口唇運動-音声間のタイミング構造を利用した非定常雑音環境での発話音声推定

川嶋　宏彰†　　堀井　　悠†　　松山　隆司†

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

あらまし　雑音環境での頑健な音声認識を目的として，視聴覚情報を統合利用する様々な手法が提案されている．これらは視聴覚情報を特徴量，状態，識別器の出力といった段階のいずれかで統合するが，雑音レベルが大きくなると，聴覚情報が失われて視覚情報の識別精度に頼ることになり，聴覚情報が十分活用されないという問題がある．そこで本研究では，まず各モダリティから得られる信号を時区間系列へと分節化して口唇運動-音声間のタイミング構造を学習し，これにより，口唇運動に合った音声候補を複数個高精度に生成する．次に，これら生成候補と観測音声との整合性を音響特徴空間で評価することにより，発話された音声を推定する手法を提案する．
キーワード　マルチモーダル，非定常雑音，タイミング，線形システム，パーティクルフィルタ

Speech Estimation in Non-Stationary Noise Environments Using Timing Structures between Mouth Movements and Sound Signals

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

† Graduate School of Informatics, Kyoto University　Yoshida-Honmachi, Sakyo-ku, Kyoto, 6068501 Japan
E-mail: †{kawashima,tm}@i.kyoto-u.ac.jp, ††horii@vision.kuee.kyoto-u.ac.jp

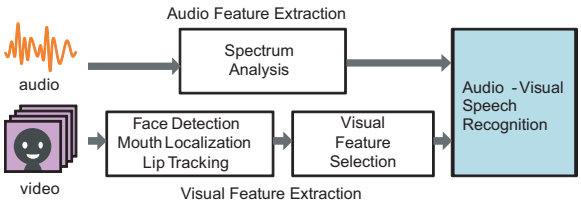
Abstract　The variety of methods of audio-visual integration, which integrates audio and visual information at the level of either features, states, or outputs of classifiers, are proposed for the purpose of robust speech recognition. However, the methods do not always fully utilize auditory information when the signal-to-noise ratio becomes too low. In this paper, we propose a novel approach to estimate speech signal in noise environments. The key idea is that it exploits clean speech candidates generated by using *timing structures* between mouth movements and sound signals. We first extract a pair of feature sequences of media signals, 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 generate clean speech candidates from the observed mouth movements.
Key words　multimodal, non-stationary noise, timing, linear system, particle filtering

1. はじめに

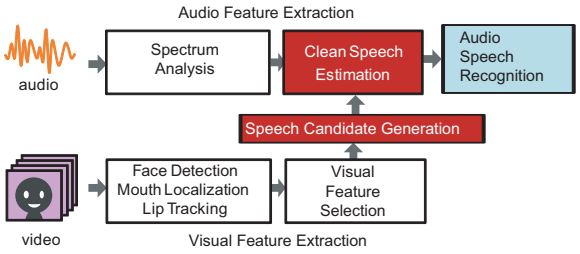
音声認識システムが用いられる状況は，電話の自動音声応答のように，雑音レベルの低い環境での単一話者の利用から，会議の議事録作成や対話分析，街角情報端末，運転環境など，複数音源もしくは複数話者が存在する場面へと広がりつつある．例えば，カーナビゲーション・システムへの音声入力を利用する運転環境では，ユーザ（運転者）による入力音声の他に，同乗者の声，カーステレオからの音，エンジン音やロードノイズなど多数の音源が存在している．

このような定常・非定常の雑音が重畳するような環境において頑健な音声認識を行う方法として，発話者の口元の動き（口唇運動）(注1)を利用することが考えられる．これらの方法は Audio-Visual Speech Recognition (AVSR) と呼ばれ，マイクとカメラにより得られる映像・音響信号から，それぞれの特徴量（顔位置や唇の動き，音響スペクトラムなど）を抽出し，同時刻や隣接時刻において音声と口唇運動の特徴量が持つ共起性をモデル化・学習する．認識時には，観測された映像・音響信号を元に，あらかじめ学習されたモデルの尤度もしくは評価値を計算することで発話音声を識別するものである．統合する段階に注目すれば，映像・音響の両メディ

(注1)：唇やその周辺，口内なども含むが，以後これらの見かけの動きを，まとめて単に口唇運動と呼ぶことにする．



(a) 従来の Audio-Visual Speech Recognition (AVSR)



(b) 視覚情報からのクリーン音声推定（口唇運動と合わせて (a) のような AVSR への入力とすることも可能）

図 1　視聴覚情報を用いた音声認識の流れ

ア信号から得られた特徴量を単純に結合するもの (early integration）や，各モダリティで独立に識別を行ったうえで結果を後段で統合するもの (late integration)，いったん Hidden Markov Model (HMM) などの状態遷移によってそれぞれの特徴量の時間変化をモデル化したうえで，状態レベルでの統合を考えるものなど様々な手法が提案されている [1], [2]（図 1(a)）．

しかしながら，これら AVSR の枠組みでは，入力される音響信号は雑音と音声とが重畳した信号であり，音声の SN 比が小さな状況においては，しばしば音声の情報が有効に利用されず，その識別精度が視覚情報に主に依存することになる．これは音声特徴と口唇運動特徴とを対等なものとして扱っていることに起因する．

そこで本研究では，視覚情報は，雑音が重畳した音響信号から音声信号（音声特徴系列）を拾い上げるための手掛かりとして利用するというアプローチをとる．具体的には，図 1(b) に示すように，話者をカメラで撮影して得られた口唇運動から，それにあった発話音声（特徴系列）の候補を複数生成し，実際に観測された音響信号との間での整合性を評価することで，雑音と音声とを分離する．推定されたクリーン音声の特徴系列は，通常の Audio Speech Recognition (ASR)，もしくは AVSR への入力に利用することが可能である．

このような視覚情報から聴覚情報を直接的に推定する手法では，口唇運動の特徴系列から音声特徴系列をいかに精度よく生成するかが重要である．特に，

（1）　一つの口唇運動には通常多数の音声が対応するため，生成候補数をいかに抑えるか

（2）　生成信号の時間的連続性や滑らかさをいかに保証するか

（3）　口唇運動と発話音声との間にしばしば生じる時間的なずれとその変動へどう対処するか

といった問題を解決する必要がある．

そこで本手法では，口唇運動から音声を推定する手法として [3] を利用する．この手法では，Hybrid Dynamical System (HDS) 呼ばれるモデルを用いて，いったん映像および音響信号から抽出した特徴系列を，それぞれ複数の線形システムの切り替わりとして表現する．そして，それら HDS において線形システムが切り替わる分節点の時間的関係を用いて，複数の信号間における変化パターン間の共起性をモデル化する（4. 節）．

再現性のある要素的な動きを線形システムで表現し，信号を記号化・分節化することで，(1) の生成候補数の爆発に対処することが可能となる．また，4. 3 節で述べるように，線形システムに基づく生成モデルでは，時系列信号を滑らかに生成することが可能であり，(2) が解決されることも期待できる．さらに，文献 [3] と同様に線形システムが切り替わる系統的時間差（タイミング構造）を別途モデル化し，時間的なずれやその変動に対応しながら信号生成を行うことで (3) の問題に対処する．

2.　問題設定と提案手法の流れ

本研究では，口唇運動と音声信号のタイミング構造に基づく音声推定の有効性を評価するため，発話者の運転状況（車内）を念頭に置いて，以下の条件を設定する．

- カメラとマイクロフォン各 1 台ずつ利用
- 口元の遮蔽のない顔映像が取得可能
- 孤立音発声に加法性非定常雑音（1 種）が重畳
- 特定話者の学習用データが取得可能

ここで，マイクロフォンアレイを併用する，残響モデル（乗法性歪み）を考慮する，多視点映像により口唇運動の追跡精度を高める，複数種類の雑音の重畳を扱う（例えば [4]），連続発話に対応する，といった様々な拡張も可能ではあるが，提案手法の基本的性能の評価に焦点を絞るために，上記のような状況設定の簡単化を行った．

さて，本研究で扱う問題は，観測映像から信頼性の高い口唇運動の特徴系列 V を抽出できることを前提として，口唇運動 V からクリーン音声の特徴系列候補 $\hat{S}^{(c)}$ $(c = 1, ..., C)$ を推定・生成し，雑音 N が重畳した入力音声（音響特徴系列）X との整合性評価により，クリーン音声の特徴系列 S を推定するというものである．この問題に対する提案手法の処理は，学習フェーズと音声候補生成フェーズ（図 2 および 4. 節と 5. 節），さらに音声候補を利用した雑音抑圧フェーズ（3. 節）の計 3 段階からなる．

まず学習フェーズでは，雑音の少ない状況での人物の発話シーンから口唇運動と音声の特徴系列を抽出し，それぞれの系列から HDS モデル [3] の学習と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

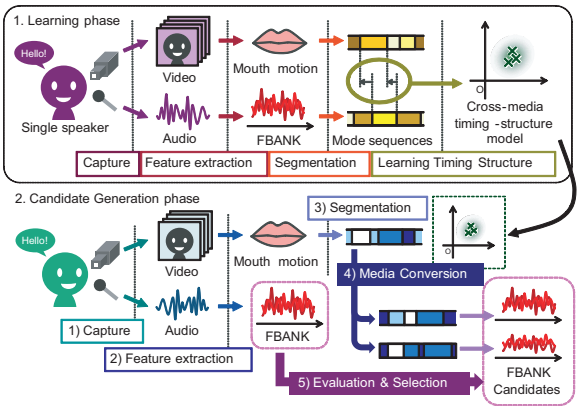


図 2　音声候補生成の流れ

て，口唇運動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，口唇運動と音声特徴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．

候補生成フェーズでは，新たに観測された映像・音響信号から特徴系列を抽出し，口唇運動について分節化処理を行う．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，その口唇運動の時区間系列に合致すると推定される音声の時区間系列の候補を生成する．文献 [3] では，本研究とは逆に音声からの口唇運動生成を扱っているため，Viterbi アルゴリズムを利用して最適な（映像側の）時区間系列をひとつ生成しているが，本手法では，ある口唇運動に対応する音声信号は一般に多数存在することから，(音声側の）時区間系列を複数候補生成できるように，Parallel List Viterbi アルゴリズム [5] を利用した拡張を行う（5. 節）．そして，候補として生成された時区間系列を元に，先に学習したメディア信号モデルを構成する線形システムを時間軸上で切り替え，音声特徴系列候補を複数生成する．

続く雑音抑圧（クリーン音声推定）フェーズでは，生成した特徴系列候補 $\hat{S}^{(c)}\ (c = 1, ..., C)$ と，観測された雑音重畳音声の特徴系列 X とから，最終的な発話音声を推定する．これには，文献 [6] で用いられているような音声の雑音抑圧手法を応用する．次節で詳述するように，これはクリーン音声として何らかの事前分布が必要な手法であり，通常は固定的な混合ガウス分布（Gaussian Mixture Model, GMM）が用いられるが，我々の提案手法では，視覚情報から生成された音声候補を利用して，動的に変化するようなクリーン音声の GMM を構成して雑音抑圧を高精度に行う．

3. パーティクルフィルタによる雑音追跡と音声分離

文献 [6] の雑音抑圧手法では，まず非定常雑音の時間変化が Random Walk 過程に従うと仮定しておく．次に，この雑音成分に，GMM に従うような音声が加わることで雑音重畳音声 X が発生するような状態空間モデルを考え，X の観測が得られた際に，雑音側 N をパーティクルフィルタで追跡することで，クリーン音声 S を分離・推定する．

音声のみを用いる従来手法では，クリーン音声の分布として固定的な GMM を仮定するため，音声の SN 比がある程度大きくなると，雑音の追跡が困難になる．そこで提案手法は，口唇運動から音声特徴系列の候補を生成し，時間的に変化する GMM を構成することで，高精度に雑音を追跡し，音響信号からクリーン音声を分離する．

以下では，まず音声のみを用いる従来の非定常雑音抑圧手法について述べ，次に視覚情報から生成した音声候補を利用する提案手法について述べる．

3.1　音声情報のみを用いた雑音抑圧手法

ここで扱う音響信号は，短時間の窓フレームによってスペクトル解析が行われ，音響特徴系列に変換されているとする（具体的な特徴量は 6.2 節で説明）．このとき，t 番目のフレームにおける観測音声，クリーン音声，加法性雑音の対数メルスペクトルをそれぞれ $\boldsymbol{x}_t, \boldsymbol{s}_t, \boldsymbol{n}_t$ とすると，各特徴量の関係は次式で表される．

$$\begin{aligned}\boldsymbol{x}_t &= \log(\exp(\boldsymbol{s}_t) + \exp(\boldsymbol{n}_t)) + \boldsymbol{v}_t \\ &= \boldsymbol{s}_t + \log(\mathbf{1} + \exp(\boldsymbol{n}_t - \boldsymbol{s}_t)) + \boldsymbol{v}_t \\ &= f(\boldsymbol{s}_t, \boldsymbol{n}_t) + \boldsymbol{v}_t \end{aligned} \quad (1)$$

ここで，$\boldsymbol{v}_t$ は加法性雑音重畳モデルの誤差成分（乗法性の歪みなど）を表し，正規分布 $\mathcal{N}(\mathbf{0}, \Sigma_x)$ に従うとする．また，Σ_s は $\boldsymbol{s}_t$ の共分散行列である．$\mathbf{1}$ は全要素が 1 であるようなベクトルであり，

$$f(\boldsymbol{s}_t, \boldsymbol{n}_t) \triangleq \boldsymbol{s}_t + \log(\mathbf{1} + \exp(\boldsymbol{n}_t - \boldsymbol{s}_t))$$

とおいた．すなわち式 (1) は，$\boldsymbol{s}_t$ を入力として持つような，雑音 $\boldsymbol{n}_t$ の非線形観測方程式である．

一方 $\boldsymbol{n}_t$ の時間推移は，$\boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \Sigma_\omega)$ を用いて，以下の Random Walk 過程とする．

$$\boldsymbol{n}_{t+1} = \boldsymbol{n}_t + \boldsymbol{\omega}_t \quad (2)$$

すなわち，観測される音響信号 $\boldsymbol{x}_t$ は，雑音 $\boldsymbol{n}_t$ を状態と考えれば，式 (1),(2) の状態空間モデルにより表される．

次に，この雑音成分を推定する方法について述べる．式 (1) の非線形性から，これには以下のようなパーティクルフィルタを用いる．まず，$\boldsymbol{n}_{0:t} = [\boldsymbol{n}_0, \ldots, \boldsymbol{n}_t]$ とすると，$\boldsymbol{x}_{0:t}$ が観測されたときの $\boldsymbol{n}_{0:t}$ の事後確率分布は，

$$p(\boldsymbol{n}_{0:t}|\boldsymbol{x}_{0:t}) = \frac{p(\boldsymbol{n}_0, \boldsymbol{x}_0)}{p(\boldsymbol{x}_{0:t})} \prod_{u=1}^{t} p(\boldsymbol{n}_u|\boldsymbol{n}_{u-1}) p(\boldsymbol{x}_u|\boldsymbol{n}_u)$$

のようにマルコフ連鎖を用いて表せるが，この事後確率密度関数を次式のように近似する．

$$p(\boldsymbol{n}_{0:t}|\boldsymbol{x}_{0:t}) \simeq \frac{1}{K}\sum_{k=1}^{K}\delta(\boldsymbol{n}_{0:t}-\boldsymbol{n}_{0:t}^{(k)}) \simeq \sum_{k=1}^{K} w_t^{(k)} p(\boldsymbol{n}_{0:t}^{(k)}|\boldsymbol{x}_{0:t})$$

ここで，k はパーティクル番号，K はパーティクルの総数，$w_t^{(k)}$ は時刻 t での第 k パーティクルに対する重み（importance weight）を示す．$p(\boldsymbol{n}_{0:t}|\boldsymbol{x}_{0:t})$ 自体からのサンプリングは困難であるため，importance density と呼ばれる $q(\boldsymbol{n}_{0:t}|\boldsymbol{x}_{0:t})$ を導入する．各パーティクルの重みは，

$$w_t^{(k)} \propto \frac{p(\boldsymbol{n}_{0:t}^{(k)}|\boldsymbol{x}_{0:t})}{q(\boldsymbol{n}_{0:t}^{(k)}|\boldsymbol{x}_{0:t})} \propto w_{t-1}^{(k)} \frac{p(\boldsymbol{n}_t^{(k)}|\boldsymbol{n}_{t-1}^{(k)})p(\boldsymbol{x}_t|\boldsymbol{n}_t^{(k)})}{q(\boldsymbol{n}_t^{(k)}|\boldsymbol{n}_{0:t}^{(k)},\boldsymbol{x}_{0:t})}$$

と表せるが，$p(\boldsymbol{n}_t^{(k)}|\boldsymbol{n}_{t-1}^{(k)}) = q(\boldsymbol{n}_t^{(k)}|\boldsymbol{n}_{0:t}^{(k)},\boldsymbol{x}_{0:t})$ と仮定することで，$w_t^{(k)} \propto w_{t-1}^{(k)}\, p(\boldsymbol{x}_t|\boldsymbol{n}_t^{(k)})$ と簡略化できる．

パーティクルフィルタによる雑音追跡において，クリーン音声成分 $\boldsymbol{s}_t$ はクリーン音声 GMM からのサンプリングによって決定する．すなわち，GMM の混合数を L_s，l 番目の混合分布の平均ベクトルを $\boldsymbol{\mu}_{s,l}$，共分散行列を $\Sigma_{s,l}$，混合重みを $w_{s,l}$ とすれば，毎フレームで以下の分布からサンプリングを行う．

$$p(\boldsymbol{s}_t) = \sum_{l=1}^{L_s} w_{s,l}\mathcal{N}(\boldsymbol{\mu}_{s,l},\Sigma_{s,l}) \tag{3}$$

式 (1) の非線形性により，各パーティクルの更新には，拡張カルマンフィルタを利用する．このとき，確率密度関数 $p(\boldsymbol{n}_{0:t}^{(k)}|\boldsymbol{x}_{0:t})$ のパラメータ（平均ベクトル $\hat{\boldsymbol{n}}_t^{(k)}$，共分散行列 $\Sigma_{n_t}^{(k)}$）推定は以下のようになる．

$$\begin{aligned}
\boldsymbol{n}_{t|t-1}^{(k)} &= \boldsymbol{n}_{t-1}^{(k)} \\
\Sigma_{n_{t|t-1}}^{(k)} &= \Sigma_{n_{t-1}}^{(k)} + \Sigma_\omega \\
K_t^{(k)} &= \Sigma_{n_{t|t-1}}^{(k)} F_t^{(k)\mathrm{T}} \left[F_t^{(k)}\Sigma_{n_{t|t-1}}^{(k)} F_t^{(k)\mathrm{T}} + \Sigma_s\right]^{-1} \\
F_t^{(k)} &= \partial f(\boldsymbol{s}_t^{(k)},\boldsymbol{n}_{t|t-1}^{(k)})/\partial \boldsymbol{n}_{t|t-1}^{(k)} \\
\hat{\boldsymbol{n}}_t^{(k)} &= \boldsymbol{n}_{t|t-1}^{(k)} + K_t^{(k)}(\boldsymbol{x}_t - f(\boldsymbol{s}_t^{(k)},\boldsymbol{n}_{t|t-1}^{(k)})) \\
\Sigma_{n_t}^{(k)} &= \Sigma_{n_{t-1}}^{(k)} - K_t^{(k)} F_t^{(k)} \Sigma_{n_{t|t-1}}^{(k)}
\end{aligned}$$

ここで $(t|t-1)$ は $(t-1)$ フレームから予測された t フレームのパラメータであることを示す．さらに，重み $w_t^{(k)}$ の値が微小であるサンプルは事後確率分布を近似するサンプルとして相応しくないため，これらを破棄し，大きな重みを持つサンプルを分割してサンプル総数を維持する Residual Resampling [7] を行う．

以上の手順で推定された雑音の確率分布を用いて，MMSE 推定法 [8] に基づくクリーン音声の推定（雑音抑圧）を行い，最終的な推定クリーン音声を得る．

$$\hat{\boldsymbol{s}}_t = \boldsymbol{x}_t - \sum_{l=1}^{L_s} P(l|\boldsymbol{x}_t)(f(\boldsymbol{\mu}_{s,l},\boldsymbol{n}_t) - \boldsymbol{\mu}_{s,l})$$

$$P(l|\boldsymbol{x}_t) = \frac{w_{s,l}\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_{x,l},\Sigma_{x,l})}{\sum_{m=1}^{L_s} w_{s,m}\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_{x,m},\Sigma_{x,m})}$$

ただし，$\boldsymbol{\mu}_{x,l},\Sigma_{x,l}$ は観測モデル（式 (1)）の平均ベクトル，共分散行列であり，これらは Vector Taylor Series 法 [9] とパラメータ $\boldsymbol{\mu}_{s,l},\Sigma_{s,l},\hat{\boldsymbol{n}}_t^{(k)},\Sigma_{n_t}^{(k)}$ を用いて近似的に推定することができる．

3.2 視覚情報を手がかりとした音声信号分離

前節では観測信号中に含まれる音声特徴の分布に関する事前知識として，式 (3) のような GMM を用いている．この分布はあらかじめ学習され，雑音追跡時には時間的に変化しないが，本研究では視覚情報を手がかりとして，この分布を動的に変化させることで，雑音追跡の精度を高めるとともに，音声を正確に推定する．具体的には，次節以降で生成する口唇運動の特徴系列より推定される音声特徴系列の候補 $\hat{S}^{(c)} = [\ldots,\hat{\boldsymbol{s}}_t^{(c)},\ldots]\,(c=1,...,C)$，および尤度より計算される混合比（信頼度）$W_c$ を用いて以下の分布を構成する．

$$p(\boldsymbol{s}_t|V) = \sum_{c=1}^{C} W_c\mathcal{N}(\hat{\boldsymbol{s}}_t^{(c)},\Sigma_{s_c}) \tag{4}$$

口唇運動の特徴系列 V を高い信頼性で抽出できる場合，$p(\boldsymbol{s}_t|V) \sim p(\boldsymbol{s}_t)$ と近似できる．次節以降では，候補の生成方法について具体的に述べる．

4. 時区間ハイブリッドダイナミカルシステム

各フレームにおいて口唇運動と音声の特徴量を抽出することで特徴系列が得られる．これら特徴系列は，再現性のある要素的な変化パターンの組み合わせによって表現できる場合が多く，本研究では，文献 [3] で用いられる時区間ハイブリッドダイナミカルシステム（以後，単に HDS）を利用して，各要素的な変化を線形システムでモデル化する．モデルの詳細，および学習，生成，分節化の具体的アルゴリズムに関しては紙面の都合から文献 [3], [10] に委ね，以下ではその概要を述べる．

4.1 システムアーキテクチャ

用いる HDS 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 層は複数の線形システム $\mathcal{D} = \{D_1,\ldots,D_N\}$ により共有される n 次元の内部状態空間である（図 3（左））．第 1 層の離散状態遷移では，各離散状態の順序だけが決まるため，各状態が活性化される物理的時間長が必要となる．そこで，これら 2 層の統合のために時区間を導入し，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，内部状態の遷移ダイナミクスとその切り替わるタイミングをオートマトンにより制御できる．

4.1.1 線形システム

本研究では線形システムとして，以下の R 次の多変量自己回帰モデルを用いる．

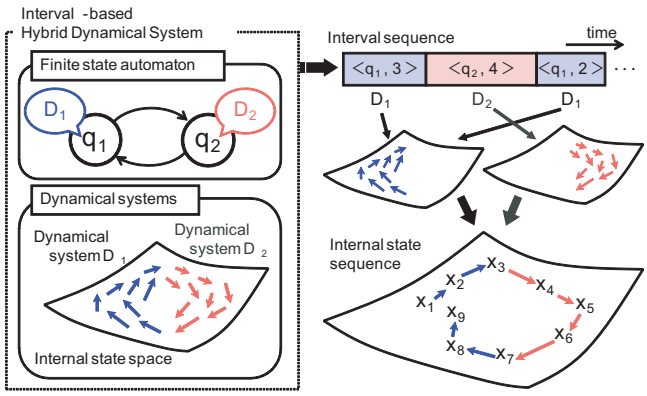


図 3　時区間ハイブリッドダイナミカルシステム

$$\boldsymbol{y}_t = \sum_{r=1}^{R} A_r^{(i)} \boldsymbol{y}_{t-r} + \boldsymbol{b}^{(i)} + \boldsymbol{\epsilon}_t^{(i)}$$

ただし，$\boldsymbol{y}_t$ は時刻 t における特徴ベクトルを表す m 次元ベクトル，$A_r^{(i)}$ は遷移行列，$\boldsymbol{b}^{(i)}$ はバイアスである（内部状態は $n = mR$ 次元となる）．また，$\boldsymbol{\epsilon}_t^{(i)}$ 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 $A_r^{(i)}$, $\boldsymbol{b}^{(i)}$, $\boldsymbol{\epsilon}_t^{(i)}$ の分布をパラメータとして持つ．

4.1.2　区間に基づくオートマトンの離散状態遷移

HDS 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．まず，生成される時区間系列 $\mathcal{I} = \{I_1, \ldots, I_K\}$ に関して，時間的に隣接する区間の離散状態および区間長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，区間長に関する分布や，状態遷移に関する分布を持つことで，4.3 節の区間系列の生成（各線形システムの活性化タイミングの決定）や，特徴ベクトルを分節化する際の尤度計算に用いることができる．

4.2　HDS モデルの 2 段階学習法

学習データとして，特徴ベクトル系列のみが与えられているとする．このとき，まず固有値制約に基づく線形システムの階層的クラスタリングにより，システムの個数およびパラメータ概値を推定する．次に，線形システムの個数を固定して (近似的な) EM アルゴリズムを行い，パラメータを調整する．この 2 段階の学習法により，HDS のシステム同定と学習データの分節化処理を同時に行うことができる [10].

4.3　HDS による信号の生成と分節化

HDS は確率的生成モデルであり，学習された HDS は学習時と類似したベクトル系列を生成することが可能となる．まず，オートマトンが区間系列を生成し，各離散状態（線形システム）の活性化タイミングを決定する．その後，活性化された線形システムは，それぞれのダイナミクスに基づいて内部状態を遷移させ，これによって観測空間における信号を生成できる（図 3（右））．

一方，観測系列（特徴ベクトル系列）が与えられると，HDS は，どのタイミングで線形システムを切り替えると，元の系列を最もよく表現できるかを，尤度に基づいて計算する．これによって観測系列は，線形システムの切り替わる時点によって分節化され，区間系列に変換することができる．

5.　タイミング構造モデルに基づく複数音声候補の生成

前節で述べた HDS を，口唇運動および音声の双方の特徴系列でそれぞれ学習しておき，これらを HDS_v, HDS_s とする．HDS を用いることで，特徴系列の背後に（線形システムのラベル付）区間系列の表現を持つことができる．この区間系列表現を用いることで，1. 節の最後で述べたような問題を解決しながら，口唇運動の特徴系列 $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{T_v}]$ から音響特徴系列候補 $\hat{S}_c = [\hat{\boldsymbol{s}}_1^{(c)}, \ldots, \hat{\boldsymbol{s}}_{T_s}^{(c)}]$ の生成が可能となる．これは文献 [3] と同様（ただしモダリティは逆）の，以下の流れとなる．

（1）　HDS_v を用いて，新たに得られた口唇運動の特徴系列 V を区間系列 $\mathcal{I}^{(v)} = \{I_1^{(v)}, ..., I_{K_v}^{(v)}\}$ へ分節化する．

（2）　区間系列 $\mathcal{I}^{(v)}$ から音声特徴系列の背後にある区間系列 $\mathcal{I}^{(s)} = \{I_1^{(s)}, ..., I_{K_s}^{(s)}\}$ を推定・生成する．

（3）　生成された区間系列 $\mathcal{I}^{(s)}$ から，HDS_s を用いて音声特徴系列 $\hat{S}$ を生成する．

このうちステップ (1), (3) は 4.3 節で概略を述べたが，ステップ (2) については次の手順で行う．まず，音声および口唇運動の区間系列において，時間的なオーバラップを伴って現れるような 2 つの区間 $I_k^{(s)} = [b_k^{(s)}, e_k^{(s)}]$ と $I_{k'}^{(v)} = [b_{k'}^{(v)}, e_{k'}^{(v)}]$ に関して，それらの線形システムのラベルを，それぞれ $m_k^{(s)}, m_{k'}^{(v)}$ とする．このとき，HDS_s と HDS_v を構成する線形システムの集合をそれぞれ $\mathcal{D}^{(s)} = \{D_i^{(s)}\}_{i=1}^{N_s}, \mathcal{D}^{(v)} = \{D_{i'}^{(v)}\}_{i'=1}^{N_v}$ とすれば，全ての線形システム対 $(D_i^{(s)}, D_{i'}^{(v)})$ に関して，

$$P(m_k^{(s)} = D_i^{(s)}, m_{k'}^{(v)} = D_{i'}^{(v)} \mid I_k^{(s)} \cap I_{k'}^{(v)} \neq \emptyset),$$
$$P(b_k^{(s)} - b_{k'}^{(v)}, e_k^{(s)} - e_{k'}^{(v)} \mid m_k^{(s)}, m_{k'}^{(v)}, I_k^{(s)} \cap I_{k'}^{(v)} \neq \emptyset)$$

という分布を学習しておく．1 つ目は，モダリティ間で共起しやすい線形システム対をモデル化し，2 つ目は，各線形システム対がどの程度の時間差で開始・終了するかをモデル化する．これらを合わせて，「タイミング構造モデル」と呼び，そのパラメタ集合を Φ で表す．

あらかじめ学習したタイミング構造モデル Φ を用い，先のステップ (2) を実現するには，区間系列 $\mathcal{I}^{(v)}$ が与えられたときに，この区間系列とともに生じ得る区間系列 $\mathcal{I}^{(s)}$ のうち，最も尤度の高いものとして以下のように推定できる．

$$\hat{\mathcal{I}}^{(s)} = \arg\max_{\mathcal{I}^{(s)}} P(\mathcal{I}^{(s)} | \mathcal{I}^{(v)}, \Phi) \qquad (5)$$

式 (5) は Viterbi アルゴリズムを用いて解くことができ，

特徴系列長を T とすれば，具体的には時間範囲 $[1,T]$ において分節化された区間数 K_s，および各区間の終了時刻 $e_k^{(s)}$ とそのモード $m_k^{(s)}(k=1,...,K_s)$ を決めることができる．これにより推定される区間系列はひとつであるが，上位 C 個の尤度を持つ区間系列を推定するために，文献 [5] の Parallel List Viterbi アルゴリズムを用いる．これは，各時刻の各状態において上位 C 個の尤度とそれを実現する前状態，およびその順位を記録していくものであり，Viterbi アルゴリズムと同様に最終時刻からのトレースバックにより C 個の候補を求めることができる．

6. 実　　験

実験に用いたデータは，日本語 5 母音/あ//い//う//え//お/の孤立音発話を各音素について 5 サンプルずつキャプチャしたものである．画像の解像度は 640×480 画素（VGA），フレームレートは 60fps，グレー画像という設定で撮影を行った．撮影した画像の例を図 4(a) に示す．音声はサンプリングレート 48kHz，16 ビット量子化で録音し，その後 16kHz にダウンサンプリングした．

6.1 口唇運動の特徴抽出

画像から抽出する特徴量は，発声する音素による口唇運動の差異を詳細に表現できることが望ましい．そのため，口唇形状だけではなく，歯の見え隠れなどの画像情報を保持する特徴量が適当であると考えられる．そこで，あらかじめ学習した Active Appearance Model (AAM) [11] を用いて，各フレームの画像から口唇輪郭の特徴点座標を抽出し（図 4(b)），この口唇輪郭の重心座標を基準として話者の口領域を矩形領域として切り出した．その後，32×32 画素にダウンサンプリングし，画像の周辺を滑らかにマスキングした（図 4(c)）．得られた口領域画像集合について主成分分析を適用することで，上位 20 の主成分を要素とするようなベクトルの時系列を求め，これを口唇運動の特徴系列として用いた．

6.2 音響信号の特徴抽出

音声認識のための特徴量としては，MFCC（Mel-Frequency Cepstrum Coefficient）が広く用いられており，特に入力音声に雑音が重畳しない場合，その有効性が確認されている．しかし，ケプストラム領域は，対数スペクトルをフーリエ変換した領域であるため，スペクトル領域のある範囲にのみ重畳した雑音であっても，ケプストラム領域ではその影響が全ての項に及んでしまう [12]．このため，複数音源からの加法性雑音に対しては，スペクトル領域での特徴を用いる方が信号レベルでの雑音分離にとって望ましいと考えられる．そこで本研究では，3. 節のスペクトル特徴量として，log フィルタバンク係数（以下 FBANK）を用いる．

一方で，FBANK のようなノンパラメトリック分析と異なるアプローチとして，音声の発生モデルを仮定し

図 4　(a) キャプチャした画像の例，(b)AAM による特徴点抽出結果の例，(c) 唇特徴点の追跡結果に基づいて口領域を切り出した画像の例．

てそのモデルパラメータを抽出するパラメトリック分析がある．4.2 節の学習および 4.3 節の分節化処理では特徴系列の時間的滑らかさが重要であり，これについてはパラメトリック分析によって得られる LSP（line spectrum pair：線スペクトル対）係数がやや優位性が高かったためこれを用いた．なお，FBANK は HMM Tool Kit (HTK) [13] を，LSP は Speech Signal Processing Toolkit (SPTK) [14] を用いて抽出した．フレーム幅はいずれも 25ms であり，フレームのステップ幅は 1/60 s として，映像と同期したレートとした（異なるレートでもよいが，実装の簡略化のために同一レートを用いた）．

6.3 発話音声の候補生成

口唇運動および音声の特徴系列より，それぞれ HDS の学習および系列の分節化を行った．各 HDS の離散状態数は，まずモデル化誤差カーブを基準に候補を出し，母音が区別される程度の精度になるように，口唇運動は 8，音声は 10 と半手動で決定した．また，線形システム（自己回帰モデル）の次数は $R=2$ とした．その後，口唇運動と音声の時区間系列より，タイミング構造モデルを学習した．ここでは，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 2 次元ガウス分布（標準偏差 3 フレーム，共分散 0）(注2) を畳み込むことによって，時間差に関する分布を得た．

続いて，学習したタイミング構造モデルを用いて，「/あ//い//う//え//お/」の孤立音発話の 1 サンプルの口唇運動データに対して，5. 節の手法により複数音声候補の生成を行った．ここでは候補系列の生成数を，その生成時の尤度を基準とした上位 50 系列とした．生成された候補系列の例を図 5 に示す．候補によって，特徴量が変化する時間領域に差が見られることが確認できる．特に上位の候補については，元の音声に比較的近い特徴系列となっていることが分かる．

6.4 音声候補の平均を用いた雑音抑制

タイミング構造モデルに基づいて口唇運動から生成した候補音声を用いて，3. 節で述べたようなパーティクル

(注2)：映像と音声の時間的ずれに関する人の許容範囲は，映像に対して音声が進む場合は約 100ms，遅れる場合は約 250ms 程度との知見があり [15]．今回は $\sigma \simeq 50$ms となるようにした．

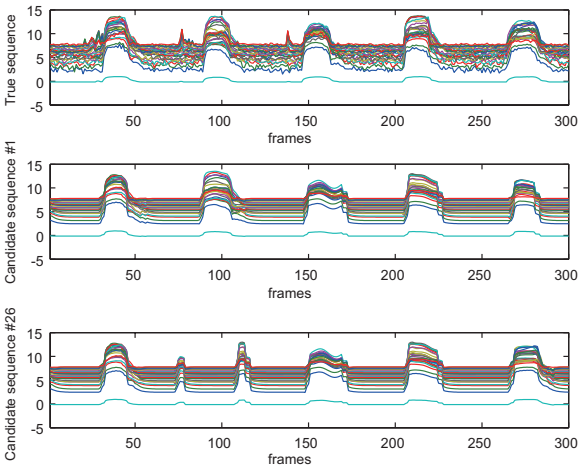


図 5　口唇運動より生成された音声特徴 FBANK 系列（24 次元+音声パワー）の候補例.（上段）クリーン音声から抽出された特徴系列（推定の真値）,（中段）生成された音声候補で尤度が最大のもの,（下段）生成された音声候補で尤度が 26 番目のもの.

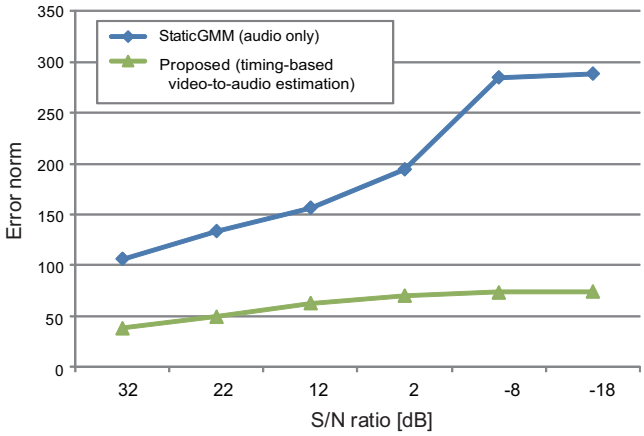


図 6　SN 比と誤差ノルムの関係（右に行くほど雑音レベルが大きい）および音声のみを用いる従来手法との比較

フィルタを用いた雑音抑制（クリーン音声推定）を行った．対象データは「/あ//い//う//え//お/」の孤立音発話（特徴空間で 300 フレーム）の音声データに，雑音を人工的に重畳させたものを使用する．重畳する雑音には空調音と工場雑音の 2 種類を用いた．空調音はエアコン吹き出し口付近で収録した定常雑音であり，工場雑音（電子協騒音 DB [16] より）はエアーレンチ等の工作機械の動作音を含む非定常性の強い雑音である．音声推定結果の特徴は大きくは変わらなかったため，以下のグラフ等では，このうち工場雑音に関する結果を記載する．

これらの雑音を，クリーン音声に対して SN 比 $-18, -8, 2, 12, 22, 32$dB(注3)の 6 通りで重畳した音響信号を作成し，それらから FBANK 特徴量の抽出を行ったのち，音声および雑音を表す GMM のパラメータの学習を HTK により行った（音声の GMM の混合数は 13 とした）.

その後，前節で求めた音声候補を使う提案手法（式 (4) を利用）と，音声のみを用いる場合（式 (3) を利用) に関して，それぞれ雑音抑制（クリーン音声推定）を行った．この際，$\Sigma_\omega = \mathrm{diag}(0.01), \Sigma_{s_c} = \mathrm{diag}(1.0)$ とし，パーティクルのサンプル総数は 50 とした．また，提案手法において用いる音声候補は，タイミング構造モデルに基づいて推定された 50 個の候補信号の平均とした（つまり式 (4) で，まずは単一ガウス分布を用いた）．このときの，正解音声と推定音声との誤差ノルムを図 6 に示す．パーティクルフィルタにおいて乱数を使用するため，各 SN 比において同じ設定で 3 回実験行い，それら誤差ノルムの平均を取っている．

SN 比が低下するほど分離の精度が低くなる様子がみられる．しかし，比較手法では特に 0dB 付近以下という雑音が大きく重畳した場合に極端な精度の低下がみられる一方で，口唇運動から推定される音声候補を利用する提案手法の場合，0dB 付近以下の状況でも音声の推定精度の低下が抑えられている．SN 比が大きな領域では，観測信号中に含まれる雑音成分が小さく，クリーン音声から学習した GMM モデルに合致する性質を持った信号が大きいため，固定的な GMM を用いても十分に音声の推定が行えている．しかし，SN 比が大きくなると，実質的に観測信号中に含まれるクリーン音声の成分が雑音に埋もれてしまい，固定的な GMM では追跡できなくなる．一方で，視覚情報に基づいてこの GMM を動的に変化させることで，雑音の追跡精度が向上したと考えられる．

6.5　複数の音声候補を用いた雑音抑制

前節の実験では，音声候補信号として，複数生成した候補の平均値を用いたが，生成されたこれらの候補の中には所望の信号をよく推定できているものもあれば，大きく異なる波形をもつものも存在する．そのため，パーティクルフィルタによる雑音追跡では，観測される音響信号の情報に基づいて，正しい候補の情報を選択的に利用した推定を行うことが必要である．このことを検証するための実験として，雑音追跡の際のクリーン音声モデル（式 (4) の時間変化する GMM）を

- 最も原音声に近い候補を利用（Best）
- 50 候補中最も原音声と異なる候補を利用（Worst）
- 上の 2 つの候補をともに利用（Best+Worst）

の 3 種類（1,2 番目は単一ガウス分布，3 番目は混合数 2 の GMM で混合比は尤度より決定）としてクリーン音声推定を行い，Best+Worst での結果が Best に近いものとなるか（複数候補から正しい候補が選択的に利用できているか）を検証する．ただし，ここでは候補の選択について焦点を絞った実験を行うため，候補の順位付けが正しく行われたことを保証できるよう，生成音声候補と原音声との誤差ノルムを用いて，生成候補の順位付けをし直した．

(注3)：dB の値に一部誤りがあったため修正．以降も同様 (2010.7.27).

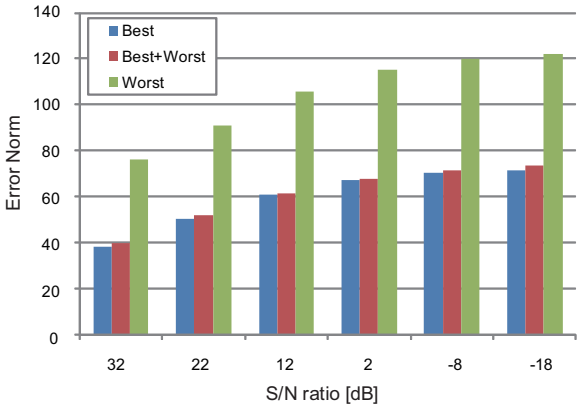


図 7　使用候補と誤差ノルムの関係（Best:原音声に最も近い系列，Worst:最も遠い系列，Best+Worst:両者の混合）

クリーン音声推定における各種パラメータは前節と同様として，SN 比を $-18, -8, 2, 12, 22, 32$dB と変えながら，使用する候補の組み合わせ 3 通り（Best, Worst, Best+Worst）についてそれぞれ誤差ノルムを求めた．これらの結果を図 7 に示す．

グラフより，使用する候補が Best+Worst の場合の音声推定精度が，Best の候補を用いた場合に近いものになっており，観測された音響信号に基づいて正しい候補を選択的に利用できているといえる．Best のみに比べてやや精度が低下しているのは，パーティクルの更新時に使用する音声に，Worst の候補からサンプリングされたものがある程度の確率で含まれるため，それらが影響していると考えられる．

今回の実験は予備的な位置づけであり，候補として Best/Worst の 2 個を選択して実験を行ったが，実際上はさらに多数のクリーン音声候補が得られるため，それらを全て利用することができる．さらに，複数の候補の中で，真のクリーン音声と近い時間範囲が部分的に現れていれば，パーティクルフィルタによる雑音追跡時には，それらが選択的に利用されていくことが期待できるため，全体としては精度の向上が可能であると考えられる．

7. おわりに

本研究では，発話時の口唇運動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口唇運動に合った発話音声の候補を生成することで，これら生成された候補と実際に観測された音響信号とから，発話音声を精度よく推定する方法を提案した．本論文に示した実験は予備的なものではあるが，音声候補を利用したクリーン音声推定を行った場合，音声モデルとして固定的な GMM を用いる場合と比較して，特徴空間での音声信号推定の精度が向上することが確認された．一方で，口唇運動から音声候補を生成・推定する方法としては，他にも HMM を利用するものや単純に回帰モデルを利用する方法などが考えられ，どのような候補生成方法が実際に有効であるかについては，孤立音だけでなく調音結合や子音を含むような発話を用いて，大規模なデータから，今後詳細な比較評価を行う必要がある．

謝辞:　本研究の一部は，科学研究費補助金 18049046 および 21680016 の補助を受けて行った．

文　献

[1] T. Chen and R. R. Rao: “Audio-visual integration in multimodal communication”, Proceedings of the IEEE, pp. 837–852 (1998).
[2] 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).
[3] 川嶋, 松山：“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).
[4] 實廣, 鳥山, 小暮：“実環境下音声認識のためのパーティクルフィルタを統合した複数モデル雑音抑圧手法”, 電子情報通信学会論文誌 D, **91**, 10, pp. 2519–2528 (2008).
[5] N. Seshadri and C.-E. Sundberg: “List Viterbi decoding algorithms with applications”, IEEE Transactions on Communications, **42**, 234, pp. 313–323 (1994).
[6] M. Fujimoto and S. Nakamura: “A non-stationary noise suppression method based on particle filtering and polyak averaging”, IEICE Transactions on Information and Systems, **89**, 3, pp. 922–930 (2006).
[7] S. Arulampalam, S. Maskell, N. Gordon and T. Clapp: “A tutorial on particle filters for on-line non-linear/non-gaussian Bayesian tracking”, IEEE Transactions on Signal Processing, **50**, 2, pp. 174–188 (2002).
[8] J. C. Segura, A. D. L. Torre, M. C. Benitez and A. M. Peinado: “Model-based compensation of the additive noise for continuous speech recognition. experiments using the AURORA II database and tasks”, Proc. EuroSpeech, **1**, pp. 221–224 (2001).
[9] P. Moreno, B. Raj and R. Stern: “A vector Taylor series approach for environment-independent speech recognition”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, **2**, pp. 733–736 (1996).
[10] 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).
[11] T. F. Cootes, G. J. Edwards and C. J. Taylor: “Active appearance model”, Proc. European Conference on Computer Vision, pp. 484–498 (1998).
[12] 西村, 篠崎, 岩野, 古井：“周波数帯域ごとの重みつき尤度を用いた雑音に頑健な音声認識”, 信学技報, SP2003-116 (2003).
[13] S. Young and et al.: “The HTK Book (for HTK Version 3.4)”, Cambridge University Engineering Department (2006).
[14] K. Tokuda and SPTK Working Group: “Reference Manual for Speech Signal Processing Toolkit Ver. 3.2” (2008).
[15] 鎧沢, 滝川, 大久保, 渡辺：“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).
[16] S. Itahashi: “A noise database and Japanese common speech data corpus”, The Journal of the Acoustical Society of Japan, **47**, 12, pp. 951–953 (1991).