声認識の検討”, 日本音響学会講演論文集 (秋), 1-Q-12 (2001).

[7] S. Furui: “Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum”, IEEE Transactions on Acoustics, Speech, and Signal Processing, **34**, 1, p. 52 (1986).

[8] 今井, 小林, 尾上, 安藤：“ニュース番組自動字幕化のための音声認識システム”, 音声言語情報処理研究会, SLP-23-11, p. 59 (1998).

[9] 小林, 今井, 安藤, 中林：“ニュース音声認識のための時期依存言語モデル”, 情処学論, **40**, 4, p. 1421 (1999).

[10] 世木, 尾上, 佐藤, 今井, 安藤：“状態共有トライフォン hmm の学習における決定木とモデル数の検討”, 日本音響学会講演論文集 (秋), 3-Q-3 (1999).