様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５月１６日現在

機関番号：１４３０１
研究種目：若手研究（Ａ）
研究期間：　2009　～　2010
課題番号：２１６８００１６
研究課題名（和文）
視覚から聴覚系へのフィードバック機構に基づく発話解析手法の構築
研究課題名（英文）
A Method for Speech Analysis Based on a Visual-to-Auditory Feedback Mechanism
研究代表者
川嶋　宏彰（KAWASHIMA HIROAKI）
京都大学・情報学研究科・講師
研究者番号：４０３４６１０１

研究成果の概要（和文）：口元の動きと音声の時間同期構造に注目した、新たな発話解析手法を構築した。本手法は、力学系と離散事象系を統合した数理モデルであるハイブリッドシステムを用いることで、カメラとマイクロフォンにより得られる信号をそれぞれ分節化して表現したうえで、両信号間の詳細な時間同期構造モデルを統計的に学習するという特徴を持つ。このモデルを用いることで、「口元の動きから音声」という信号生成機構を実現でき、非定常雑音環境下での高精度な音声分離・推定を可能とした。

研究成果の概要（英文）：We have developed a novel speech-analysis method based on the detail modeling of temporal relationship between mouth movements and speech signals. First, we use a hybrid system, which is an integrated model of dynamical systems and discrete-event systems, as a mathematical tool to segment and model multimedia signals such as captured mouth motion and speech data. Then, we build a statistical cross-media timing model that can be learned from those segmented data. The proposed method realizes the mechanism of signal generation “from mouth motion to speech”, which enables highly accurate speech estimation in non-stationary noise environment.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	4,600,000	1,380,000	5,980,000
２０１０年度	2,100,000	630,000	2,730,000
総　計	6,700,000	2,010,000	8,710,000

研究分野：　情報学
科研費の分科・細目：　情報学　知覚情報処理・知能ロボティクス
キーワード：　音声推定・分離、口唇運動、線形システム、ハイブリッドシステム、タイミング構造、視聴覚統合、マルチモダリティ、時系列の分節化

１．研究開始当初の背景

音声対話システムが用いられる状況は、電話の自動音声応答をはじめとする単一話者の利用から、会議の議事録作成や対話分析、街角情報パネル、運転環境など、非定常雑音や複数話者が存在する場面へ広がりつつある。このような状況では、いつどの話者が発話したかを知るとともに、複数話者の発話のオーバラップや雑音の重畳に対処しながら音声分離や音声認識を行う必要があるが、話者に常に接話型マイクを装着してもらうことは負担が大きい。そこで、認識の前段階でマイクロフォンアレイや独立成分分析などを用いて、話者位置同定や発話区間同定、発話分離を行う手法が提案されている。ところが、音源（話者）が密集する場合や、反響音が存在する場合には、これら音響信号処理のみでは十分対応できないという問題がある。

一方、テレビ会議や運転状況など、話者の動きが比較的限られている場合は、話者の映

像を取得することは比較的容易である。そこで、音声と映像情報を統合することによる話者位置・発話区間同定などの発話解析や音声認識手法(audio-visual speech recognition, AVSR)が提案されている。これらは、カメラとマイクロフォンにより得られた映像・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、隠れマルコフモデルやパーティクルフィルタと呼ばれるような手法を用いてその時間変化を記述する（いずれも特徴量の時間変化を状態遷移によってモデル化する手法である)。その上で、同時刻や隣接時刻において、映像・音響の特徴量（もしくは状態）が持つ共起性をモデル化・評価する。

顔追跡やマイクロフォンアレイを統合することで、話者位置や発話区間の同定精度は向上する。しかし、講義や会議などの話者が密集する場合には、顔と音源位置を結び付けることが難しい。より詳細な情報として、唇の動きと音響特徴の共起性を見ることが考えられ、AVSR にも広く利用されているが、複数話者発話や強い非定常雑音が存在する状況に対してはしばしば精度が確保できず、音声分離への応用も限られているのが現状である。これらの問題は、口元の動きと音声とを十分詳細に結び付けられていないことに、ひとつの原因があると考えられる。

２．研究の目的

特に従来の AVSR の枠組みでは、音声特徴と口唇運動特徴とを対等なものとして扱ってモデル化する場合が多く、認識時には視覚・聴覚情報のうち信頼性の低い情報に低い重みを与えて統合することで、音声や視覚情報を単独で用いるよりも高い認識精度を実現する。たとえば、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 比が小さくなるほど、音声側に低い重みを設定する。しかし音声の SN 比がある程度小さい場合、その識別精度が視覚情報に主に依存することになる。つまり、音響情報側の切り捨てが行われ、聴覚情報が有効に利用されないことになる。

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。特に、コンピュータビジョンの手法によって話者の口形状が精度よく追跡できることを前提として、「口元の動きから音声を生成する機構」を積極的に利用する。具体的には、話者をカメラで撮影して得られた口元の動きから、それに合った発話音声の候補を複数生成し、実際に観測された音響信号との整合性を評価することで、雑音と音声とを分離する。

ただし、これには口元の動きから精細な音声を生成できる機構が必要であり、１で述べたように、口元の動きと音声との詳細な対応関係が求められるため、従来は困難であった。そこで本研究では、(1)ハイブリッドシステムと呼ばれる数理モデルを用いることで、口元の動きと音声のそれぞれを、力学系の切り替わりにより生じるとしてモデル化・分節化し、(2)分節化された口元の動きと音声に基づき、両信号間の詳細な時間同期構造（完全な同期だけでなく系統的なずれも含めるため本研究ではタイミング構造と呼ぶ）を統計的に学習する。こうして、口元の映像からそれに対応する音声候補生成を行うことで、非定常環境下での高精度な発話分離・音声推定の実現を目指す。

なお、口元の動きや音声、さらにそれらの間の対応関係は個人差が大きいため、本研究では、個々の利用者のデータからあらかじめこれらのモデルが学習できるような状況（特定話者の状況）を前提とする。推定された音声特徴系列は、音声認識エンジンの入力として用いることができる。さらに、本研究の基本アイデアは、複雑な環境下での様々なマルチモーダル音声解析の基盤技術として広く利用することが可能である。

３．研究の方法

研究目的で述べたような、口元の動きからの音声推定手法の流れを図１にまとめる。まず、カメラおよびマイクロフォンで取得した映像および音響信号より、口元の動きと音声の特徴系列を得る。このとき、音声には雑音が重畳されているが、口元の動き情報は高精度に取得できると仮定する。次に、口元の動きから、それに対応する音声特徴系列の候補を複数生成する。そして、生成された候補系列と、実際にマイクロフォンで観測された（雑音と音声が重畳されている）音響信号との整合性を信号（音響特徴量）レベルで評価することで、雑音が重畳する前のクリーン音声（本来の発話音声）を推定する。このような枠組みを構築するために、以下で述べる３つの観点から研究を展開した。

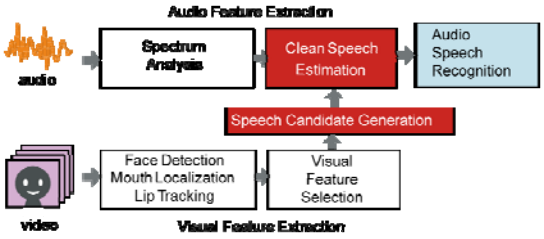

図 1　視覚情報からのクリーン音声推定

(1)高精度な口唇画像特徴抽出：　音声を推定するのに十分な情報を含むような口元の動きをとらえようとすると、顔向きやまばたき、表情の変化（顔に生ずるしわなど）といった、姿勢変化や様々な非剛体変形の影響を

受けずに、唇を含む口元領域を高精度に追跡し、映像として取得する必要がある。本研究ではこれをコンピュータビジョンの分野で提案されている非剛体変形のモデル化手法を応用することで実現する。

(2)口元の動きと対応する音声候補の生成アルゴリズムの構築： 音声推定時には、カメラより得られた口元の動きを、学習されたモデルとともに用いることで、その口元の動きと対応するような音声候補系列を複数生成する。このようなモデルおよび音声候補系列生成アルゴリズムを考案する。

(3)非定常雑音環境下での音声推定手法の開発： 雑音の重畳した音声と、(2)によって生成された音声候補系列を用いて、本来のクリーン音声を推定する手法を開発する。特に、非定常性の雑音はそのスペクトル分布が時間的に変化する。そこで、視覚的な情報から生成された音声候補を手掛かりとして使いながら、雑音の時間変動に対応した音声推定を行う。

図１に示すような、非定常雑音下での音声推定手法の枠組みは、本研究で新たに提案するものであり、特に(2)の音声候補生成手法は、従来の音声解析には無い独自性の高い手法である。

４．研究成果

(1)高精度な口唇画像特徴抽出

非剛体変形のモデルとして、コンピュータビジョンの分野で提案されている Active Appearance Model (AAM)を用いることで、口元特徴点の追跡を行った。あらかじめ話者の発話時の顔映像データを獲得し、この顔映像のいくつかのフレームにおいて、話者の顔特徴点を抽出しておくことで、それら特徴点と顔画像(appearance)とから AAM を学習する。このとき、特徴点と appearance の両者の共起関係が、線形な部分空間として構築されるため、新たな顔画像が撮影されれば、それに対応する顔特徴点を推定することが可能となる。本研究では、AAM によって口元（口唇輪郭）の特徴点を追跡することで、口元の動画像を矩形領域として切り出し、正規化、周辺のマスキング、次元削減といった処理を行うことで、後段の音声推定に利用可能な画像特徴系列を抽出できることを確認した（図 2 に切りだされた矩形領域画像の例を示す）。

図 2　AAM による特徴点追跡と口元画像の切りだし例

(2) 口元の動きと対応する音声候補の生成アルゴリズムの構築

①手法の特徴

口元の動き（以下では口唇運動と呼ぶ）の特徴系列から、それと対応する音声特徴系列の候補を生成するためには、特に、

(a) 一つの口唇運動には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

(b) 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

(c) 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、ハイブリッドダイナミカルシステム（hybrid dynamical system, HDS, もしくはハイブリッドシステム)と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、それぞれ複数の力学系（ここでは線形システムを用いる）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関係を用いて、両信号間における変化パターン間の、時間ずれを伴うような共起性をモデル化する。

再現性のある要素的な動き（プリミティブ）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、(a)の生成候補数の爆発に対処することが可能となる。一方、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、(b)が解決されることも期待できる。さらに、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(c)の問題に対処できる。以下では、学習フェーズ（図 3 上段）と候補生成フェーズ（図 3 下段）に分けて、これらの基本的な考え方をそれぞれ述べ、その成果をまとめる。

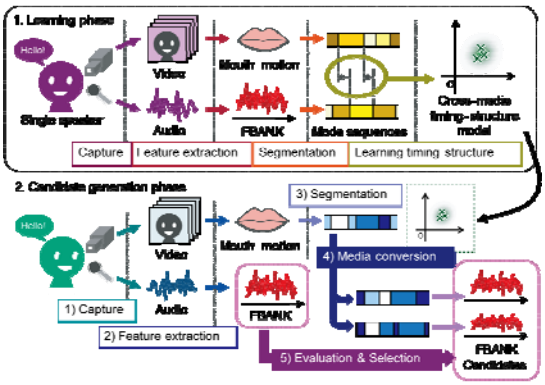

図 3 画像からの音声候補生成の流れ（上段：学習フェーズ、下段：候補生成フェーズ）

②学習フェーズ

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する。この段階で学習するモデルは、(i)音声と口唇運動それぞれの「HDS モデル」、および(ii)両メディア信号間の「タイミング構造モデル」である。まずそれぞれの特徴系列から HDS モデルの学習と系列の分節化を行う。そうして得られた区間系列対を学習データとして、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。これにより、口唇運動と音声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することが可能となる。すなわち、子音の破裂音や母音の/i/では口唇運動と発話音が同期しやすく、母音の/a/や/o/では時間的にずれることがある、といった現象をモデルとして獲得できる（①で述べた(c)の特徴)。さらに、学習された HDS を用いることで、新たに撮影された口唇運動の分節化や、分節化された区間系列表現からの滑らかな信号生成（①(a)の特徴）が可能となる。

③候補生成フェーズ

候補生成時には、新たに観測された映像・音響信号から特徴系列を抽出し、そのうち口唇運動の特徴系列に対し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、その口唇運動の区間系列に合致すると推定される音声の区間系列の候補を生成する。つまり、口唇運動からそれに対応する音声候補を直接生成するのではなく、

口唇運動の特徴系列

→口唇運動の分節化（区間系列）表現

→音声の分節化（区間系列）表現

→音声特徴系列

のように、いったん分節化された段階を経ることで、学習された時間同期構造を満足するような信号を生成できるだけでなく、生成候補の数を大幅に抑えることができる（①(b)の特徴)。なお、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、区間系列を複数候補生成できるように、Parallel List Viterbi アルゴリズムと呼ばれる方法を応用した。

候補として生成された複数の区間系列のそれぞれに対し、音声から学習した HDS を用いることで、最終的に音声特徴系列候補を複数生成することができる。図 4 は、母音のみの発話を用いて音声候補生成を行った結果であり、上段は口唇運動と本来対応していたクリーン音声、中段および下段は、生成された候補の一例である。中段に示したものは、口唇運動と本来対応していたクリーン音声（上段）と非常に近いことが分かる。このように、生成された候補の一部は本来の音声をうまく推定できることが確認された。

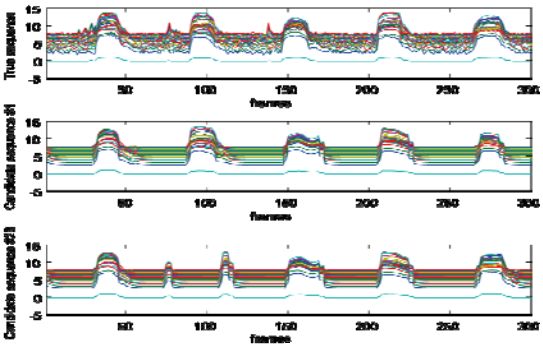

図 4 口唇運動と本来対応する音声の特徴系列（上段）および生成された音声特徴系列の例（中、下段）

(3) 非定常雑音環境下での音声推定手法の開発

生成した特徴系列候補と、観測された音響信号（雑音と音声が重畳した信号）の特徴系列とからクリーン音声を推定するには、生成された候補のうち、音響信号と整合のとれるものを見つける必要がある。調査の結果、これには従来手法である、パーティクルフィルタに基づく音声の雑音抑圧手法を応用できることが分かった。この手法は、音響信号のみを用いて雑音抑制を行う手法であり、具体的には、加法性非定常雑音を仮定し、雑音の時間変化をランダムウォークモデルで表現しておく。すると、雑音を状態と考えれば、これはパーティクルフィルタが前提とする状態空間モデルとなる。つまり状態方程式がランダムウォーク過程、観測方程式が雑音とクリーン音声との非線形演算（非線形性は対数スペクトルの特徴量を考えることから）となる。その結果、雑音の変動をパーティクルフィルタで追跡できれば、元の音響信号と推定された雑音の残差としてクリーン音声が得られるという仕組みである。

本研究では、視覚情報から得られた音声候補を用いて、この手法を拡張した。従来手法は、クリーン音声として何らかの事前分布を仮定する必要があり、通常は固定的な混合ガウス分布が用いられる。しかしながら、SN 比が小さくなるほど、より正確な事前分布が必要となる。そこで、本研究の提案手法では、視覚情報から生成された音声候補を利用して、時間的に変化できるようなクリーン音声の動的な事前分布を構成して、パーティクルフィルタの観測方程式に入力として与える。その結果、図 5 の赤線に示すように高精度なクリーン音声推定が可能となった（縦軸は正解音声からの誤差)。これは、音声情報のみを用いる手法による結果（図では紫線および青線で表示）に対してはもちろんのこと、回帰モデルを用いて口唇運動の特徴系列から音声特徴系列を推定した場合の結果（図中の緑線）に比べても大幅に精度が向上している。すなわち、HDS モデルやタイミング構造モデ

ルを用いることで、それぞれの信号やその対応関係を詳細に表現でき、より本来のクリーン音声に近い特徴系列を推定することができるといえる。

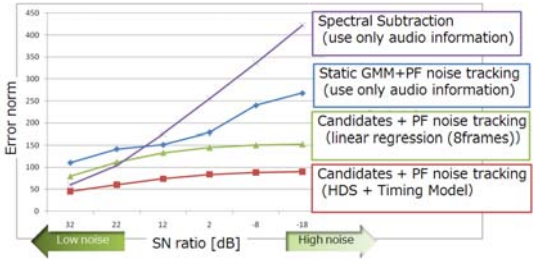

図 5 クリーン音声推定の比較結果

２の研究目的で述べたように、本研究の基本アイデアは、複雑な環境下での様々なマルチモーダル音声解析へ応用することが可能であり、今後研究展開を行う予定である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計３件）
① H. Kawashima、Interval-based Modeling of Human Communication Dynamics via Hybrid Dynamical Systems、Workshop on Human Communication Dynamics (NIPS WS)、2010.12.10、Whistler Canada
② H. Kawashima、Speech Estimation in Non-Stationary Noise Environments Using Timing Structures Between Mouth Movements and Sound Signals、Interspeech2010、2010.9.27、Makuhari Japan
③川嶋宏彰、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、第 13 回画像の認識・理解シンポジウム (MIRU)、2010.7.29、北海道（釧路）
〔その他〕
ホームページ
http://vision.kuee.kyoto-u.ac.jp/~hiroaki/research/

６．研究組織
(1)研究代表者
川嶋　宏彰（KAWASHIMA HIROAKI）
京都大学・情報学研究科・講師
研究者番号：４０３４６１０１

(2)研究分担者：なし

(3)連携研究者：なし

(4)研究協力者：なし