A Frame-Based Context-Dependent Acoustic Modeling for Speech Recognition

Ryuta Terashima Non-member (Nagoya Institute of Technology, Toyota Central R&D Labs., INC. ryuta@sp.nitech.ac.jp)
Heiga Zen Non-member (Nagoya Institute of Technology, Toshiba Research Europe Ltd, heiga.zen@crl.tochiba.co.uk)
Yoshihiko Nankaku Non-member (Nagoya Institute of Technology, nankaku@sp.nitech.ac.jp)
Keiichi Tokuda Non-member (Nagoya Institute of Technology, tokuda@nitech.ac.jp)

Keywords: speech recognition, acoustic model, context clustering

We propose a novel acoustic model for speech recognition, named FCD (Frame-based Context Dependent) model. It can obtain a probability distribution by using a top-down clustering technique to simultaneously consider the local frame position in phoneme, phoneme duration, and phoneme context The model topology is derived from connecting left-to-right HMM models without self-loop transition for each phoneme duration.

Fig. 1 shows an example of the phoneme FCD model.

The left white node is the head node of this model, and the transition probabilities from this to the colored nodes correspond with the probabilities of the phoneme duration. The character of each node is "state ID", which means that the nodes with same ID are tied. This example clearly shows that FCD model can change the probability distribution into a sequence corresponding with one phoneme duration

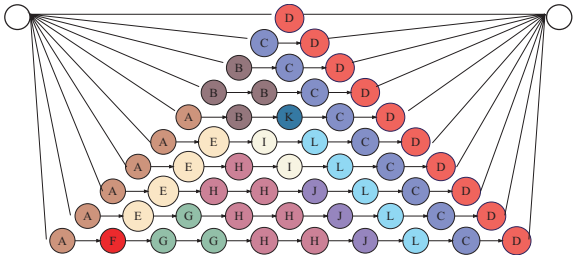


Fig. 1. An examples of phoneme model by FCD (/m-i+r/)

Because of this characteristic, the FCD model can generate a smooth trajectory of speech feature vector. Fig. 2 shows that the trajectory generated by the FCD model. The solid line is the first coefficient of the MFCC of the actual utterance, and the dashed lines are the standard deviation obtained from the FCD model. The vertical dashed lines denote the phoneme boundaries. The trajectory of the fifth section /m-i+r/ was generated from the state sequence "A-E-H-H-J-L-C-D" by the model as shown in Fig. 1.

We also performed an experiment to evaluate the performance of speech recognition for the model. In the experiment, we used 132 questions for frame position, 66 questions for phoneme duration and 134 questions for phoneme context were used to train the sub-phoneme FCD model. In order to compare the performance, left-to-right HMM and two types of HSMM models with almost same number of states were also trained. As a result, 18% of relative improvement of tri-phone accuracy was achieved by the FCD model.

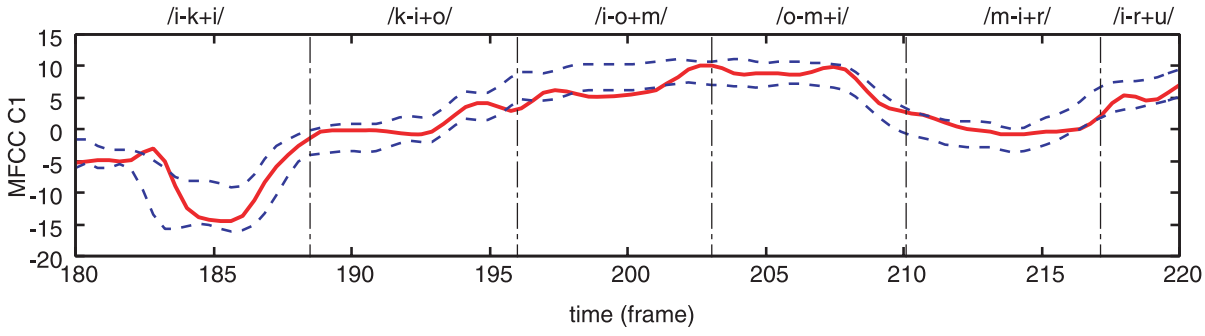


Fig. 2. An trajectory of FCD model

論　文

フレーム単位のコンテキスト依存構造に基づく音声認識のための音響モデル

非会員　寺嶌　立太[*,**]　非会員　全　炳河[*,***]
非会員　南角　吉彦[*]　非会員　徳田　恵一[*]

A Frame-Based Context-Dependent Acoustic Modeling for Speech Recognition

Ryuta Terashima[*,**], Non-member, Heiga Zen[*,***], Non-member, Yoshihiko Nankaku[*], Non-member, Keiichi Tokuda[*], Non-member

We propose a novel acoustic model for speech recognition, named FCD (Frame-based Context Dependent) model. It can obtain a probability distribution by using a top-down clustering technique to simultaneously consider the local frame position in phoneme, phoneme duration, and phoneme context. The model topology is derived from connecting left-to-right HMM models without self-loop transition for each phoneme duration. Because the FCD model can change the probability distribution into a sequence corresponding with one phoneme duration, it can has the ability to generate a smooth trajectory of speech feature vector. We also performed an experiment to evaluate the performance of speech recognition for the model. In the experiment, 132 questions for frame position, 66 questions for phoneme duration and 134 questions for phoneme context were used to train the sub-phoneme FCD model. In order to compare the performance, left-to-right HMM and two types of HSMM models with almost same number of states were also trained. As a result, 18% of relative improvement of tri-phone accuracy was achieved by the FCD model.

キーワード：音声認識，音響モデル，コンテキストクラスタリング
Keywords: speech recognition, acoustic model, context clustering

1. はじめに

現在，音声認識においては，隠れマルコフモデル (HMM) を用いた手法が広く用いられている。HMM を用いる利点の一つとしては，音声信号の時間的変化を確率的な状態遷移として表現できるため，発話のバリエーションに対して柔軟に対応できるということが挙げられる。

しかしながら，さらに認識性能を向上させるためにいくつかの課題がある。その一つとして HMM では同一の状態内で出力確率分布が一定という制約があることが挙げられる。音声は主に音素コンテキスト，話速，発話それ自身のばらつきによって変動するが，こうした現象に対応するためには，各状態の確率分布の表現能力を向上させる一方で，コンテキストおよび時間方向に対してより精緻なモデル化が必要になる。

そのための手段としてこれまでに様々な手法が提案されてきた。音素コンテキストに関するモデル化に対しては前後の音素環境別に音素モデルを構築する手法が試みられており[(1)]，クラスタリングによって多様なコンテキストに対するデータスパースネスの問題を解消する手法[(2)(3)]，FST によってより広範囲の音素コンテキストを効率的に処理する手法[(4)(5)] などが提案されている。さらに韻律を考慮するモデル[(6)] や特徴量次元を考慮する手法[(7)] も提案されている。

また時間方向のモデル化に関しては，音素継続長のモデル化に対して多くの研究がなされてきた[(8)~(13)]。その代表的なものは HMM が仮定する状態継続長が指数分布に従うという仮定を改良するものである。

別のアプローチとして，観測ベクトルのトラジェクトリを考慮する手法が挙げられる。線形回帰モデルによりトラ

* 名古屋工業大学
〒 466-8555　名古屋市昭和区御器所町
Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555
** (株) 豊田中央研究所
〒 480-1192　愛知県愛知郡長久手町
Toyota Central R&D Labs., INC.
Nagakute-cho, Aichi-gun, Aichi 480-1192
*** 東芝欧州研究所
Toshiba Research Europe Ltd
208 Cambridge Science Park, Milton Road, Cambridge, CB4 0GZ, UK

© 2010 The Institute of Electrical Engineers of Japan.

ジェクトリを表現する手法[14]，観測ベクトルの依存関係を現時刻の状態だけでなく前時刻の観測ベクトルも考慮した上で，さらに現時刻の状態の依存関係にも反映させる手法[15][16]，前後数フレームの回帰係数として導出される動的特徴量を用いて，観測ベクトルのトラジェクトリを求め，そのトラジェクトリに基づいて分散を再計算する手法[17]，さらには静的特徴量と動的特徴量の関係をモデル化することで HMM の時間方向の制約を解消する手法[18] などが試みられてきた。しかしながらこれらの方法はモデルパラメータが多くなる，あるいはデコーディング時の計算コストが大きくなるといった課題があった。

我々はフレームレベルでコンテキスト依存構造を考慮した新しい音響モデル—以下，FCD(Frame-based Context Dependent) モデル—を提案する。これは音素内のフレーム位置，音素継続長，音素コンテキスト間の依存構造を考慮することにより，トラジェクトリモデルで表現されるような観測ベクトルの時間的変化を静的にモデル化できるという特徴をもつ。さらに前述の依存構造に基づいて状態を共有する事により，発話の変動をコンパクトに表現できる。

本論文では，実際の音声信号の音素内における時間的な変化を状態遷移構造として表すことで構成される FCD モデルの有効性を示すとともに，音声認識のための状態の構成単位について行った検討結果を述べる。

2. フレーム単位のコンテキスト依存構造に基づく音響モデル

始めにフレーム位置，音素継続長，音素コンテキストを考慮する事により，どのようなモデル構造が導出されるかについて述べる。

まず継続長とフレーム位置を独立に考慮するため，ある音素のフレーム継続長別に観測ベクトルの遷移を考え，さらに各継続長のフレーム位置毎に一つの確率分布を割り当てる。この確率分布を，ある音素コンテキストを一つの単位として，継続長別にフレーム位置の順に接続すると，Fig. 1 に示すような自己ループを持たない left-to-right 型の HMM を音素継続長別に並列に接続した構造が得られる。

ここで始点から音素継続長に対応する各遷移系列の先頭状態への遷移確率は音素継続長分布に対応する値をもつものとし，その他の状態間の遷移確率は 1.0 とする。また始点と終点は観測ベクトルを出力しない null 遷移を仮定する。

さらに別の音素コンテキストに対しても，同様のトポロジをもつモデルを与える事によって，フレーム位置，音素継続長，音素コンテキスト別に異なる確率分布をもつモデルを構成する事が可能となる。このモデル構造はフレーム毎に確率分布を切り替えることで特徴量の時間的な変化をより精緻に表現することが可能である。

しかしながらこのようなトポロジを学習するためには大量の学習データが必要となる。また母音の定常部分での表現などで複数の状態を持つことはモデルとして冗長になることが予想される。

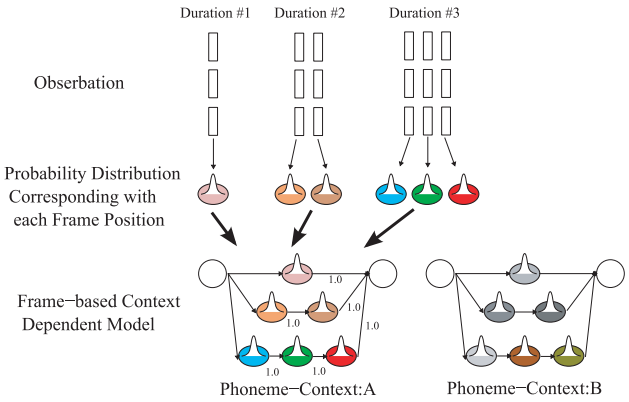


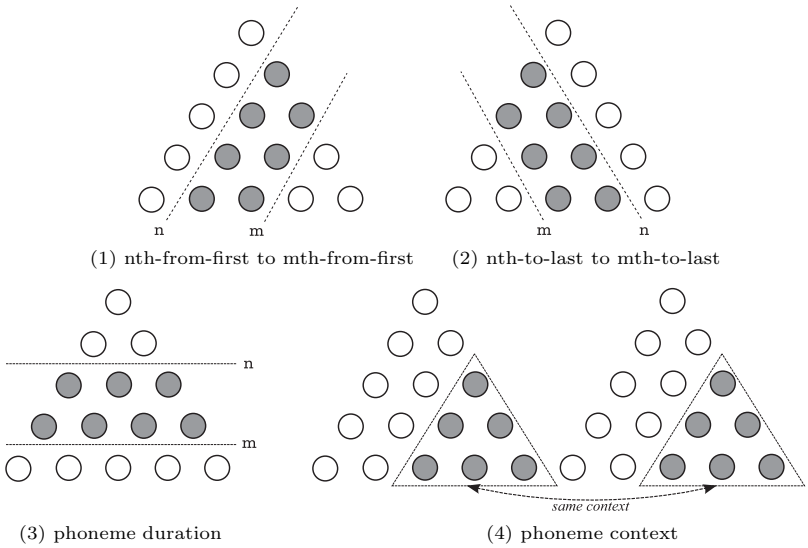
Fig. 1. Structure of Frame-based Context Dependent Model

そこで音素内，音素間で状態を共有することにより，音素内の音声信号の時間変化に対し十分な表現力を保ったまま，よりコンパクトにモデルを構成することを考える。具体的には，Odell 等が提案した ML 基準による Tree-based クラスタリング[2][3] をベースとして FCD モデルが仮定するコンテキストをより良く反映できるように改良を行った。Odell の方法は，一般的には音素コンテキストに依存した状態の共有構造を構築するために用いられる。今回は，音素コンテキストのみならず，状態継続長，音素内のフレーム位置をも各状態のコンテキストとみなして，クラスタリングを行うという方法をとった。すなわち以下の 4 種類の質問に基づいて質問木の各節において最尤となる共有構造を網羅的に探索し，さらに篠田等[19] が提案した MDL 基準により分割後の記述長が減少しなくなるまで繰り返し状態の分割を行うことで FCD モデルの状態共有構造を決定する (Fig. 2)。

（1） 系列先頭 n 番目から m 番目までの状態
（2） 系列終端 n 番目から m 番目までの状態
（3） 継続長 n から m までの系列
（4） 音素コンテキスト

状態位置に関する質問 (1) と (2) は長母音の定常部分や，音素の始まりや終わりなど周辺音素環境の影響の大きい部分を共有する場合に用いられる質問である。また音素継続長に関する質問 (3) は，音素継続長が似通っている場合，特徴量系列は同じような時系列変化をすると考えられることから，近接した音素継続長系列を共有するために用いられるものである。音素コンテキストに関する質問 (4) は通常のクラスタリング同様，音素コンテキストに依存した状態共有のための質問として適用される。後述するようにこれらの質問が複数組み合わされて適用されることで，音素内における発声の変化を柔軟に共有することが可能となる。

ここで，前述の 4 種類の質問は，どの質問が状態の共有構造を決定する上で有効となるかは自明ではない。そこで各質問を適用した際の尤度変化の計算式を次式のように改良した。



(1) nth-from-first to mth-from-first　(2) nth-to-last to mth-to-last

(3) phoneme duration　(4) phoneme context

Fig. 2.　State sharing for FCD model

$$\delta\mathcal{L} = \lambda_c \left\{ \sum_{d\in D} \frac{1}{2} \ln(|\Sigma_d|) \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_d^e(t) \right\} - \frac{1}{2} \ln(|\Sigma_p|) \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_p^e(t)$$

λ_c は質問種別毎の重み係数，$\gamma_d^e(t)$ はコンテキスト d における観測ベクトル e の時刻 t における状態占有確率，Σ_d は共分散行列であり，$\gamma_d^e(t)$ と観測ベクトル e から計算される。

これにより，例えばフレーム位置，音素継続長といった音素内の共有よりも音素コンテキストのような音素間の共有をより優先させるといった処理が可能となった。具体的には λ_c の値が大きい質問は，他の質問に比べより適用されやすくするといった制御を行うことが可能となった。

3. 音素単位の FCD モデル

〈3·1〉 モデルの学習手順　本手法によって音素内の状態変化が適切にモデル化できるかを検証する実験†を行った。

最初にモデルの学習手順について述べる。

（1） 初期モデルの学習

（2） クラスタリング

（3） 状態継続長分布の計算

（4） 連結学習

(1) の初期モデルの学習はハンドラベリングが施された学習用データに対して 1 フレーム 1 状態を仮定した状態遷移系列とを対応づけることで行った。具体的には，平均が各フレームの観測ベクトル，分散が十分小さい値††で表現されたガウス分布をもつ状態の系列を，学習データに含まれる全てのデータに対応づけて初期モデルを構成した。

(2) のクラスタリングプロセスでは，まず始めに FCD モデルの最長音素継続長を決定してモデルトポロジを固定した後，2 章で示した質問を用いてクラスタリングを行った。

最長音素継続長はハンドラベリングしたデータに基づいてほとんど全ての音素の継続長を包含する 31 フレームとした。唯一/t/だけは 71 フレームのデータが存在したが，これはショートポーズ前後の無声化した発声であると考え，今回は考慮に入れなかった。また，文頭，文末の無音モデル，ショートポーズ，促音に関しては継続長のバリエーションが多いこと，そもそも無音であるため精密なモデル化はそれほど必要とされないと考えられることから，モデル化の対象からは除外して 3 状態の left-to-right モデルとして別にクラスタリングを行った。

クラスタリングのための質問のうち，フレーム位置に関する質問と，継続長に関する質問を最大音素継続長 31 フレームを仮定した全ての組み合わせに関して生成したところ，それぞれ 992 種類，496 種類の質問が得られた。音素コンテキストに関する質問は左右の音素コンテキストに関して，それぞれ 67 種類†††合計 134 種類の質問を作成した。

(3) の状態継続長分布はラベルから得られた triphone 音素別の継続長分布をガウス分布で近似して，前述した音素コンテキストに関する質問によって共有することで求めた。

そして最後に得られた各音素の継続長を FCD モデルの先頭状態への遷移確率へ共有したまま割当て，無音の left-to-right モデルと接続した後，連結学習を行った。

その他の詳細な実験条件は Table 1 に示す。

〈3·2〉 音素 FCD モデル　Fig. 3 に FCD モデルによる中心音素/i/の決定木の一部の例を示す。実線で囲まれたノードは質問に対応し，ノード中の表記 “Hn_m”，“Tn_m”，“Ln_m” は，それぞれ 2 章で示した先頭からの状態位置，後端からの状態位置，音素継続長の各質問種別に対応する。

† 実験には HTK [20] および HTS [21] を用いた。

†† 本実験では全学習データの各ベクトルの分散値を 10^{-6} 倍した値を用いた

††† 例えば，音素そのものの質問や，口舌母音 (/o/, /u/, /a/) といった発声形態などを分ける質問などを用いた。

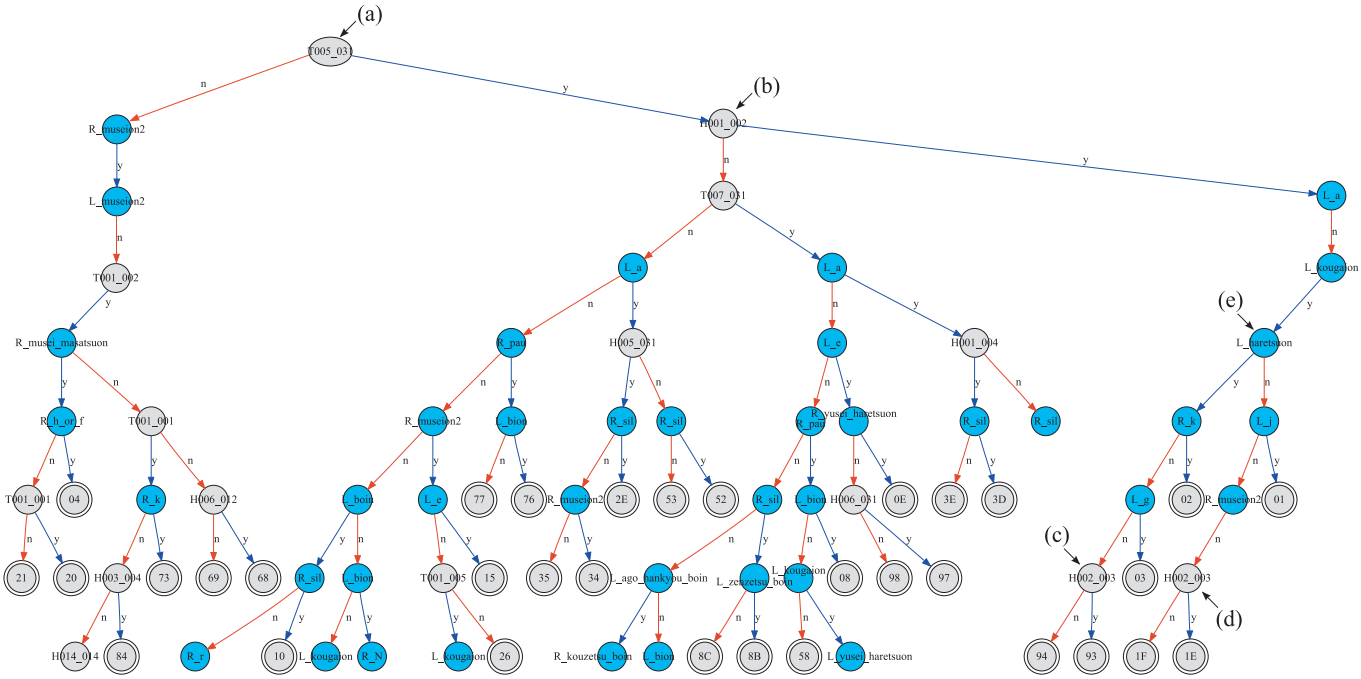


Fig. 3. Decision tree of FCD model (/i/)

Table 1. Experimental condition

特徴量	MFCC 12 次元+パワおよびその動的特徴量
サンプリング周波数	16kHz
分析窓長	25ms
フレームシフト	10ms
コーパス	ATR 音素バランス文 話者 mht
データ数	学習用:450 文 (a〜i セット), 評価用:53 文 (j セット)
音素体系	IPA 37 音素 (長母音は除く)

“L_”または “R_”の後ろにアルファベットが表記されているものは/i/を中心とした左音素コンテキスト，右音素コンテキストに対応する。

二重線で囲まれたノードは末端ノードを表し，ノード中に表記されている記号が状態の異なりを表す ID†に対応する。

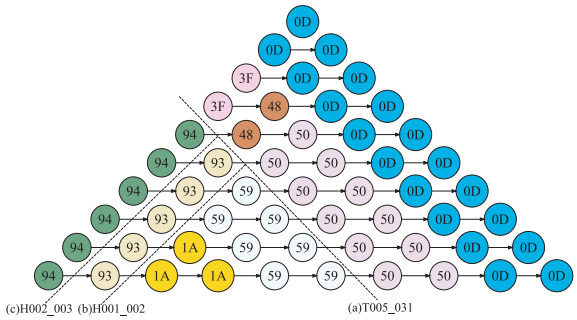
今回の実験では中心音素/i/を表現する状態数は 180 であった。最大継続長が 31 の FCD モデルを表現するには 496 状態必要であることから 36%の状態を共有して表現可能であることがわかる。また音素全体では 2275 状態であり，これは全状態数の 13%に相当する。なお，クラスタリングは中心音素毎に行った。

次に Fig. 3 で得られた決定木から決定された状態共有の例を示す††(Fig. 4)。ノード中に示されている記号は状態 ID を示しており，Fig. 3 の末端ノードで示されている ID に対応する。

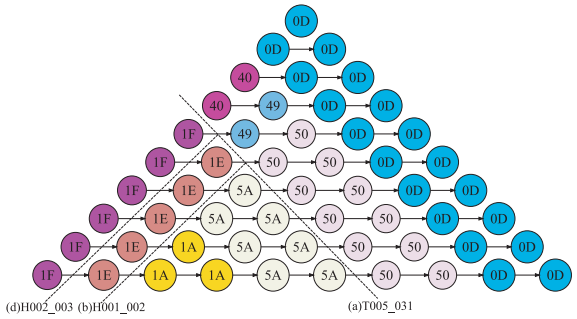
例えば，/k-i+o/(Fig. 4) (a) の状態 ID94 および 93 は，Fig. 3 のルートノード (a) 音素内位置が後ろから 5 番目から 31 番目まで (**T005_031**)，(b) 音素内位置が先頭から

† 状態別に異なるように対応づけた数値を 16 進表示したものを用いた

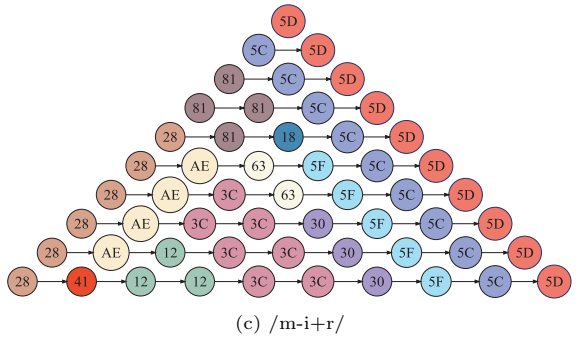
†† 実際は，最大音素長 31 の FCD モデルであるが，紙面の都合上，音素長 10 までの共有状態を図示した。



(a) /k-i+o/



(b) /ch-i+o/



(c) /m-i+r/

Fig. 4. Examples of sharing state

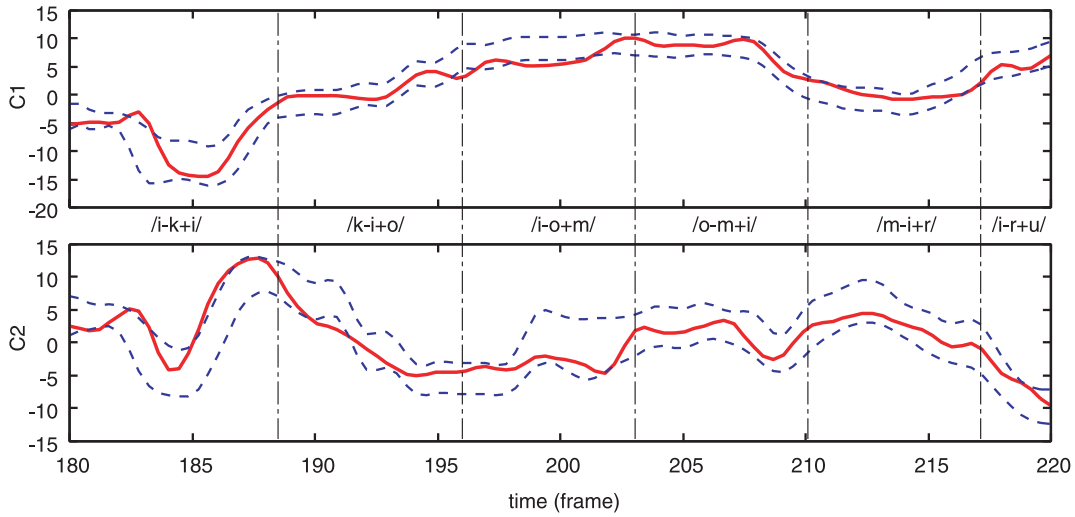


Fig. 5. An example of a trajectory generated by FCD model

1 番目から 2 番目 (**H001_002**) を経て，左音素が/a/でない (**L_a**)，左音素が口蓋音 (**L_kougaion**)，(e) 左音素が破裂音 (**L_haretsuon**)，右音素が/k/でない (**R_k**)，左音素が/g/でない (**L_g**) をたどり，(c) 音素内位置が先頭から 2 番目から 3 番目か否か (**H002_003**) によって決定されている。

同様に/k-i+o/とほぼ同じ共有構造を得られた/ch-i+o/ (Fig. 4 (b)) において，/k-i+o/の状態 94, 93 と同位置にある状態 1F および 1E の違いは Fig. 3 の (e) 左音素が破裂音 (**L_haretsuon**) で分岐して，(d) 音素内位置が先頭から 2 番目から 3 番目か否か (**H002_003**) によって決定された。

一方，Fig. 5 に FCD モデルによって得られる確率分布のトラジェクトリの例を示す†。上のグラフが MFCC の第一次の係数，下のグラフが二次の係数を表し，実線が実発声，点線が FCD モデルから得られる分布平均からの標準偏差の幅を表している。

この結果から同じ中心音素をもつ/k-i+o/と/m-i+r/は，同一の音素継続長であるにもかかわらず，それぞれの音素コンテキストによって表現されるトラジェクトリが異なることがわかる。これは Fig. 4 (a) と Fig. 4 (c) に示した状態共有構造の違いからもこのような結果が得られることが確認できた。

このように FCD モデルによって得られるトラジェクトリは実際の発声をなめらかに近似できることがわかった。

〈3·3〉 音素 FCD モデルによる音声認識性能 さらに，FCD モデルと left-to-right 型の HMM との音声認識性能の比較を行った。Table 1 で示した通り，450 文の学習データに対し，それとは異なる 53 文のテストデータを用意しその認識性能を評価した。

比較に用いた left-to-right モデルは中間状態数 3 の 1 状態 1 混合ガウス分布型のモデルを Table 1 の条件で連結学習したモデルを用いた。また音素コンテキストは triphone を仮定し FCD モデルのクラスタリング時に用いた質問セットのうち，音素コンテキストに関する質問だけを用いてコンテキストクラスタリングを行った。一般にモデル毎にその表現能力を最適にする状態数が存在すると考えられるが，今回はモデル毎の認識性能の差が，モデルの表現能力にあるのか状態数の違いにあるのかを切り分けるために，状態共有後の異なり状態数が比較モデル間でほぼ等しくなるようにクラスタリングパラメータを調整し，状態数をそろえて実験を行った。具体的には最大の認識性能が得られた HMM の状態数と等しくなるように FCD モデルに適用する MDL の α パラメータを調整し，両モデルの状態数が等しくなるような条件下において実験を行った。

そして認識性能を音素挿入誤りと削除誤りの数がほぼ等しくなるよう挿入ペナルティを調整して triphone レベルの音素認識精度で評価した。その結果，left-to-right 型の HMM が 88.14%であったのに対し FCD モデルの音素認識精度は 84.87%であり，left-to-right モデルに対して低くなるという結果であった。

次にフレーム平均尤度によってテストデータに対するモデルの適合精度を評価した。フレーム平均尤度は評価データの正解トランスクリプションに対する forward-backward 確率のフレーム平均値を算出した。

Fig. 6 の横軸は FCD モデルのフレーム平均尤度，縦軸は left-to-right モデルのフレーム平均尤度を示す。各点はテストデータとして用いた 1 文発声データに対応する。その結果，53 文中 49 文で FCD モデルの平均尤度が高い，すなわちモデルとしての平均適合精度が高いことがわかった。

〈3·4〉 サブ音素 FCD モデル 前節のフレーム平均尤度の実験でモデルとしての適合能力は left-to-right モデルに比べて FCD モデルが高いのにも関わらず，逆に音声認識性能は FCD モデルが低いという結果が得られた。

FCD モデルによって生成されたトラジェクトリは，テストデータをよく近似するトラジェクトリを多く含んでいる

† ATR 503 文 j53「周りの美しい雪景色を見る (/k i o m i r/) 余裕も出た」の一部を示した

ためフレーム平均尤度は高くなる一方，音素単位でモデルを構築したことから，一継続時間長内の状態遷移が固定されているために音声の時系列変化に対して十分に対応できなくなる場合があるためと考えられる。それに対し，left-to-right モデルはサブ音素単位で状態を構成したことから，一状態あたりの適合精度は高くないが，各状態が時間方向に伸縮できるため，結果として音声認識性能が高くなったのではないかと思われる。

そこで，FCD モデルもサブ音素単位で構成し，これらを音素内で組み合わせることによってこの問題の解決を図る(Fig. 7)。すなわち音声信号のなめらかなトラジェクトリの生成は FCD モデルの高い表現能力を利用し，時間方向の伸縮に対してはサブ音素の FCD モデルの組み合わせで表現することにより音声認識性能の向上をねらうものである。

以上の考察に基づいて，サブ音素 FCD モデルによる認識性能の評価実験を行った。

サブ音素 FCD モデルの学習は，まず初めに left-to-right 型の HMM を学習した後，HMM の各状態を対応するサブ音素 FCD モデル中の全ての状態に割当て，そのモデルトポロジの下で音素別に連結学習を行った後クラスタリングを行った。クラスタリングはサブ音素別に行なったが，〈3·1〉節で述べた質問と同一ものを用い，同一音素内でのサブ音素間の状態の共有は考慮しなかった。なおサブ音素モデルの最大状態継続長は 11 とした。

さらに比較対象としてサブ音素単位で構成した HSMM (Hidden Semi-Markov Model) [22][23] と，中川等の提案した HSMM 型のモデル [24] の学習方法を参考に連結学習時に平均パラメータと遷移確率のみを更新したモデル（以下，HSMM'）の認識性能を評価した。両モデルとも初期モデルはサブ音素 FCD モデルと同一のものを用い，クラスタリング時の音素環境に関する質問も同一のものを使用した。

すなわち今回比較に用いた 4 モデルの差異をまとめると

（1） HSMM は left-to-right HMM に継続時間長分布モデルを加えたもの

（2） HSMM' は HSMM の平均パラメータと遷移確率のみを更新したもの

（3） サブ音素 FCD は HSMM に対して状態毎に音素位置，継続時間別のコンテキスト依存構造を加えたもの

となる。

さらに〈3·3〉節の実験と同様にモデルを構成する状態数が比較モデル間でほぼ等しくなるようにモデルの学習を行った。具体的には，質問木の大きさを制御することによって最大の認識性能が得られる HSMM モデルを求め，その状態数と等しくなるように他のモデルにおける MDL の α パラメータや FCD モデルの質問重み係数を調整した。

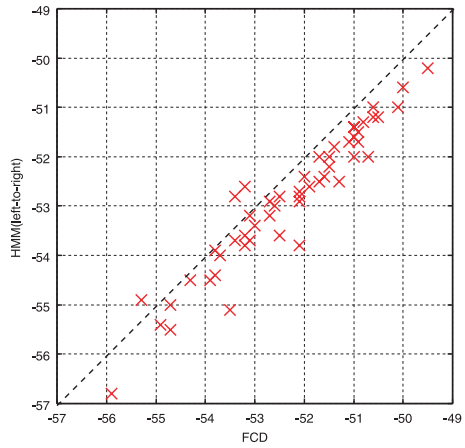


Fig. 6. Comparison of utterance log likelihood per frame between FCD and left-to-right HMM

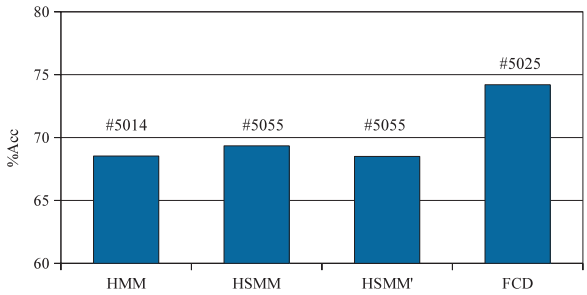


Fig. 8. Comparison of tri-phone accuracy with speaker-independent model

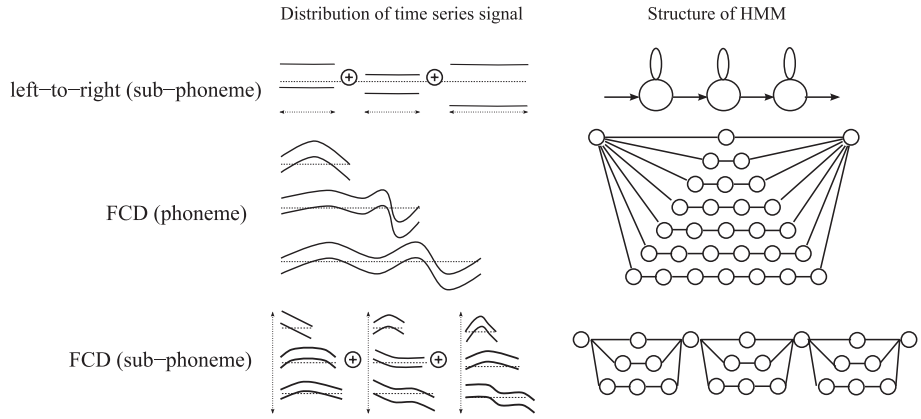


Fig. 7. Comaparison between sub-phoneme left-to-right model, phoneme FCD model and sub-phoneme FCD model

また評価実験はATR日本語音声データベースB set中に含まれる全ての話者を用いた不特定話者モデルにより行った。認識性能は〈3·3〉節の実験と同様，53文の評価データに対するopen-testにおいて，音素挿入誤りと削除誤りの数がほぼ等しくなるよう挿入ペナルティを調整したtriphoneレベルでの音素認識精度で評価した。

Fig. 8にHMM，HSMM，HSMM'，FCDモデルのtriphone音素認識精度の比較結果を示す。バーの上に示されている数値は各モデルの状態数を表す。これからわかるようにサブ音素FCDモデルはHMMに比べて18.02%，HSMMと比較して15.85%の誤り改善率[26]を得る事が出来た。またHSMMはHMMよりも認識精度が高かったが，この実験ではHSMM'の効果は得られなかった。

この結果からサブ音素FCDモデルが構成するトラジェクトリが音声認識性能の点でも有効に寄与することが示された。

〈3·5〉 考　　察　ここでFCDモデルと他のモデルとの違いについて考察する。

従来のleft-to-right型の音素コンテキスト依存HMMは，決定木により音素コンテキストと状態位置の質問を与えると，尤度最大化基準に基づいて一つの確率分布が決定された。一方，FCDモデルは音素コンテキスト，継続長，フレーム位置を与えると一つの確率分布が与えられるモデルだとみなす事が出来る。

またモデルトポロジに関して，前者は状態位置の順に接続されるためleft-to-right型のモデルとして与えられるが，FCDモデルは継続長，フレーム位置に対して独立に確率分布が与えられるため，結果としてFig. 1のような接続関係をもつモデルとして表現される。

一方，HSMMは，FCDモデルと同じような接続構造のモデルとして表現できる。しかしながら，HSMMは音素コンテキストと継続長を与える事で確率分布が決定されるモデルであるため，各音素継続長における時間変化はleft-to-right型のHMMと同様に固定された構造になる。

またHSMM'は分散パラメータのオーバフィッティングによる時間方向のばらつきに対する記述能力低下を抑止する効果があると考えられるが，本実験では動的特徴量を用いているため効果が現れにくかったのではないかと推測される。

したがって，これらのモデルに対するFCDモデルの優位性は，状態間のより多くの依存関係を考慮した上で，Fig. 5で示したような時間方向に対してなめらかに変化するトラジェクトリを構成することが可能である点にあると考えられる。

さらにこの仮説を検証するため〈3·4〉節の実験で用いたモデルを学習した際に得られた決定木を分析した。

Table 2はHSMMおよびFCDモデルを学習した際に得られた決定木において音素環境，継続長，音素位置の各質問がどれだけ含まれていたかを示したものである。括弧内の数値はそれぞれのモデル別の質問種別毎の構成割合を表す。この表から分かるようにFCDモデルでは質問のうちの約40%が音素環境以外のクラスタリングに割り当てられている。特にフレーム位置に関する質問は全体の37%を占めることから，音素環境による変化と同様に音素内でも時間方向の変化が数多く生じていることが見てとれる。

Table 2. A number of types of applied questions for decision tree

Context	Phoneme	Duration	Frame Position	# of questions
HSMM	4953(100%)	0(0%)	0(0%)	4953
FCD	2917(59%)	193(4%)	1813(37%)	4923

Table 3. Relationship between the number of states and the recognition performance on FCD model

# of states	5025	10877	20103	40655
%Acc	74.20	76.32	77.45	78.88

フレーム位置による分割はなめらかなトラジェクトリを構成することにつながるため，本分析からもフレームレベルのコンテキストを考慮することの有効性が示唆された。

なお〈3·4〉節の実験では，モデル毎の表現能力の差異を比較するために，FCDモデルの状態数をHSMMで最大の認識性能が得られた状態数に揃えて実験を行ったが，FCDモデルのポテンシャルを検討するために，状態数と認識性能との関係に関しても調査を行った。

その結果HSMMでは約5000状態程度で認識性能のピークがあったが，FCDモデルではさらに状態数を増やす事により認識性能が向上する事がわかった(Table 3)。

さらにモデルの表現能力という観点では，状態の分布表現を単混合ガウス分布から多混合分布へと拡張する事が考えられる。特にフレームレベルの共有構造を決定する際に多混合分布を考慮する事で，提案手法の利点を更に発揮する事ができると思われるが，現時点ではそうしたアルゴリズムの導出，実装ができていないため，今後モデルを拡張した上で従来法との比較評価を試みたい。

また，モデルの構成単位という観点では，Willettが音素または単語を1単位とする構成単位別に最大継続長分用意し，各状態系列間の状態遷移確率を一つ前の状態系列に依存して構成するESHMM(Expanded-State Hidden Markov Model)を提案している[25]。Willettは同時に構成単位内のいくつかの状態を共有することにより，よりコンパクトにESHMMを構築できることを示している。しかしながらESHMMの状態共有はヒューリスティクスに基づくもので，どのような基準で状態を共有すれば良いかが明確ではない。

我々は音素コンテキスト，状態継続長，フレーム位置を各状態のコンテキストとみなしてML基準によるクラスタリングを行なうことで音声認識性能が向上することを明らかにした。これは特定のヒューリスティクスに基づかない汎用的な手法であり，Willettが実験を行った数字認識タスクのように音素よりもより大きな単位においても応用が可

能であるが，この点については今後の検討課題としたい。

4. むすび

フレームレベルのコンテキスト依存構造を考慮したFrame-based Context Dependent モデルを提案した。このモデルではフレーム位置，音素継続長をコンテキストとみなす事で，自己ループを持たない left-to-right 型の HMM を音素継続長別に並列に接続した構造が得られる。通常の left-to-right 型の HMM や HSMM では状態系列内での確率分布は一定であるが，提案法は一系列内で確率分布を切り替えることで特徴量の時間的な変化をより精密に表現可能であるという特徴を持つ。さらに，フレーム位置，音素継続長，音素コンテキストで状態を共有し，それぞれの寄与に応じたクラスタリングを行うことで，よりコンパクトにモデルを表現できる方法を提案した。

モデルの適合性能と音声認識性能を評価した結果，モデルの適合性能に関しては，53 文発声中 49 文において従来法と比較してフレーム平均尤度を高くモデル化できることがわかった。一方，音声認識性能に関しては，サブ音素単位でモデルを構築することにより不特定話者の音素認識タスクにおいて約 18%の誤り改善率を得ることが出来た。

本手法の大きな特長として音声信号の時系列変化を滑らかに近似できる能力が高いことがある。この特性から，今後音声合成へ本手法を適用することで高い品質の音声合成器の実現が期待できる。今後の課題として大規模な学習データを用いた場合の性能評価や，多混合分布への対応，FST などの手法を適用することによる大語彙音声認識への応用を検討していきたい。

謝　辞

本研究の推進にあたり貴重なご助言をいただいた (株) 東芝の田村正統博士および (株) 豊田中央研究所の吉村貴克博士に謹んで感謝の意を表する。

(平成 21 年 10 月 19 日受付，平成 22 年 6 月 9 日再受付)

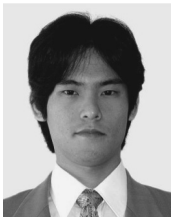
文　献

（1） K. F. Lee: “Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition”, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.38, No.4, pp.599–609 (1990)
（2） J. J. Odell: “The Use of Context in Large Vocabulary Speech Recognition”, PhD thesis, University of Cambridge (1995)
（3） S. J. Young, J. J. Odell, and P. C. Woodland: “Tree-based state tying for high accuracy acoustic modeling”, In Proc. ICASSP 94, pp.307–311 (1994)
（4） S. F. Chen: “Compiling large-context phonetic decision trees into finite-state transducers”, In Proc. Eurospeech 2003, pp.1169–1172 (2003)
（5） S. Sakti, S. Nakamura, and K. Markov: “Incorporating a bayesian wide phonetic context model for acoustic rescoring”, In Proc. INTERSPEECH 2005, pp.1629–1632 (2005)
（6） K. Chen, S. Borys, M. Hasegawa-johnson, and J. Cole: “Prosody dependent speech recognition with explicit duration modelling at intonational phrase boundaries”, In Proc. EUROSPEECH 2003, pp.232–245 (2003)
（7） H. Zen, K. Tokuda, and T. Kitamura: “Decision tree based simultaneous clustering of phonetic contexts, dimensions, and state positions for acoustic modeling”, *IEICE Trans. Inf. & Syst.*, Vol. J87-D-II, No.8, pp.1593–1602 (2004) (in Japanese)
全　炳河・徳田恵一・北村　正：「決定木に基づく音素コンテキスト・次元・状態位置の同時クラスタリングによる音響モデリング」, 信学論 (D-II), Vol.J-87-DII, No.8, pp.1593–1602 (2004)
（8） J. Pylkkonen and M. Kurimo: “Duration modeling techniques for continuous speech recognition”, In Proc. INTERSPEECH 2004, pp.385–388 (2004)
（9） M. J. Russell and R. K. Moore: “Explicit modelling of state occupancy in hidden markov models for automatic speech recognition”, In Proc. ICASSP 85 (1985)
（10） Y. Zhu and T. Lee: “Explicit duration modeling for cantonese connected-digit recognition”, In Proc. INTERSPEECH 2004, pp.685–688 (2004)
（11） Y. Takada, E. Tsuboka, and H. Wakita: “A comparative study on duration models — Poison, Gamma, and Logarithmic normal distributions —”, The 1991 Autumn Meeting of ASJ, 2-5-14, pp.75–76 (1991) (in Japanese)
高田義広・坪田英一・脇田　壽：「継続時間分布モデルの比較検討ーポアソン分布，ガンマ分布，対数正規分布」, 日本音響学会 秋季講演論文集，2-5-14, pp.75–76 (1991) (in Japanese)
（12） Y. Wakita and E. Tsuboka: “State duration constraint using syllable duration for speech recognition”, In Proc. ICSLP, pp.195–198 (1994)
（13） Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura: “State duration modeling based on Gamma distribution for HMM-based speech synthesis”, The 2001 Autumn Meeting of ASJ, 3-2-4, pp.311–312 (2001) (in Japanese)
石松喜伸・徳田恵一・益子貴史・小林隆夫・北村　正：「HMM 音声合成のためのガンマ分布を用いた状態継続長のモデル化」, 日本音響学会講演論文集，3-2-4, pp.311–312 (2001) (in Japanese)
（14） H. Gish and K. Ng: “Parametric trajectory models for speech recognition”, In Proc. ICSLP, Vol.1, pp.466–469 (1996)
（15） T. Kobayashi, J. Furuyama, and K. Masumitsu: “Partly hidden markov model and its application to speech recognition”, In Proc. ICASSP, Vol.1, pp.121–124 (1999)
（16） J. A. Bilmes: “Buried markov models: a graphical modeling approach for automatic speech recognition”, Computer Speech & Language, Vol.17, No. 2-3, pp.213–231 (2003)
（17） Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri: “Recognition method with parametric trajectory generated from mixture distribution HMMs”, In Proc. ICASSP, Vol.1, pp.124–127 (2003)
（18） H. Zen, K. Tokuda, and T. Kitamura: “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences”, *Computer Speech & Language*, Vol.21, No.1, pp.153–173 (2007)
（19） K. Shinoda and T. Watanabe: “MDL-based context-dependent subword modeling for speech recognition”, *J. Acoust. Soc. Jpn. (E)*, Vol.21, No.2, pp.79–86 (2000)
（20） Cambridge University Engineering Department *HTK Speech Recogntion Toolkit* , http://htk.eng.cam.ac.uk/
（21） HTS working group *HMM-based Speech Synthesis System (HTS)* , http://hts.sp.nitech.ac.jp/
（22） S. E. Levinson: “Continuously variable duration hidden markov models for automatic speech recognition”, *Comput. Speech Lang.*, Vol.1, No.1, pp.29–45 (1986)
（23） J. Ferguson: “Variable duration models for speech”, In Proc. Symposium on the Application of Hidden Markov Models to Text and Speech, pp.143–179 (1980)
（24） S. Nakagawa, Y. Hirata, and Y. Hashimoto: “Japanese phoneme recognition using continuous parameter hidden markov models”, J. Acoust. Soc. Jpn. (J), Vol.46, No.6, pp.486–496 (1990) (in Japanese)
中川聖一・平田好充・橋本泰秀：「連続出力分布型 HMM による日本語音韻認識」, 日本音響学会論文誌，Vol.46, No.6, pp.486–496 (1990)
（25） D. Willett: “Context-dependent duration modeling”, In *Proc. ICASSP*, Vol.1, pp.421–424 (2005)
（26） IPSJ SIG-SLP 雑音下音声認識評価 WG：「CENSREC-2 ベースラインスクリプトの使用方法について」, CENSREC-2 添付資料, http://sp.shinshu-u.ac.jp/CENSREC/ja/CENSREC/CENSREC-2/

寺 嶌 立 太（非会員） 1992 年愛知教育大学教育学部卒。同年 (株) 豊田中央研究所入社。2002 年名古屋工業大学電気情報工学専攻博士課程入学。2010 年満期退学。音声認識，音声対話システム，自動車の予防安全システムに関する研究開発に従事。情報処理学会員。



南 角 吉 彦（非会員） 1999 年名古屋工業大学工学部知能情報システム学科卒業。2004 年名古屋工業大学大学院博士課程修了。2005 年名古屋工業大学 大学院工学研究科 助手。工学博士。音声言語情報処理，マルチモーダル情報処理，統計的学習理論の研究に従事。電子情報通信学会，日本音響学会各会員。



全 炳 河（非会員） 2001 年名古屋工業大学工学部知能情報システム学科卒業。2006 年名古屋工業大学大学院博士課程修了。2008 年東芝欧州研究所・ケンブリッジ研究所 研究技術員。工学博士。統計的音声認識・合成の研究に従事。日本音響学会，情報処理学会，ISCA 会員。



徳 田 恵 一（非会員） 1989 年東京工業大学大学院博士課程修了。同年東京工業大学電気電子工学科助手。1996 年名古屋工業大学知能情報システム学科助教授。2004 年名古屋工業大学大学院情報工学専攻教授。工学博士。音声言語情報処理，マルチモーダル情報処理，統計的学習理論の研究に従事。電子情報通信学会，日本音響学会，人工知能学会，情報処理学会，IEEE, ISCA 各会員。

