口唇動作と音声のタイミング構造に基づく話者検出

堀井　悠† 川嶋　宏彰† 松山　隆司†

† 京都大学大学院情報学研究科　〒 606-8501 京都市左京区吉田本町
E-mail: †horii@vision.kuee.kyoto-u.ac.jp, ††{kawashima,tm}@i.kyoto-u.ac.jp

あらまし　人の発話における口唇動作と音声変化の間には，ずれを伴う複雑な時間的構造が存在するため，これらの共起性をフレーム単位のモデルで表現することはしばしば困難である．そこで本研究では，話者を撮影して得た口唇動作と音声の特徴量系列をそれぞれ時間的に分節化して時区間系列の対とし，これら時区間の時間関係に基づいてメディア間の系統的時間差（タイミング構造）をモデル化することで，複数人物の発話シーンに対して話者検出（どの人物が発話しているかの判別）を行う手法を提案する．本手法を用いて，近接した複数人物のうちいずれが発話しているかを，単一のカメラとマイクのみでも高精度に検出できることを実験により確認した．

キーワード　話者検出，タイミング構造，視聴覚統合，hybrid dynamical system

Speaker Detection Using the Timing Structure between Lip Motion and Speech Signal

Yu HORII†, Hiroaki KAWASHIMA†, and Takashi MATSUYAMA†

† Graduate School of Informatics, Kyoto University　Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan
E-mail: †horii@vision.kuee.kyoto-u.ac.jp, ††{kawashima,tm}@i.kyoto-u.ac.jp

Abstract　In this paper, we propose a novel approach to speaker detection using the cue of *timing structure* between audio and visual information. We first extract a pair of feature sequences of lip motion and sound, and segment each sequence into temporal intervals. Then, we construct a cross-media timing-structure model of human speech by learning the temporal relations of overlapping intervals. Based on the learned model, we realize speaker detection by evaluating the timing structure between the observed video and audio. Our experimental result shows the effectiveness of using temporal relations of intervals for speaker detection.

Key words　speaker detection, timing structure, auditory-visual integration, hybrid dynamical system

1. はじめに

我々人間は，話者の顔位置や口唇動作などの視覚情報と，音源方向や音声変化などの聴覚情報を，その共起性に基づいて相互補完的に統合することで，他者の発話状態（いつ誰が話したか）の認識を行っている．機械による人の発話状態認識においても，視聴覚メディアの統合による実現が期待されており，例えば，会議や講義といったシーンに対して，その自動撮影のために話者位置を検出する方法が広く研究されている [1]〜[3].

これらの手法の多くは，背景差分を用いた人物位置推定とマイクロホンアレイによる音源定位の結果を，Coupled HMM（Hidden Marcov Model）や DBN（Dynamic Baysian Network）を用いて統合することにより，話者検出を実現している [4], [5]. しかし，音源定位の空間分解能以上に人物が近接している場合や，検出対象空間でのカメラおよびマイクロホン配置に制約がある場合など，検出困難となる状況もしばしば存在する．

このような状況でも高精度に発話状態を認識するためには，人や音源の位置情報だけではなく，人の発話における口唇動作と音声変化の間の共起性を利用する方法が考えられる．発話認識で用いられる従来のメディア統合手法は，サンプリングしたデータや特徴抽出時のフレームを単位として，同一フレームや隣接フレームでの共起性や特徴量相関をモデル化している [6]. しかし，/a/や/o/などの母音の発声において口唇動作の開始が音声よりも先行するように，これらは必ずしも完全に同期するものではなく，従来のフレームベースの統合手法では，このようなずれを伴う共起性を十分に表現できない [7]. 実際に，人の知覚においても，許容される視覚情報と聴覚情報の時間的ずれには，ある程度の広がりがあることが知られている [8], [9].

本研究ではこの点に着目し，口唇動作と音声変化の間の系統的時間差を伴う時間的構造を**タイミング構造**と呼び，これをよ

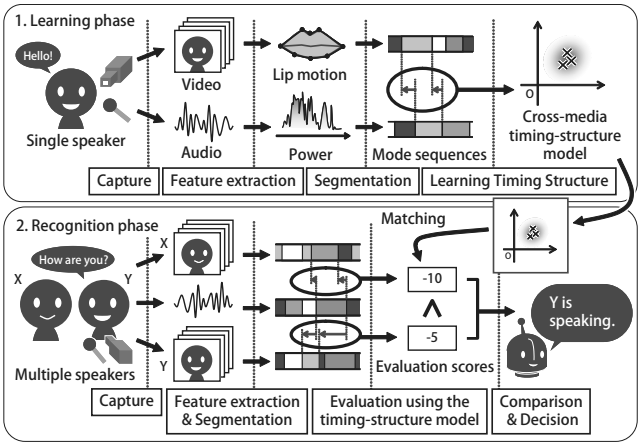

図 1　全体の処理の流れ

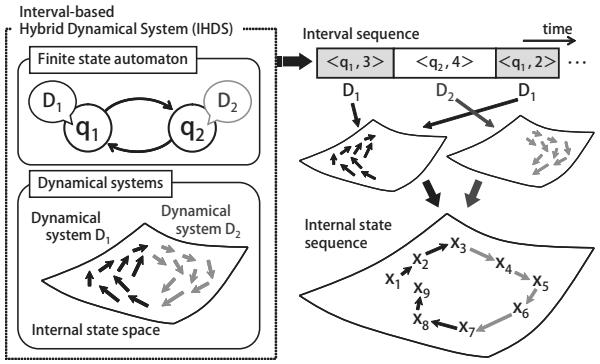

図 2　時区間ハイブリッドダイナミカルシステム（IHDS）

り直接的にモデル化し，話者検出に利用する手法を提案する．具体的には，映像中の複数人物のうち，いずれかの人物が発話している区間について，発話者の正確な判別を目的とする．

メディア信号間のタイミング構造をモデル化することで，音声にあった口唇映像を生成する先行研究として文献 [10] がある．これに対して，本研究ではこのようなタイミング構造を．メディア信号の変換・生成ではなく，信号間の時間的構造の整合性評価に用いる．すなわち，発話というイベントを認識する上で，タイミング構造がどの程度の有効性を有するかを検討する．

本手法は，人物同士が非常に近接している場合や，カメラとマイクが各 1 台のみの状況でも実現可能であり，遠隔会議システムでの話者追跡による自動撮影だけでなく，アーカイブされた映像コンテンツの分析などへの応用が期待できる．

2. 状況設定とアプローチ

本稿では，話者検出にタイミング構造を利用することの有効性の評価に焦点をあてるため，以下のように単純化した状況を設定する．

- 話者検出の対象とするデータは，カメラ方向を向いている複数人物の発話シーンを，カメラとマイク各 1 台ずつを用いて同時キャプチャした時系列画像および音声信号である．
- 画像には唇の遮蔽がない正面顔が映っている（video）．
- 話者以外の人物の口唇動作も同時に起こる（video）．
- 音声に関して一定レベルのノイズは許容する（audio）．
- 同時発話は起こらない（audio）．

この状況設定における話者検出のアプローチを以下に示す（図 1 参照）．処理は学習フェーズと認識フェーズの 2 段階からなる．

まず，学習フェーズでは，人物の発話シーンから口唇動作と音声の特徴量系列を抽出し，各メディア信号のモデル学習と特徴量系列の分節化を行う．そして，得られた区間系列対を学習データとして，口唇動作と音声の間に存在する時間的構造を学習することによって，発話のタイミング構造モデルを獲得する．これは，口唇動作と音声強度の変化パターンに関して，両者の時間的ずれの許容範囲を確率分布として表現したものである．

認識フェーズでは，新たな観測データに対して，シーンにいる人物全員の口唇動作と音声の特徴量系列を抽出し，分節化を行う．次に，あらかじめ学習フェーズで獲得したタイミング構造モデルを用いて，各人物の口唇動作と音声の間のタイミング構造を評価し，評価値の最も高い人物を話者として検出する．

なお，本研究では不特定人物を対象とした話者検出を目指すため，個人差の大きな口唇映像や形状パラメタ，音素変化の特徴量ではなく，口唇の特徴点の垂直方向の変化や音声パワーといった，比較的個人差の少ない特徴量を利用する．

続く第 3 節では特徴量系列を分節化するための単一メディア信号のモデル化法について述べ，第 4 節ではメディア間のタイミング構造をモデルとして表現する方法を述べる．第 5 節では学習したモデルを用いた新たな観測データの評価方法について述べる．第 6 節では実験により提案手法の有効性を検証する．

3. 単一メディア信号の時区間表現

本研究では，各フレームにおいて口唇動作および音声の特徴量を抽出し，それぞれのベクトル系列を時系列データとして用いる．これらの時系列データは，繰り返し出現する要素的な変化の組み合わせによって表現することができると仮定する．本稿では，時系列データを複数の線形システムで表現される時区間に分割して扱うためのモデルとして，文献 [10], [11] で提案されている時区間ハイブリッドダイナミカルシステム（IHDS, Interval-based Hybrid Dynamical System）を用いる．

3.1 時区間ハイブリッドダイナミカルシステム

IHDS は，信号の要素的な変化を表現する力学系（ここでは線形システム）と，線形システム間の遷移を表す有限状態確率オートマトンから構成される．これに類する構造を持つモデルとして，SLDS（Switching Linear Dynamical System）[12] が挙げられる．しかし，SLDS は線形システム内の状態変化と離散的状態遷移を共に物理的時間でモデル化しており，一つの線形システムが持続する時間が短いほど尤度が高くなるという点で，分節化結果が観測ノイズの影響を受けやすいと考えられる．

一方，離散状態の持続時間を確率分布として直接モデル化する枠組みとして Segment Model [7] があり，本稿で用いる IHDS はこの一形態であるといえる（Segment 内の信号変化のモデルに力学系を利用する）．一般に Segment Model などのように多くのパラメタを有する場合，与えられた観測信号からモデルを推定することは困難であるが，IHDS では，要素的な

信号変化を表現する線形システムを観測信号からボトムアップに推定する手法を提供しており，さらに，新たな観測信号の分節化を精度よく行うことが可能である．

以上の理由から，本稿では観測信号を分節化する具体的モデルとして IHDS を用いるが，第 4 節以降で述べるメディア信号間のタイミング構造のモデル化には，上述の SLDS や Segment Model を含む，他のモデルおよび分節化手法も利用できる．

3.1.1 システムアーキテクチャ

IHDS は 2 層構造を持ち，第 1 層は複数の離散状態間の確率的遷移をモデル化する有限状態確率オートマトンであり，第 2 層は各離散状態に付随する複数の線形システム $\mathcal{D} = \{D_1, \ldots, D_N\}$ を持つ内部状態空間（IHDS に属す全線形システムが共有する n 次元ベクトル空間）である（図 2 参照）．さらに，これら 2 層の統合のために時区間の概念を導入し，各時区間に，属性としてオートマトンの離散状態 q_i とその持続時間 τ を持たせる．このとき，離散状態 q_i と線形システム D_i を対応づけることで，オートマトンによる内部状態のダイナミクスを制御できる．以下では，離散状態 q_i, 持続時間 τ の時区間を $\langle q_i, \tau \rangle$ と表す．

3.1.2 線形システム

線形システム D_i の内部状態遷移を，

$$\boldsymbol{x}_t \;=\; F^{(i)}\boldsymbol{x}_{t-1} + \boldsymbol{g}^{(i)} + \boldsymbol{\omega}_t^{(i)} \qquad (1)$$

と表す．ただし，$\boldsymbol{x}_t$ は時刻 t における内部状態を表す n 次元ベクトル，$F^{(i)}$ は遷移行列，$\boldsymbol{g}^{(i)}$ はバイアスである．また，$\boldsymbol{\omega}_t^{(i)}$ はプロセスノイズであり，ガウス分布でモデル化される．各線形システムは $F^{(i)}$, $\boldsymbol{g}^{(i)}$, $\boldsymbol{\omega}_t^{(i)}$ の分布をパラメタとして持つ．

3.1.3 区間に基づくオートマトンの離散状態遷移

IHDS でのオートマトンは，持続長を持つ離散状態系列（時区間系列）を生成する．生成される時区間系列 $\mathcal{I} = [I_1, \ldots, I_K]$ に単純マルコフ性を仮定し，さらに，区間同士にはギャップやオーバーラップがないものとする．このとき，時区間 $\langle q_j, \tau \rangle$ が時区間 $\langle q_i, \tau_p \rangle$ の後に続いて起こる確率（区間遷移確率）$P(I_k = \langle q_j, \tau \rangle | I_{k-1} = \langle q_i, \tau_p \rangle)$ のモデル化を行う．

3.2 IHDS の学習

学習データとして，特徴量ベクトル系列のみが与えられているとする．このとき，線形システムの固有値制約に基づく階層的クラスタリングによるシステム個数およびパラメタ概値の推定と，システム個数を固定しての EM（Expectation-Maximization）アルゴリズム [13] によるパラメタ調整という 2 段階の学習により，IHDS のシステム同定が行える．また，同時に学習データの特徴量ベクトル系列に対する分節化処理も行える．このアルゴリズムの詳細については，紙面の都合から文献 [11] に委ねる．

3.3 単一メディア信号の分節化

観測系列が与えられると，IHDS は，観測した信号系列を最もよく表現できる時区間系列を，Viterbi アルゴリズムに基づく尤度計算 [11] によって求める．これにより，新たに観測された口唇動作および音声に対して，学習した IHDS を用いて特徴量ベクトル系列の分節化を行い，時区間系列に変換することが可能である．

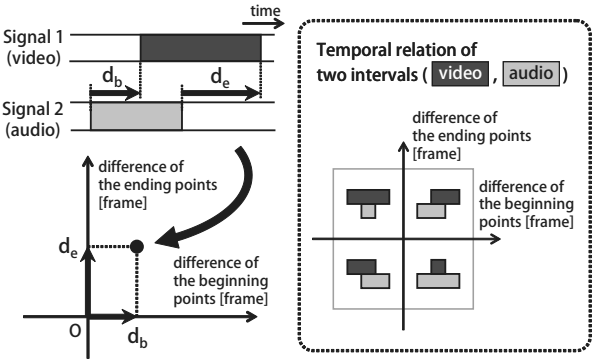

図 3　区間対の始点差・終点差によるタイミング構造の表現

4. メディア間のタイミング構造のモデル化

各メディア信号をそれぞれ別の IHDS でモデル化することで，それぞれの信号は区間系列として表現される．以下では，2 つのメディア信号 S, S' の時間的構造について扱うため，それぞれのメディア信号を「′」の有無によって区別して表記する．

4.1 メディア信号におけるモードおよびモード対時間差分布

メディア信号に含まれる要素的な変化の事象を**モード**と呼び，ここでは各モード M_i を IHDS に属す線形システム D_i と一対一に対応させる．このとき，メディア信号 S におけるモード M_i（$i = 1, \ldots, N$）の時区間と，メディア信号 S' におけるモード M'_p（$p = 1, \ldots, N'$）の時区間とが，どのような時間的関係で開始・終了するかをモデル化したものを，メディア信号 S と S' のタイミング構造モデルと定義する．

メディア信号 S, S' の区間系列を，$\mathcal{I} = [I_1, \ldots, I_K]$, $\mathcal{I}' = [I'_1, \ldots, I'_{K'}]$ とすると，区間対の総数は $K \times K'$ となるが，これらの中には時間的に非常に離れた区間対も存在する．そこで，ここでは時間的に離れたところにある区間同士の相互依存性は小さいと仮定し，オーバーラップする区間対についてのみモデル化を行う．すなわち，考慮する区間対 $(I_k, I'_{k'})$ の条件を，

$$[b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \emptyset \qquad (2)$$

とする．ただし，b_k, e_k は区間 I_k の開始時刻，終了時刻を表す．

具体的には，2 つのメディア信号 S, S' のオーバーラップする区間対 $(I_k, I'_{k'})$ において，それぞれのモードが M_i, M'_p であるときに，両区間の始点差が $b_k - b'_{k'}$，終点差が $e_k - e'_{k'}$ になる同時確率をモデルとする．すなわち，式 (2) の条件を満たす区間対 $(I_k, I'_{k'})$ に表れるモード対 (M_i, M'_p) について，その開始時刻の差 d_b および終了時刻の差 d_e の同時分布

$$P(b_k - b'_{k'} = d_\mathrm{b}, e_k - e'_{k'} = d_\mathrm{e} | m_k = M_i, \\ m'_{k'} = M'_p, [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \emptyset) \qquad (3)$$

をタイミング構造のモデルとし，これを**モード対時間差分布**と呼ぶ．これは始点差と終点差を軸とする 2 次元ユークリッド空間における分布となり（図 3 参照），例えば，分布が原点付近に高い山を持つ場合は，その 2 つのモードは開始と終了が共に同期する傾向があることを意味する．

4.2 タイミング構造モデルの学習

学習データとして，2 つのメディア S, S' のモード区間系列 $\mathcal{I}, \mathcal{I}'$ が与えられたとする．このとき，モード対 (M_i, M'_p) の時間差分布は，学習データの区間系列においてオーバーラップを含む区間対のうち，モード対 (M_i, M'_p) を持つものを見つけ，始点差と終点差を軸とする 2 次元ユークリッド空間に投票することによって学習する．ただし，学習サンプル数は有限なので，実際はガウス分布等の分布関数を各投票点に畳み込むことで全体の分布を得る．評価実験では，2 次元ガウス関数を用いた．

これを全モード対について行うことで，モード対時間差分布が学習できる．メディア信号 S, S' におけるモード集合の要素数がそれぞれ N, N' の場合，合計 $N \times N'$ 個の分布が得られる．これによって，次式に示すような区間対 $(I_k, I'_{k'})$ についての関数 $F(I_k, I'_{k'})$ を得たことになる．

$$F(I_k, I'_{k'}) = P(b_k - b'_{k'} = d_\mathrm{b}, e_k - e'_{k'} = d_\mathrm{e} | m_k = M_i, \\ m'_{k'} = M'_p, [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \emptyset) \quad (4)$$

ただし，区間対 $(I_k, I'_{k'})$ が持つモード対を (M_i, M'_p)，区間対の始点差，終点差をそれぞれ d_b, d_e とした．

5. タイミング構造評価に基づく話者検出

2 つのメディア信号 S, S' として, 区間系列 $\mathcal{I} = [I_1, \ldots, I_K]$, $\mathcal{I}' = [I'_1, \ldots, I'_{K'}]$ が観測されたとする．このとき，区間系列対 $(\mathcal{I}, \mathcal{I}')$ が持つ時間的構造が，学習によって獲得したタイミング構造モデルとどの程度一致しているかを評価する．

区間系列 $\mathcal{I}, \mathcal{I}'$ のうち，オーバーラップを含む区間対 $(I_k, I'_{k'})$ の集合を $\mathcal{P}$ とする．このとき，この集合 $\mathcal{P}$ が持つタイミング構造の評価値を，$\mathcal{P}$ に含まれる区間対がそれぞれの始点差・終点差を持つ同時確率により定める．すなわち，区間対 $(I_k, I'_{k'})$ の評価値を式 (4) の $F(I_k, I'_{k'})$ とし，$\mathcal{P}$ の評価値 $\hat{F}(\mathcal{P})$ を集合 $\mathcal{P}$ に含まれる全区間対について $F(I_k, I'_{k'})$ を掛け合わせて，

$$\hat{F}(\mathcal{P}) = \left[\prod_{(I_k, I'_{k'}) \in \mathcal{P}} F(I_k, I'_{k'}) \right]^{\frac{1}{n(\mathcal{P})}} \quad (5)$$

と定める．ただし，$\mathcal{P}$ の要素数 $n(\mathcal{P})$ による正規化を行っている．さらに，メディア信号対 (S, S') 全体の評価値 $E(S, S')$ を，

$$E(S, S') = \log \hat{F}(\mathcal{P}) = \frac{1}{n(\mathcal{P})} \sum_{(I_k, I'_{k'}) \in \mathcal{P}} \log F(I_k, I'_{k'}) \quad (6)$$

と定義する．ここで対数をとるのは，実際の計算を行う際に，計算機上でのアンダーフローを防ぐためである．

話者検出にあたって，2 者 X, Y の口唇動作を表すメディア信号として $S^{(\mathrm{v,X})}$, $S^{(\mathrm{v,Y})}$, 音声信号として $S^{(\mathrm{a})}$ が観測されたとする．このとき，$E(S^{(\mathrm{v,X})}, S^{(\mathrm{a})})$ と $E(S^{(\mathrm{v,Y})}, S^{(\mathrm{a})})$ を比較することによって，その評価値の大きい方の口唇動作が，音声との間に学習したタイミング構造モデルと近い時間的構造を持っている，すなわち，発話者のデータであると判断できる．

6. 評 価 実 験

提案手法の評価のため，実験を行った．まず，学習データとして 2 者の発話シーンをキャプチャしたデータを用い，発話のタイミング構造モデルを学習した（6.1 節）．続いて，学習データと同一の 2 者をキャプチャした新たなデータに対して，学習したモデルを用いたタイミング構造評価を行い，話者検出を行った（6.2 節）．さらに，学習データとは別の 5 者の発話シーンを用い，学習したモデルの汎化性能を検証した（6.3 節）．

図 4　発話シーンをキャプチャした画像の例（学習データ）

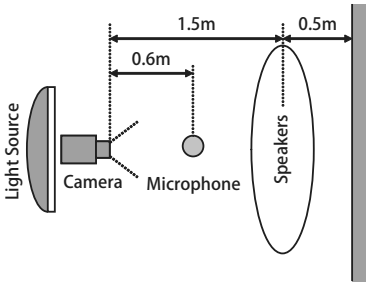

図 5　撮影環境のレイアウト

6.1 発話のタイミング構造モデルの学習

学習データとして 2 者の発話シーンをキャプチャし，以下に述べる手順で発話のタイミング構造モデルを学習した．なお，ここで学習するモデルは人物間で共通のものとした．

6.1.1 時系列画像および音声データのキャプチャ

図 4 に示すような 2 者（人物 X, Y とする）の発話シーンをキャプチャした．キャプチャは，図 5 に示すようにカメラ(注1)とマイク(注2)を各 1 台ずつ配置し，多少の残響と PC ファン音程度の音声ノイズのみが存在する室内環境で行った．撮影画像は解像度 640×480 pixels （口唇領域が約 40×20 pixels），フレームレート 60fps，8 ビット量子化，グレースケール画像とし，音声は標本化レート 48kHz，16 ビット量子化とした．

キャプチャしたデータは計 17010 フレーム（≃ 4.7 分）で，2 者は 1〜2 センテンス（約 20〜30 秒）毎に発話を交代した．

6.1.2 口唇動作と音声の特徴量抽出

本研究は，不特定人物に対する話者検出の実現を目的とするため，不特定話者に対応したモデルを用いる必要があり，そのためには個人差の小さな特徴量を用いることが望ましい．

口唇動作に関しては，個人差が表れやすい口唇形状そのものや，口形状の横方向（x 方向）の変化は特徴量として適当で

(注1)：Flea（Point Grey Research 社），固定焦点 4mm CCTV レンズ（FUJINON, YF4A-2）を使用.

(注2)：コンデンサマイク（SONY, ECM-23F5），USB Audio Capture（Roland, UA-1000）を使用.

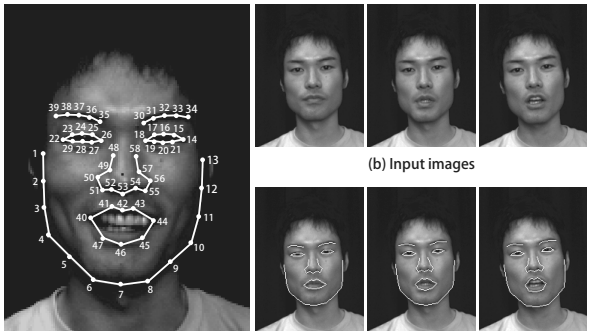

図 6　(a) AAM の学習用画像の例．特徴点は全 58 点（うち口唇領域に 8 点），(b) キャプチャした顔画像系列の例，(c) AAM による特徴点抽出結果の例.

はない．そこで，口唇動作の特徴量として，下唇に対応する特徴点の y 座標のフレーム間差分を用いた．具体的には，まず，撮影画像の一部に特徴点を手動で与え，顔の AAM（Active Appearance Model）[14] を学習し，それを用いて時系列画像全体の顔特徴点座標（全 58 点，口唇輪郭に 8 点）を抽出した（図 6 参照）．そして，特徴点座標系列に対して，全特徴点が同一平面上に存在するという仮定をおき，両目端点と鼻下の計 5 点（図 6(a), 14, 18, 22, 26, 53）を基準として特異値分解によるスケールの正規化 [15] を行ったのち，上唇（図 6(a), 40〜44）の重心の位置合わせと回転補正を行った．このとき，下唇の 5 点（図 6(a), 40, 44〜47）の y 座標のフレーム間差分を抽出した 5 次元ベクトル系列を，口唇動作の特徴量系列とした．

一方，音声の特徴量には音声信号のパワーを使用した．これは，本研究では複数人の音声を単一のメディア信号として取り扱うため，個人差に影響されないパワーベースの特徴量が望ましいと考えられるためである．音声に関しては，幅 33.3ms の窓を 16.6ms 間隔で掛け，パワーを抽出した．処理には HTK Ver.3.4 [16] を用い，1 次元特徴量系列を得た．

6.1.3　特徴量ベクトル系列の分節化

口唇動作および音声の特徴量ベクトル系列のうち，いずれかの人物が発話状態にある区間（17010 フレーム中 13533 フレーム(注3)）を手動で切り出し，各時点での発話者の口唇動作と音声信号を用いて，各メディア信号を表現する IHDS の学習および学習データの分節化を行った（3.2 節）．なお，各 IHDS のモード数は，モデル化誤差カーブを基準にして，口唇動作・音声ともに 5 とした．分節化結果の例を図 7, 図 8 に示す.

6.1.4　発話のタイミング構造モデルの学習

口唇動作と音声変化を表す時区間系列をそれぞれメディア信号 S, S' とおき，4.2 節に述べた方法で，モード対時間差分布を学習した．すなわち，オーバーラップを含む区間対の始点差と終点差を 2 次元平面にプロットし，各投票点に標準偏差 3, 共分散 0 の 2 次元ガウス分布を畳み込む(注4)ことによって，全モード対についてモード対時間差分布を得た．結果を図 9 に示す.

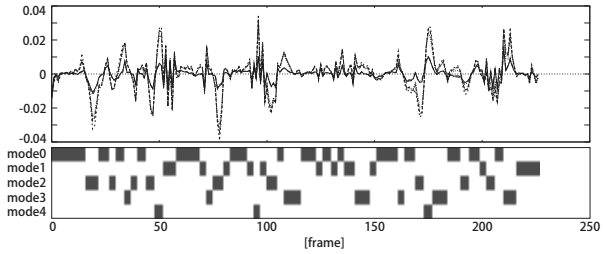

図 7　口唇動作の特徴量系列の分節化の例

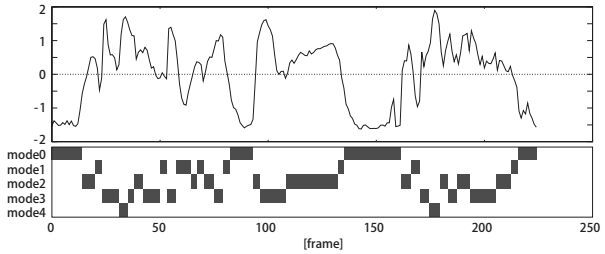

図 8　音声の特徴量系列の分節化の例

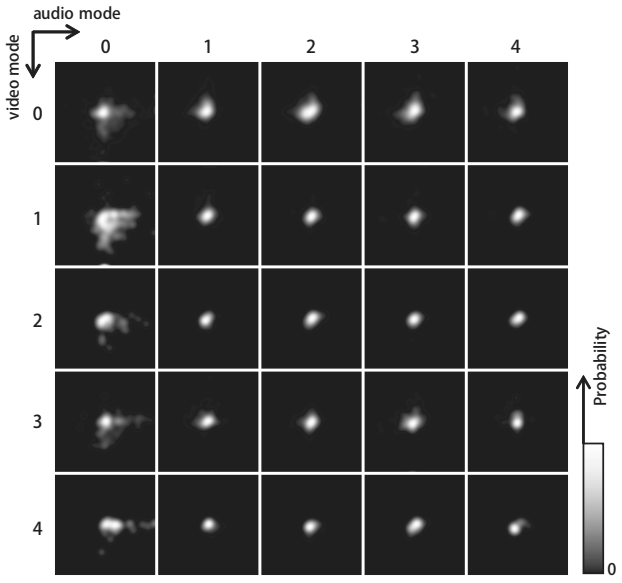

図 9　モード対時間差分布表．横軸が始点差，縦軸が終点差で，それぞれ ±50 フレームの範囲の分布を濃淡により可視化したもの.

例えば，video mode 4 と audio mode 0 の時間差分布をみると，始点差は分散が大きいが，終点は同期する傾向にあることがわかる．video mode 4 は口を開く動作に，audio mode 0 は無音区間にそれぞれ対応しているので，この学習結果は，人の発話に「口を開き終わってから，発声が起こる」という傾向があることに対応していると解釈できる.

6.2　タイミング構造評価に基づく話者検出

学習したタイミング構造モデルを用いて，新たな観測データに対する話者検出を行った．テスト用データとして，学習データと同一の 2 者 X, Y の発話シーンを新たにキャプチャし，3.3 節に述べた方法で分節化を行った時区間系列データ 12 本（そ

(注3)：人物 X の発話が 6761 フレーム，人物 Y の発話が 6772 フレーム．モデルが一方の人物に特化することを防ぐため，2 者のデータをほぼ同量とした.

(注4)：テレビ視聴の際の映像と音声について，どの程度のずれであれば不自然さを感じないかを調べた実験 [8] によると，その範囲は映像に対して音声が進む場合は約 100ms，遅れる場合は約 250ms までである．今回は $2\sigma \simeq 100$ms となるよう，標準偏差を 3 とした（フレームレート 60fps なので，$\sigma \simeq 50$ms）.

れぞれ約 2000 フレーム，計 23831 フレーム）を用いた．なお，キャプチャの環境は学習データと同じとした．

以下では，口唇動作，音声の時区間系列をそれぞれ video 信号，audio 信号と呼ぶ．さらに，音声に対する口唇動作の状態として，次の 3 つの用語を定義する．

- **発話**（Utterance）：音声に対応した口唇動作をしている．
- **沈黙**（Silence）：口唇動作をしていない．
- **口パク**（Fake lip motion）：音声と無関係な口唇動作をしている．表情変化や，呟きに伴う口唇動作などが含まれる．

6.2.1 観測信号の評価方法

video 信号 $S^{(\mathrm{v})}$ と audio 信号 $S^{(\mathrm{a})}$ のタイミング構造評価関数を $E_i(S^{(\mathrm{v})}, S^{(\mathrm{a})})$ とし，$E_i(S^{(\mathrm{v,X})}, S^{(\mathrm{a})})$ と $E_i(S^{(\mathrm{v,Y})}, S^{(\mathrm{a})})$ のうち，値が大きい方の video 信号が，発話状態に対応するデータであると判断する．以下では，$S^{(\mathrm{v})}$ と $S^{(\mathrm{a})}$ に幅 T フレームの窓をかけ，その時間範囲の評価値を求める関数として，E_1〜E_3（E_3 が提案手法）の 3 種類を定義する（図 10 参照）．

a） 同一フレームでのモード対共起確率による評価

video 信号，audio 信号の同一フレームにおけるモード対 $(m_t^{(\mathrm{v})}, m_t^{(\mathrm{a})})$ が，$(v_\mathrm{c}, a_\mathrm{c})$ となる確率 $P(m_t^{(\mathrm{v})} = v_\mathrm{c}, m_t^{(\mathrm{a})} = a_\mathrm{c})$（$t$ によらず一定）を，あらかじめ学習データから求めておく．$S^{(\mathrm{v,X})}$, $S^{(\mathrm{a})}$ の時刻 t における実際のモード対を (v_t, a_t) とするとき，評価関数 E_1 を次式により定義する．

$$E_1(S^{(\mathrm{v})}, S^{(\mathrm{a})}) = \frac{1}{T}\sum_{t=0}^{T-1} \log P\Big(m_t^{(\mathrm{v})} = v_t, m_t^{(\mathrm{a})} = a_t\Big) \quad (7)$$

b） 隣接フレーム間のモード遷移確率による評価

Coupled HMM で用いられているような隣接フレーム間のモード遷移確率を用いて評価を行う．a）において，前フレームのモード対 $(m_{t-1}^{(\mathrm{v})}, m_{t-1}^{(\mathrm{a})})$ が，$(v_\mathrm{p}, a_\mathrm{p})$ であるときのモード遷移確率 $P(m_t^{(\mathrm{v})} = v_\mathrm{c}, m_t^{(\mathrm{a})} = a_\mathrm{c} | m_{t-1}^{(\mathrm{v})} = v_\mathrm{p}, m_{t-1}^{(\mathrm{a})} = a_\mathrm{p})$（$t$ によらず一定）を学習データから得ておく(注5)．このとき，評価関数 E_2 を次式により定義する．

$$E_2(S^{(\mathrm{v})}, S^{(\mathrm{a})}) = \frac{1}{T-1}\sum_{t=1}^{T-1} \log P\Big(m_t^{(\mathrm{v})} = v_t, m_t^{(\mathrm{a})} = a_t \Big| m_{t-1}^{(\mathrm{v})} = v_{t-1}, m_{t-1}^{(\mathrm{a})} = a_{t-1}\Big) \quad (8)$$

c） モード対時間差分布による評価

本研究で提案するモード対時間差分布を用いた評価の方法として，式 (6) において，次の式 (9) に定義する集合 $\mathcal{P}$ を用いたものを評価関数 E_3 とする．すなわち，

$$\mathcal{P} = \Big\{(I_{k_\mathrm{v}}^{(\mathrm{v})}, I_{k_\mathrm{a}}^{(\mathrm{a})}) \Big| I_{k_\mathrm{v}}^{(\mathrm{v})} \in \mathcal{I}^{(\mathrm{v})}, I_{k_\mathrm{a}}^{(\mathrm{a})} \in \mathcal{I}^{(\mathrm{a})}, [b_{k_\mathrm{v}}^{(\mathrm{v})}, e_{k_\mathrm{v}}^{(\mathrm{v})}] \cap [b_{k_\mathrm{a}}^{(\mathrm{a})}, e_{k_\mathrm{a}}^{(\mathrm{a})}] \neq \emptyset\Big\}, \quad (9)$$

$$E_3(S^{(\mathrm{v})}, S^{(\mathrm{a})}) = \frac{1}{n(\mathcal{P})} \sum_{(I_{k_\mathrm{v}}^{(\mathrm{v})}, I_{k_\mathrm{a}}^{(\mathrm{a})}) \in \mathcal{P}} \log F(I_{k_\mathrm{v}}^{(\mathrm{v})}, I_{k_\mathrm{a}}^{(\mathrm{a})}) \quad (10)$$

(注5)：学習時に出現しないモード対の出現確率を 0 としないよう，補正した値を用いる必要がある（ゼロ頻度問題のディスカウンティング）．ここでは，全モード対の出現回数に一律に 0.5 を加えるという単純な加算法を用いた．なお，a）では，出現回数 0 のモード対がなかったため補正は行っていない．

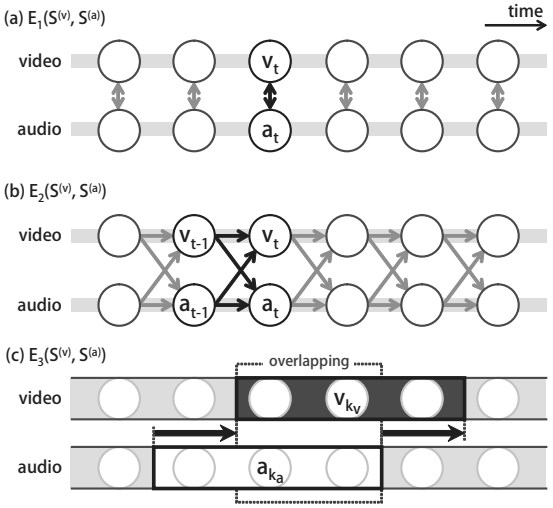

図 10　各評価関数で用いられる時間的構造．(a) 同一フレームでのモード対共起確率による評価，(b) 隣接フレーム間のモード遷移確率による評価，(c) モード対時間差分布による評価（提案手法）．

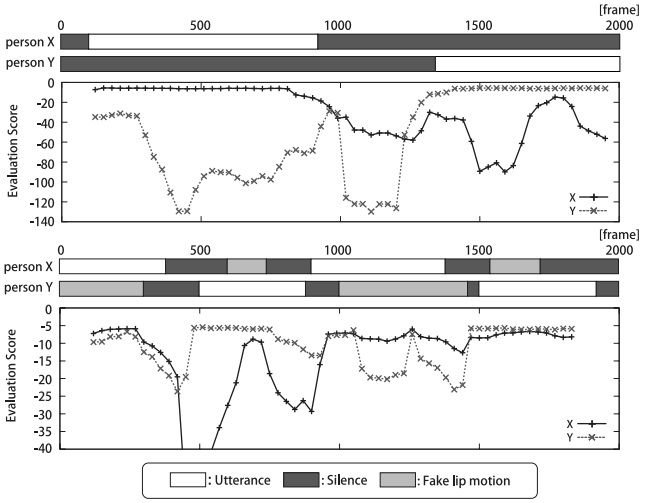

図 11　テスト用データに対する評価値の時間的変化の例．上段は各人物の口唇動作状態の変化を示したものである．窓幅 180 フレーム（3 秒），間隔 30 フレームで窓をかけ，その窓範囲内のオーバーラップする区間系列対について，評価関数 E_3 によりタイミング構造評価を行い，各窓の中央の時刻における評価値とした．上段は口パク状態が出現しないデータ（$R = 100.0\%$）．下段は口パク状態を十分に含むデータ（$R = 94.1\%$）．

とする．ただし，$F(I_{k_\mathrm{v}}^{(\mathrm{v})}, I_{k_\mathrm{a}}^{(\mathrm{a})})$ の定義は式 (4) の通りである．

6.2.2 実験結果および考察

評価値の時間変化をみるため，テスト用データに対して幅 $T = 180$ フレーム（3 秒）の窓を 30 フレーム間隔で掛け，各窓範囲のデータについて，各人物の評価値を求めた．ただし，各窓範囲の評価値を，その窓の中央の時刻における評価値と定めた．提案手法（E_3）による評価結果の例を図 11 に示す．

図 11 のグラフから，発話状態の人物の評価値が他方の人物よりも大きくなっている様子が確認できる．非発話者が沈黙状態にあるとき（図 11 上段），その評価値の差は特に大きい．

続いて，検出精度の定量的評価のため，いずれかの人物が発話している区間のうち，その発話者の評価値が他者の評価値よ

表 1　評価方法による検出正解率の比較．単位は%．括弧内はデータ点の数．E_3 が提案手法である（6.2.1 節を参照）．R_{sil} = 非発話者が沈黙状態のときの検出正解率，R_{fak} = 非発話者が口パク状態のときの検出正解率，R_{X} = 人物 X が発話者のときの検出正解率，R_{Y} = 人物 Y が発話者のときの検出正解率，R = 全体での検出正解率を表す．

使用した評価関数	R_{sil} [%]		R_{fak} [%]		R_{X} [%]		R_{Y} [%]		R [%]	
E_1	55.0	(170)	51.4	(179)	86.1	(309)	13.4	(40)	53.1	(349)
E_2	36.2	(112)	49.1	(171)	77.7	(279)	1.3	(4)	43.1	(283)
E_3（**提案手法**）	94.2	(291)	81.6	(284)	82.2	(295)	94.0	(280)	87.5	(575)

りも高くなっている区間の割合を求めた．以下では，この割合を**検出正解率**と呼び，R と表記する．また，特に非発話者が沈黙，口パク状態にある区間での検出正解率を順に $R_{\mathrm{sil}}, R_{\mathrm{fak}}$ とし，発話者が人物 X, Y である区間での検出正解率を順に $R_{\mathrm{X}}, R_{\mathrm{Y}}$ とする．テスト用データに対して，各評価関数を用いて R_{sil}, R_{fak} を求めた結果を図 12 に，$R_{\mathrm{sil}}, R_{\mathrm{fak}}, R_{\mathrm{X}}, R_{\mathrm{Y}}, R$ を求めた結果を表 1 にそれぞれ示す．

図 12 から，提案手法では非発話者が沈黙，口パクのいずれの状態であっても，他の比較手法よりも高い精度で話者検出が行えていることがわかる．非発話者が沈黙状態の場合と口パク状態の場合を比較すると，R_{sil} の方が R_{fak} より 10 ポイント以上大きい．加えて，非発話者が口パク状態での正解検出時の評価値の差が平均 2.2 であったのに対し，沈黙状態の場合には平均 34.4 となり，顕著な差が見られた．これは，沈黙状態では video 信号に長い時区間が多くなり，時間差分布の中心から離れるような時間的構造が現れやすいためだと考えられる．

さらに，表 1 で人物ごとの検出正解率を比較すると，評価関数 E_1, E_2 による評価では R_{Y} が R_{X} に対して著しく低く，その結果として全体についての R の値も 50% 程度となっており，話者検出が正しく行えていないことがわかる．人物ごとの検出正解率に大きな差が生じた原因として，これらの評価方法がモード対の出現頻度そのものをモデルとしていることが挙げられる．個人差の小さい特徴量を選ぶことで，両人物の口唇動作を共通の IHDS で記述できるが，個人ごとの口唇動作の癖の違いのため，時区間系列のモード遷移やその出現頻度には，ある程度の個人差が現れる．今回の実験では，(キャプチャ映像からの主観的な判断ではあるが）人物 X よりも人物 Y のほうが口をはっきり動かして発話する傾向にあった．その結果として人物 Y の方がモード遷移が激しく起こったため，人物 X の発話に現れやすいモードを含むモード対の出現頻度が相対的に高くなり，検出正解率に差が生じたと考えられる．

これに対して，提案手法（E_3）では，いずれの人物に対しても 80%以上が正しく検出できている．これは，モード対の出現頻度を考慮せず，時間的構造の妥当性あるいは崩れに基づいて評価値を算出することによって，このような口唇動作の個人差を吸収できたためだと考えられる．

6.3　汎化性能に関する実験

学習したタイミング構造モデルが不特定人物に対して適用できるかを評価するため，モデル学習に用いたデータとは別の 5 者（人物 1, ..., 人物 5）の発話シーンをキャプチャした（図 13 参照）．撮影環境は学習用データと同じく図 5 の通りとした．

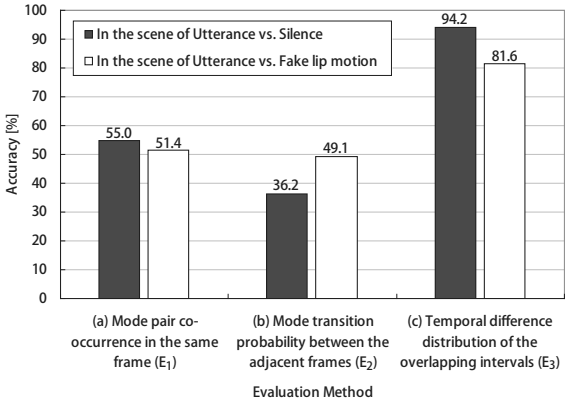

図 12　タイミング構造評価方法の比較．非発話者が沈黙状態であるときの検出正解率 R_{sil}（黒），および非発話者が口パク状態であるときの検出正解率 R_{fak}（白）を各評価方法ごとに示す．

図 13　学習データとは異なる人物をキャプチャした画像の例

撮影したデータは 8 本，計 12837 フレームで，その大部分で少なくとも 1 名が無視できない程度の口パク動作を行った．

これらのデータに対して，6.1 節で得たタイミング構造モデル（人物 X, Y から学習したもの）を用いて，タイミング構造の評価を行い，各時点で最大の評価値をもつ人物を話者として検出した．このときの各評価関数による検出正解率の結果を表 2 に示す．ただし，R_i（$i = 1, \ldots, 5$）は人物 i が発話者のときの検出正解率，R は全体の検出正解率を表す．

表 2 より，提案手法（E_3）が，他の手法（E_1, E_2）よりも高精度に話者検出を実現していることがわかる．評価関数 E_1, E_2 を用いた場合は，6.2 節と同様に人物ごとの検出正解率に大きなばらつきがあり，結果として全体の検出正解率も低くなった．

一方，提案手法では人物 2 と人物 5 の検出正解率がともに 80%程度となった．これは 6.2 節の実験の R_{fak}（口パク状態に対する検出正解率）と同程度の精度であり，他の人物のデー

表 2　学習データとは異なる人物に対する検出正解率．単位は%．括弧内はデータ点の数．E_3 が提案手法である（6.2.1 節を参照）．R_i = 人物 i が発話者のときの検出正解率を表す．

使用した評価関数	R_1 [%]		R_2 [%]		R_3 [%]		R_4 [%]		R_5 [%]		R [%]	
E_1	61.8	(21)	6.3	(4)	53.1	(34)	6.0	(6)	3.5	(2)	21.0	(67)
E_2	47.1	(16)	0.0	(0)	32.8	(21)	0.0	(0)	3.5	(2)	12.2	(39)
E_3（**提案手法**）	51.5	(17)	78.1	(50)	67.2	(43)	64.0	(64)	84.2	(48)	69.8	(222)

タで学習したモデルが適用できたと判断できる．しかし，他の人物 1, 3, 4 の検出正解率は 50〜60%前後と低い結果となった．発話者ごとの平均話速を比較したところ，人物 2 および 5 の平均話速は他の人物よりも速く，学習データ（人物 X, Y）に近い値であった．このことは，発話者の話し方によって，タイミング構造にも違いが現れる可能性を示している．ただし，個人間の検出正解率の分散は，他の手法と比較して小さくなっており，話者の個性の影響をある程度吸収できているといえる．

7. おわりに

本研究では，発話時の口唇動作と音声の変化パターンの間に存在する共起性や系統的時間差（タイミング構造）をモデル化し，このモデルに基づいて観測信号の評価を行うことで，話者検出を行う手法を提案した．本稿に示した実験は予備的な規模のものではあるが，口唇動作と音声変化の間で許容されるタイミング構造を，モード対時間差分布によって直接モデル化することによって，フレーム単位で共起性をモデル化する場合に比べて，高い精度で話者検出が可能となることを確認した．

一方で，今回提案したタイミング構造モデルは，オーバーラップする区間対のみから時間差分布を学習するという，比較的単純なものにとどまっている．非オーバーラップ区間対への拡張として，オーバーラップはしないが近傍にある区間対も含めてモデル化する方法が考えられる．そこで実際に，15 フレーム以内の時間差範囲にある区間対も含めて(注6)時間差分布を学習し，6.2 節と同様の実験を行ったが，検出精度に明確な差は見られなかった．今回の評価方法では，考慮する時間範囲を広げると評価に使う区間対の数も急激に増加するため，特徴的な情報を持つモード対が逆に埋もれてしまった可能性がある．このため，精度の向上のためには，単に時間範囲を広げるだけではなく，評価に用いるラベル対の選別やタイミング構造モデル自体の拡張をあわせて検討していく必要がある．

さらに，タイミング構造の有効性評価に焦点をあてるため，本稿では比較的単純な状況を設定したが，実際の応用を行う上では，全員が沈黙状態にある場合の考慮や，一時的な唇の遮蔽・同時発話への対応，多人数のデータに基づく不特定話者モデルの学習などが必要となる．また，6.3 に述べたように，発話者の話し方がタイミング構造に与える影響については興味深く，両者の関係を深く調べる必要がある．より大規模な定量評価とあわせ，これらを今後の課題とする．

謝辞　本研究の一部は，科学研究費補助金 No.18049046 の補助を受けて行った．

(注6)：式 (2) の区間対 $I_k, I'_{k'}$ の条件を，$[b_k - \alpha, e_k + \alpha] \cap [b'_{k'}, e'_{k'}] \neq \emptyset$ とした．α は考慮する時間範囲（$\alpha = 15$ フレーム）である．

文　献

[1] 大西正輝, 影林岳彦, 福永邦雄：“視聴覚情報の統合による会議映像の自動撮影”, 電子情報通信学会論文誌 D-II, **85**, 3, pp. 537–542 (2002).

[2] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez and D. Moore: “Audio-visual speaker tracking with importance particle filters”, IEEE International Conference on Image Processing (ICIP) (2003).

[3] 西口敏司, 東和秀, 亀田能成, 角所考, 美濃導彦：“講義自動撮影における話者位置推定のための視聴覚情報の統合”, 電気学会論文誌 C, **124**, 3, pp. 729–739 (2004).

[4] V. Pavlović, A. Garg, J. Rehg and T. Huang: “Multimodal speaker detection using error feedback dynamic Bayesian networks”, Proc. Computer Vision and Pattern Recognition, pp. 34–43 (2000).

[5] 吉村隆, 浅野太, 本村陽一, 麻生英樹, 市村直幸, 山本潔, 中村哲：“実環境における発話区間検出のための音響情報と画像情報の統合”, 電子情報通信学会技術研究報告. SP, **103**, 26, pp. 13–18 (2003).

[6] A. V. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy: “Dynamic Bayesian networks for audio-visual speech recognition”, EURASIP Journal on Applied Signal Processing, **2002**, 11, pp. 1–15 (2002).

[7] M. Ostendorf, V. Digalakis and O. A. Kimball: “From HMMs to segment models: a unified view of stochastic modeling for speech recognition”, IEEE Transactions on Acoustics, Speech and Signal Processing, **4**, 5, pp. 360–378 (1996).

[8] 鎧沢勇, 滝川啓, 大久保栄, 渡辺義郎：“衛星通信を利用した画像会議におけるエコー及び伝搬遅延の影響”, 電子通信学会論文誌, **J64-B**, 11, pp. 1281–1288 (1981).

[9] A. Peregudov, K. Glasman and A. Logunov: “Relative timing of sound and vision: evaluation and correction”, Proceedings of the Ninth International Symposium on Consumer Electronics, pp. 198–202 (2005).

[10] 川嶋宏彰, 松山隆司：“時区間ハイブリッドダイナミカルシステムを用いたマルチメディア・タイミング構造のモデル化”, 情報処理学会論文誌, **48**, 12, pp. 3680–3691 (2007).

[11] H. Kawashima and T. Matsuyama: “Multiphase learning for an interval-based hybrid dynamical system”, IEICE transactions on fundamentals of electronics, communications and computer sciences, **88**, 11, pp. 3022–3035 (2005).

[12] V. Pavlović, J. M. Rehg and J. MacCormick: “Learning switching linear models of human motion”, Proc. Neural Information Processing Systems (2000).

[13] A. P. Dempster, N. M. Laird and D. B. Rubin: “Maximum likelihood from incomplete data via the EM algorithm”, J. R. Statist. Soc. B, **39**, pp. 1–38 (1977).

[14] T. F. Cootes, G. J. Edwards and C. J. Taylor: “Active appearance model”, Proc. European Conference on Computer Vision, pp. 484–498 (1998).

[15] K. S. Arun, T. S. Huang and S. D. Blostein: “Least-squares fitting of two 3-d point sets”, IEEE Trans. on Pattern Analysis and Machine Intelligence, **9**, 5, pp. 698–700 (1987).

[16] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland: “The HTK Book (for HTK Version 3.4)”, Cambridge University Engineering Department (2006).