マルチモーダルタスク推定のための
調音運動 HMM 音声認識合成方式

2015 年
博士（工学）

木村　優志
豊橋技術科学大学

目次

図目次

表目次

1章　序　論

1.1 研究の背景

近年の技術発展に伴い, 音声認識合成技術はカーナビゲーションシステムやスマートフォンの音声エージェントなどに利用されるようになってきた. 将来は人間が遂行するタスクを支援するようなロボットにも音声対話技術が利用されるようになるであろう. その際には音声認識合成，及び，対話技術のみでなく，周囲環境から対話相手の状態や取り組んでいるタスク，更には，行動を推定する技術が必要になると考えられる．本研究ではこうした応用が現実となる状況に備えるべく，特に音声合成，及び，タスク推定技術に焦点を当て，それらの性能向上に取り組む.

隠れマルコフモデル(Hidden Markov Model: HMM)に基づく音声認識は，近年，幾つかの分野で成功を収めたが，多くが音声スペクトル由来の特徴を使用するため，話者，音素コンテキスト，ノイズ重畳による変動を抱える．そのため，モデル近似に多くのデータと混合分布を要するという欠点を持つ．他方，乳幼児は，身辺の僅かな人たちから音声を模倣するだけで音声言語を獲得している [1]. 親の音声と乳幼児の模倣音声では，声道形状等の違いから物理的には異なる音響信号になる．このため，音韻知識が未確立の乳幼児がこれら二つの音声を同一と判断するには, 音韻に関係しない何らかの不変量が内在することになる．この説明として，音声知覚が音響信号そのものではなく，調音器官に送られる運動指令を参照して行なわれるとし, 音声の知覚と生成が一つのシステムで構成されているという見解が古くから提唱されてきた [2]．人間の音声生成と音声知覚が 1-model か 2-models かは，長年論争され未だ決着がついていないが[3]，近年の脳研究は 1-model 説を支持する結果を示しつつある[4]．人間が備える仕組みをロボットにも備えることで, 人間と同様の学習過程で言語を習得させることが可能になることから, ロボットによる音声言語の獲得の際にも, 人間同様に知覚と生成の 1-model が望ましいモデルと考えられる．著者はこの見解を基に，音声認識と音声合成の双方に，共通の音響モデル(Acoustic Model: AM) を使用する「ワンモデル音声認識合成方式」を開発している [3, 4, 5]．また，ワンモデル音声認識合成器により，自身で聴取した音声を話者不変の特徴量に変換し，同じモデルから発話者の同音声を合成し，スペクトル等を比較することにより，音声認識誤りの訂正が行えるようになると考えられる．音声認識と音声合成で単一モデルを利用しようとする試みは, 古く 1970 年代の初めに販売された Threshold Technology 社の音声認識装置に見られたが（当時の装置技術資料による），近年に至って数多くの方式が提案されるようになり [6, 7, 8, 9, 10] ，音素認識については多数話者音声で学習した標準的 MFCC ベース HMM を上まわる性能も得られるようになっている.

認識と合成の双方を一つのモデルで実現するには, 音声から話者に依存しない特徴を抽出して HMM を構成すると共に，合成の際は HMM が生成する特徴系列に話者性を与える処理が必要になる．我々の研究グループでは，先に多数話者データからニューラルネットを学習することで，音声から話者不変な調音特徴を抽出し，調音特徴の時間変化を表現する AM を構成することで，高精度不特定話者音声認識が 1 名の話者データで達成可能なことを示した[10]．音声合成では，同じ AM から調音特徴系列を生成し，

これを声道音響パラメータに変換（以後 VT(Vocal Tract) 変換と呼ぶ）してデジタルフィルタの係数とすると共に，別途，音声の残差信号から設計した音源符号帳でフィルタを駆動し音声を合成する．この方式は，調音様式や調音部位を表す調音特徴(Articulatory Feature: AF) と発声システムを分離できるため，数文の音声試料を適応学習するだけで特定話者の音声を合成できる[11].

本論文では，音声認識のための調音運動モデルを HMM で実現し，同じモデルを使用して音声を合成する方式を提案する．これまでに提案された標準的 HMM 音声合成 [13] は，スペクトル由来の特徴を使用するため特定話者の多量の音声を必要とする．また，特定話者モデルであるが故に不特定話者音声を認識することができなかった．提案方式は，話者共通の調音運動を HMM で表現すると同時に，HMM から得られる調音特徴系列を，多層ニューラルネット(Multi-Layer Neural Network: MLN) を用いて声道音響パラメータである線スペクトル対(Line Spectrum Pair: LSP) [14, 15] に変換した後，LSP 合成フィルタにより音声を合成する．この合成方式は，音声合成と認識を一つのモデルで実現できることに加え，少量データで音声を合成できることから，工学的にも有用性が高いと考えられる．

一方，ロボット研究，脳研究，及び，音声言語を含むマルチモーダル対話研究の発展に伴い，ロボットが人と共生することを最終目標とした研究が盛んに行われ始めている [16, 17, 18, 19, 20]. 将来は身近な場面で，家事ロボット，ナビゲーションロボット，秘書エージェントなどのパーソナルエージェントが実世界の事物を対象に人とコミュニケーションする時代が到来すると予想される．

ロボットが人間と共生していくには，様々な課題を解決する必要があるが，その中でも人間との共同作業課題 (タスク) を耳と眼で的確に判断する機能が重要になる．工場などの限定された環境ではロボットに信号を送ることで，従事すべきタスクを指示することができる．しかし，家庭内でロボットを利用するシーンでは，従事すべき多種多様なタスクについて，その都度指令を与える必要があるため，このようなアプローチは現実的でない．こうした背景から，人間(達)が従事するタスクを音声対話や映像を通してロボットに判断させる研究は大変重要であると考えられるが，この分野の研究はこれまであまり実施されてこなかった．

関連する研究としては，映像解析に関してビデオシーンから特定のシーンを検出する研究 [21, 22, 23]，特にサッカーのゴールシーン等を検出するなどの研究 [24, 25, 26, 27, 28] が行われている．しかし，これらは人間の対話を対象とする研究ではない．人間が行う対話の認識に関しては Alatan らが，映像から対話シーンを検出する研究を行なっている[21]．この手法では，映像から場所の切り替わりや顔が写っているかを認識し，DP マッチングにより対話シーンを検出している．しかし，こうしたアプローチでは他のタスクへ応用できないという問題がある．

そこで本論文では，ロボットやエージェントに人間が従事するタスクを認識させることのできる，より一般性のある方法を提案する．提案方法は，最初に，タスクに従事中現れる画像オブジェクトと発話単語という二つのメディアの出現頻度をベクトル空間上に表現し，潜在意味解析(Latent Semantic Analysis; LSA)を適用してベクトルを圧縮した特徴空間上に表現する．続いて，タスクが未知の映像と発話からなる入力ベクトルを特徴空間上に射影し，タスク毎に予め学習した参照ベクトルとの類似度を計算することでタスクを推定する．本手法は，人間の動作や意図に関する推定を行う必要がないた

め簡易にタスクを推定することができる.

1.2 論文の構成

図 1 ，及び，以下に本論文の構成を示す．本論文では，2 章では，音声認識と画像認識を含むマルチモーダル情報に基づくタスク推定に関して説明する．ここでは潜在意味解析を用いたタスク推定法を提案すると共に評価実験からその有効性を示す．続いて 3 章と 4 章では，調音特徴を用いた音声合成システムの改良について述べる．我々の音声合成手法は声帯振動と声道形状のモデルから音声を合成するボコーダ方式を採用しており，音声の合成には声道音響パラメータと駆動音源（声帯振動）を必要とする．3 章では，声道音響パラメータを 3 段のニューラルネットワークで抽出する手法を用いる．提案手法では，ニューラルネットワークを隠れマルコフモデルからの出力（調音特徴）で学習することによって音質向上が実現できることを示す．4 章のデジタルフィルタ駆動音源の改良では CELP(Code Exited Linear Prediction)方式を駆動音源の選択に利用することによって，主観評価値が上昇することを示す．その際，駆動音源を LBG (Linde-Buzo-Gray) 法によってクラスタリングを行い，もっとも近い残差素片を駆動音源として選択する． 5 章は本論文の結論と今後の課題である．

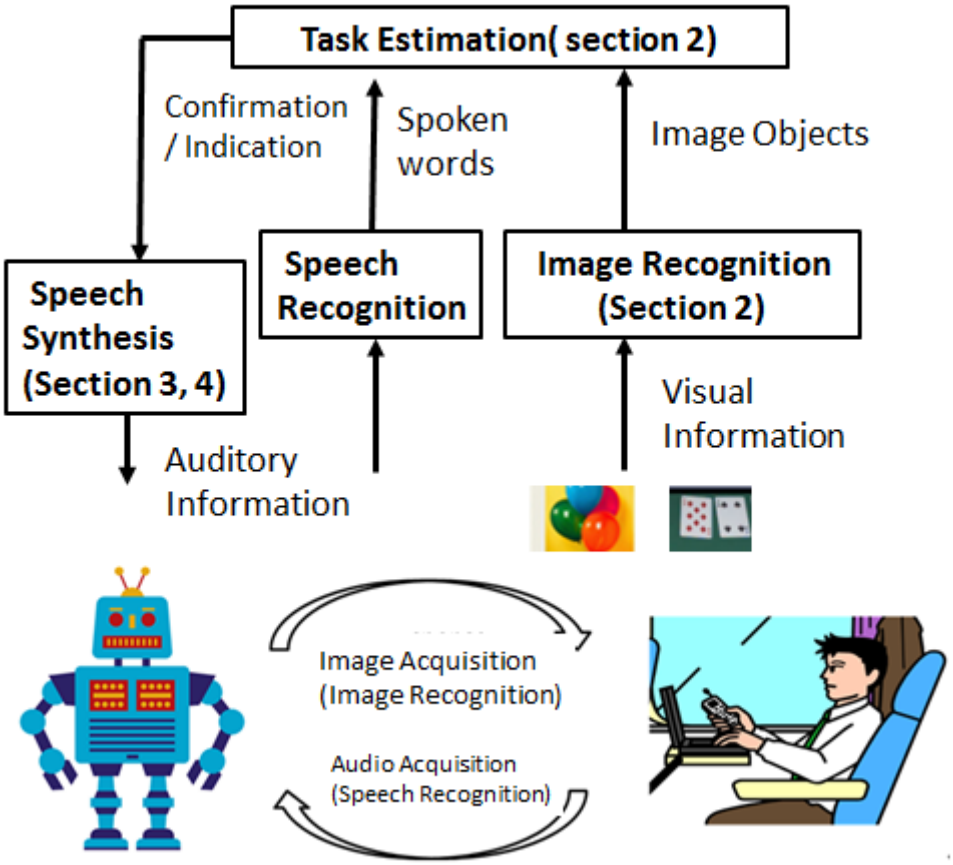


図 1 提案するマルチモーダル対話システムと論文の構成

2 章　マルチモーダル情報を用いたタスク推定

本章では，マルチモーダル音声対話システムのための，タスク推定手法について説明する．ここではタスクを人々が行う作業のことと定義する．近年，人間と共生するロボットを目指す研究開発が進められている．こうした近未来の知的ロボット（以後，知的エージェントと呼ぶ）では，遭遇する状況に即して行動を選択し，人間活動を支援する能力が要請される．知的エージェント達は，視覚と聴覚のマルチモーダル情報から状況を解釈し，必要とされるタスクを瞬時に判断し，場合によっては確認対話の後，適切な行動を起こさなければならない．例えば，視覚情報から「赤くて丸いオブジェクト」が推定されたとしよう．解釈は状況に依り異なる．それが交差点であれば，道路を渡ってはいけないことを意味し，行動「止まれ」を選択するだろう．工場内のように限定された環境であれば，予め決められた信号を送ることでロボットに人間を支援する行動をさせることが可能であるが，実環境下で人間を支援するには視覚と聴覚情報からタスクを推定することが好ましい．このように，人間と共生する未来のロボットには，情景と発話に即して行動する能力が要請される．

本論文では，情景と発話のマルチモーダル情報に潜在意味解析（Latent Semantic Analysis; LSA）を適用し，タスクを推定する手法を提案する． LSA は，これまでテキスト解析の分野で文書分類に大きな成功を収めてきた [29, 30, 31, 32, 33]．近年では，LSA をテキスト情報だけでなく，画素ヒストグラムのような画像特徴に適用して，画像分類を行う手法が提案されている [34, 35, 36]．本論文では，情景と発話内容という異なるメディア（マルチモーダル情報）を，タスクを表現する行列上に配置し，LSA によりタスクを推定することを目指す．

我々は，何らかのタスクを遂行する際，情景からタスク推定に必要な画像オブジェクト（以後，オブジェクトと呼ぶ）を確認すると同時に，発話からも関連情報を聞き取り推定に役立てている．例えば，丸いボールがプレーヤの間を飛び交い，「パス」，「シュート」といった発話が聴こえることで，我々はサッカー競技というタスクを推定する．

本論文ではタスク遂行中のオブジェクトと発話内容の二つから，人間が遂行しているタスクを推定する手法を提案し，その評価実験結果について考察する．図 2 に，ゲームタスクを対象とする，マルチモーダル情報に基づくタスク推定システムについて示す．まずタスク遂行中の画像からオブジェクトと発話単語を抽出し，これを同じベクトル上に表現する．このベクトルを用いてタスクの推定を行えば，これを元にマルチモーダル対話システムで発話する内容を決定することができる．本論文ではタスク推定の部分について研究を行った．

2.1 先行研究と本研究の優位性

これまでに，人間の行う動作や機械の組み立て工程を画像から認識する手法が提案されている [22, 23]．これらの手法では，オブジェクトの結合状態の推移や，手とオブジェクトの位置関係を観測することで，動作を認識している．しかし，オブジェクトの結合状態などを認識するには，複雑な画像処理や，特殊な器具が必要となる．これに対して，提案するタスク遂行中のオブジェクトと発話からタスクを推定する手法は，ユーザの身体動作や対象を認識することなく，遂行中のタスクを推定することが可能である．また，人間が行うタスクの推定を LSA を用いて行う手法はこれまで提案されていない．

提案手法は，タスク遂行中にユーザが扱うオブジェクトの出現頻度と，発話中の単語頻度を取得し，これら二つを要素とする列ベクトル(以後，タスクベクトルと呼ぶ)をタ

スク毎に構成した行列としてタスクを表現する．以後は，これをタスク行列と呼ぶ．続いて，マルチタスクを表現するベクトルが張る空間に対して LSA を適用することにより，潜在意味空間としてのタスク固有の部分空間（以後，タスク固有部分空間と呼ぶ）を抽出することを試みる．タスク推定の際には，未知のタスクベクトル入力と，全てのタスク固有部分空間との間で類似度を計算し，最大値を与えるものを推定タスクと判定することになる．

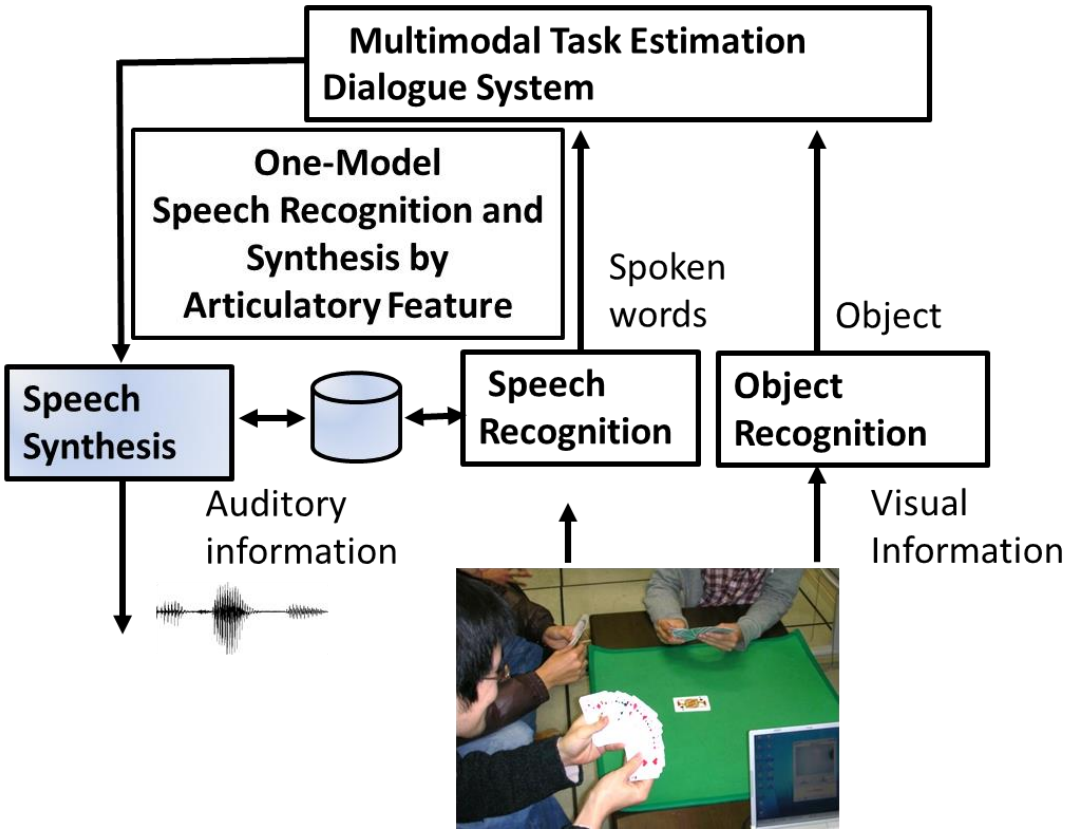


図 2 ゲームタスク推定を対象とする MMI システムとワンモデル音声認識合成

2.2 タスクのベクトル空間表現

以下では，提案する LSA を用いたタスク推定手法について説明する．提案手法の概要を図 3 に示す.

我々は，他人が行う共同作業を見たり，会話を聞いたりする中で，彼らがどのようなタスクを遂行しているのかを理解する．共同作業中は，タスクに関連する話をしたり，タスクに関連する物体を操作したりすることが多い．そこで，以下ではタスク遂行中の発話と，出現するオブジェクトを用いてタスクを表現し，同定することを検討する．なお，本論文では既知のタスク分類を対象とする.

文書分類などのテキスト解析研究では, 文書をベクトル空間上に表現する手法が用いられてきた．これらの手法では，単語出現頻度を要素とするベクトルとして文書を表現する．提案手法では，タスク遂行中の発話に含まれる単語と共に，作業中に出現するオブジェクトとその数を要素にすることで，タスクをベクトル空間上に表現する．以下，発話中の単語とオブジェクトをあわせてタームと呼ぶ．タスクベクトル t^k は，式(1)のように表される.

$$t^k = \left(w_1\ w_2 \cdots w_{\mathrm{I}}\ o_1\ \mathrm{o}_2 \cdots \mathrm{o}_{\mathrm{J}}\right)^{\mathrm{T}} \quad \text{.....(1)}$$

ここで, w_{i} はコーパス k 中 i 番目(i=1, 2,..., I)の単語の正規化頻度, o_{j} は j 番目(j=1, 2, …, J) のオブジェクトの正規化頻度である． 両頻度は大きく値が異なるため，これを正規化することによってバランスをとる.

このように本手法ではタスクをベクトル空間上に表現することを特徴に持つ. ベクトルの要素として表現しづらい特徴を扱うことは困難である．例えば，タスク遂行中のユーザの意図，動作等はベクトルの要素として表現しづらい．また，タスクの推定は時間に依存しないため自己回帰モデルのような直前の時間の要素との関連性が重要であるような特徴はタスクベクトル上でうまく表現できない.

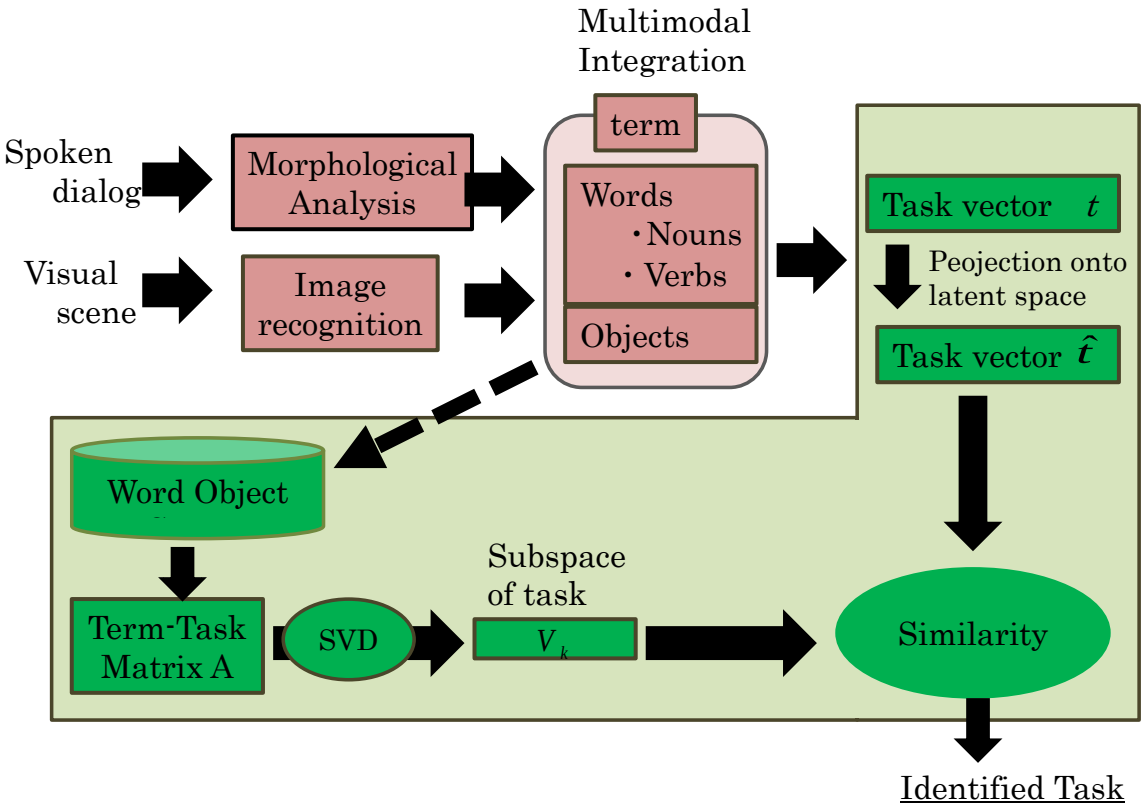


図 3 タスク推定のシステムフロー

2.3 発話単語とオブジェクトの抽出

オブジェクトの頻度 o_j を得るには，タスク遂行中の画像からオブジェクトを認識し，オブジェクトの種類毎に出現数を数える必要がある．オブジェクト認識は以下の手順で行う．図 4 にオブジェクト認識の流れを示す．

1. オブジェクトの領域を抽出する．
2. オブジェクト領域から画像を抽出し，形状特徴を算出する．
3. オブジェクトの各平均パターンと，上記オブジェクトの形状特徴のマハラノビス距離を算出する．
4. 距離値からオブジェクトの種類を判別する．
5. オブジェクトを種類毎に集計する．

まず，オブジェクトが存在する領域を抽出するため，画像から背景領域を除去する．背景領域の除去には，Morita らの背景差分に基づく注目物体の検出手法[36] を採用した．ただし，本論文では背景統計量の更新は行っていない．

Morita らは，カメラのゲインアップによるゴマ塩ノイズ，蛍光灯などのフリッカーによる明度変化を考慮し，注目物体を検出する．まず，入力画像中の各背景画素の輝度値を以下のように定式化する．

$$I = \bar{I} + \sigma \sin(2\pi f t) + k\zeta \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

ここで，$\bar{I}$ は輝度値の時間平均，σ は輝度の振幅，f は輝度の周波数，t は時間，k は -1〜1の値を取る係数，ζはカメラや人の影の映り込みに依存したノイズの最大値である．上式で$\sigma \sin(2\pi ft)$ の項は蛍光灯のフリッカー等の影響を，また $k\zeta$の項はカメラなどに依存するノイズを表す．また，輝度値 I が式(3)で表す範囲に収まっていれば背景の画素，そうでなければオブジェクトの画素とする．

$$\bar{I} - \sigma - \zeta \leq I \leq \bar{I} + \sigma + \zeta \quad \cdots\cdots\cdots\cdots\cdots\cdots (3)$$

Morita らは，このあと各フレームに対し，$\bar{I}$と σの更新を行なっている．しかし，我々の実験では事前に背景動画を撮影して $\bar{I}$とσ，ζを決定する．その後は全てのフレームについて同じ値を用いている．

得られたオブジェクト画像と背景画像を二値化しオブジェクトの形状画像を得る．その後，得られた形状画像に対して膨張・縮小処理を行い，ノイズを除去している．

次に，オブジェクトの画像特徴として次の五つを抽出する．

- オブジェクト領域の周辺長: l
- オブジェクト領域の面積: s
- 曲率係数の絶対値の平均: c
- バウンディングボックスの 2 辺の長さ: 長辺 e_1，短辺 e_2
- 凸包の面積 : h

ここで，バウンディングボックスとはオブジェクト領域を囲う最少の矩形であり，凸包はオブジェクト領域を囲う最小の凸図形である．

凸包を求めるには，Sklansky のアルゴリズムを用いる[37]．図 5 に Sklansky のアルゴリズムを示す．このアルゴリズムでは 2 段階の処理で凸包を決定する．

はじめのステップでは，上，下，左，右の最大の頂点を見つけ，それぞれ，T, B, R, L とする．次に，L から始まり半時計回りに辺を探索して，凸包の頂点に保持するか，あるいは棄却するかを決定する．決定のアルゴリズムを以下に示す．まず，頂点 L を i，頂点 L の反時計回りの隣の点を i+1，さらにその隣を i+2 とする．ここで以下のアルゴリズムにより，i+2 を保持するか棄却するかを決める．なお，頂点で囲まれた外部の領域 R_j の j は 1 から始まり，頂点 i が B,R,T を周る毎に 1 ずつ増えるものとする．また A1, A2, A3 は，図 5 に示すようにそれぞれ、着目頂点の左下の領域、右下の領域、上の領域を指す．

(1) If $i+2 \notin R_j$, i+2 を棄却する．
(2) Else if $i+2 \in A3$, i+2 を保持する．
(3) Else if (i+2 ∈ A2) OR (i+2∈A1 AND i+2 is above i+1 AND a vertex was discarded due to line 3 on the immediately preceding iteration), i+2 を棄却する．
(4) Else if $i+2 \in A1$, i+1 を棄却する．

もし(3)を実行していた場合は，i+2 を i+1 とし，i+1 を i とする．
それ以外の場合は i を一つ進め， (1) に戻る．

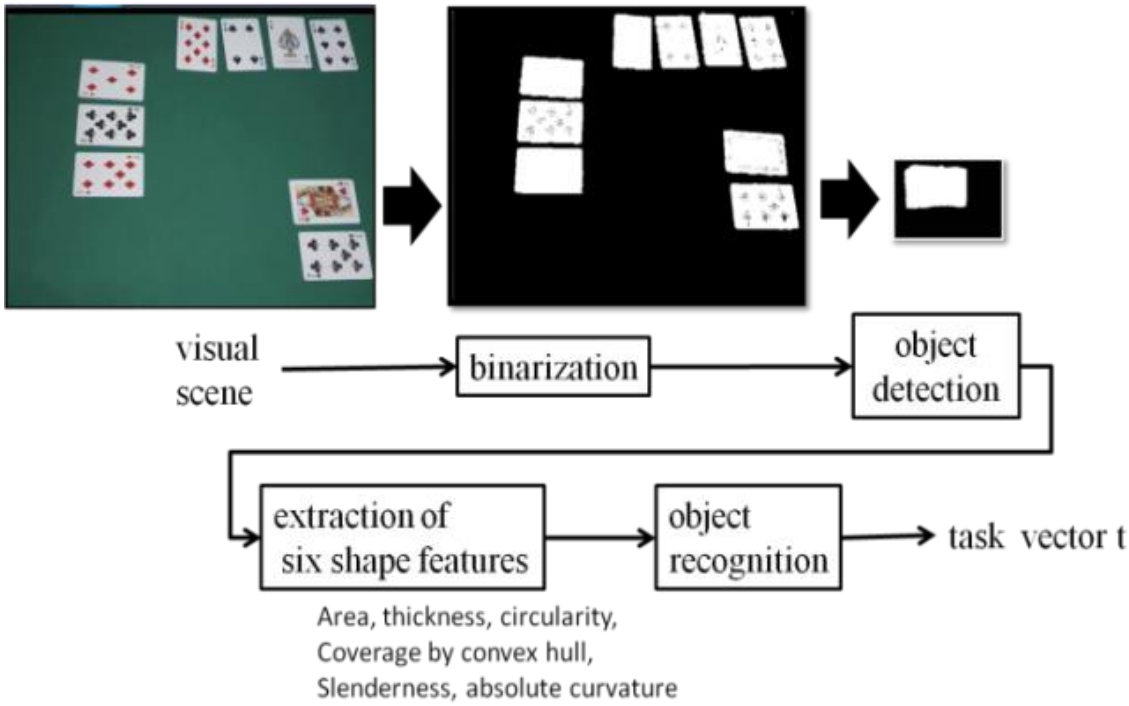
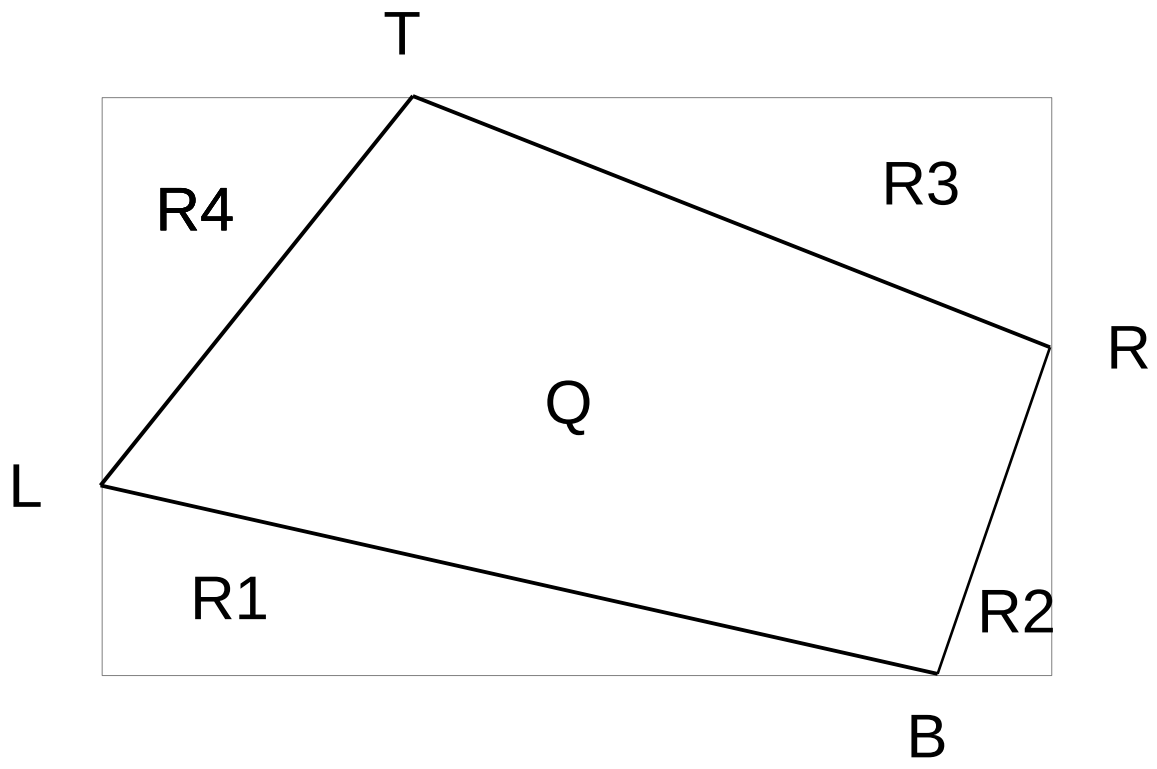
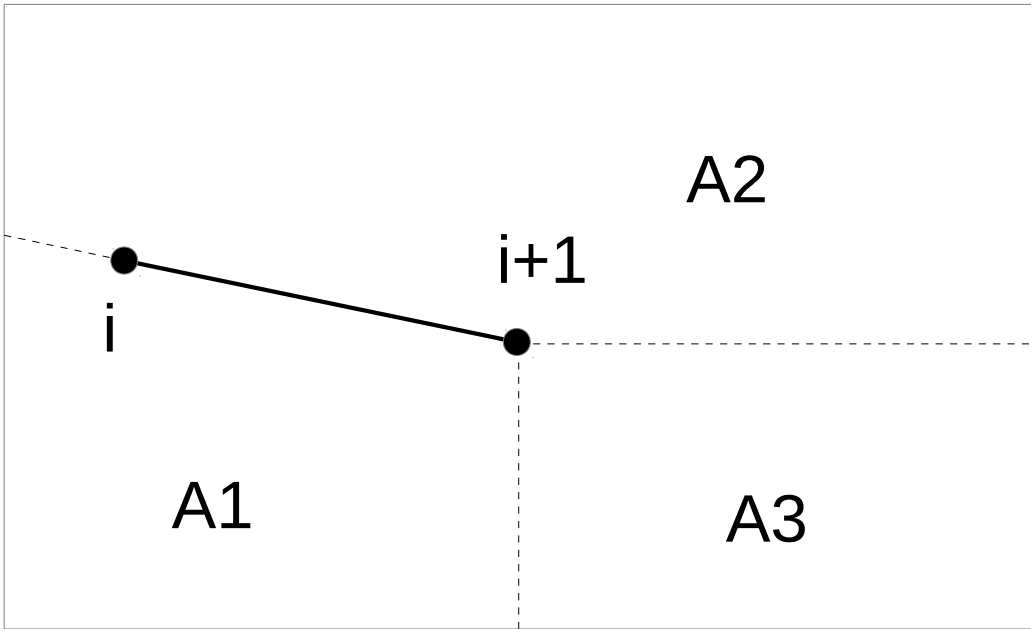


図 4 オブジェクト認識の流れ



Step 1. 最大頂点に基づく分割
T,B,L,R はそれぞれ上，下，左，右の最大の頂点



Step 2. 頂点 i+2 の保持・棄却，R_1 の場合
R_2,R_3,R_4 の場合は線分 BR, RT, TL の方向に合わせて回転させて考える

図 5 Sklansky のアルゴリズム:　凸包の探索

曲率係数とは局所的な特徴であり，各画素の 8 近傍，計 9 画素の様相に基づく特徴である．曲率係数は式(4) で求める[38].

$$1-\frac{1}{2}\sum_{k\in s}x_k+\frac{1}{4}\sum_{k\in s}x_k x_{k+1}x_{k+2} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4)$$

次に，形状特徴として以下の六つを計算する[38]．図 6 でこれらの特徴を説明する．

- 面積： s
- 平均太さ: $2s/l$
- 円形度 ： $4\pi s/l^2$
- 凸包の面積に占める割合: s/h
- 細長さ: e_1/e_2
- 曲率係数の絶対値の平均: c

以上の特徴を使用して認識対象オブジェクトと，各 10 個の学習パターンの平均形状とのマハラノビス距離を計算する．距離が最も近いオブジェクトを認識結果とする．これを平均パターン毎に数えたものを 2.4 節で示すタスク行列のオブジェクト要素として用いる．

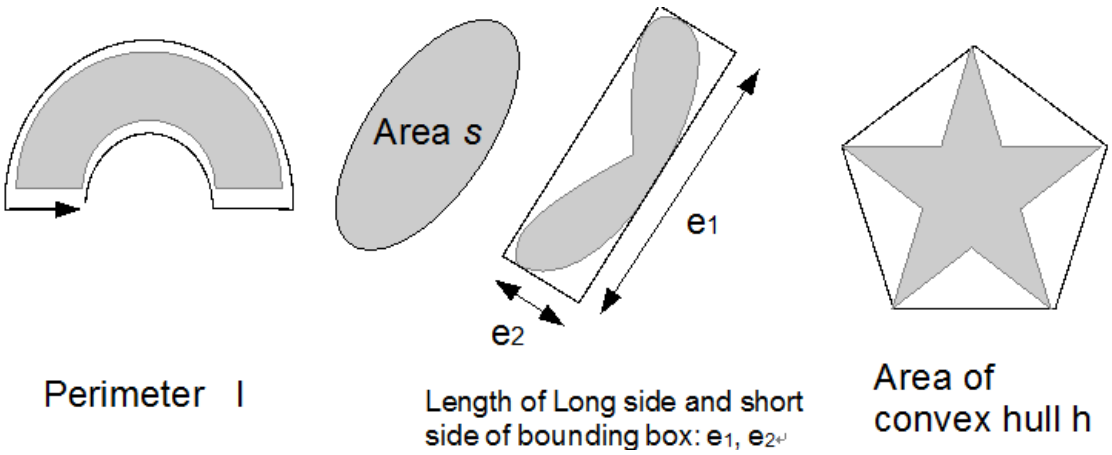


図 6 オブジェクトの画像特徴
灰色部分がオブジェクト領域

2.4 潜在意味解析

本論文では発話文中の単語とオブジェクトをまとめてタームと呼んでいる．タスクは，タームを行，タスクベクトルを列とする，タスク毎のタームから表現される行列（タスク行列) によって表現される．タスク行列には，タスク推定に関連の無いタームが含まれていたり，同じことを表現するために違うタームが使われていたりするなどの問題がある．これらの問題を解決するために，タスク行列に対して特異値分解(SVD: Singular Value Decomposition) を行ったあと，階数の低減を行う手法が知られている[39,40].

SVD は，行列を分解する手法の一つで，例えば m×n の行列 **A** を分解すると，

$$\mathbf{A} = \mathbf{USV}^{\mathbf{T}} \cdots\cdots (5)$$

となる．ここで，**U** は m×r，**S** は r×r，**V** は n×r の行列で，r =min(m, n)である．また，$\mathbf{U}$，$\mathbf{V}^{\mathbf{T}}$ の各列ベクトルはそれぞれ左特異ベクトル，右特異ベクトルと呼ばれる．また，S は対角行列で，その対角成分を特異値という．SVD によって分解された行列から，上位 k 個の特異値とそれに対応する特異ベクトルを用いて，タスク固有部分空間 $\widetilde{\mathrm{A}}$ を得ることが出来る．

$$\mathbf{A} \cong \widetilde{\mathbf{A}} = \mathbf{U}_{\mathrm{k}}\mathbf{S}_{\mathrm{k}}\mathbf{V}_{\mathrm{k}}^{\mathrm{T}} \cdots\cdots (6)$$

ここで，$\mathbf{U}_{\mathrm{k}}$は m×k 行列，$\mathbf{S}_{\mathrm{k}}$は k×k 行列，$\mathbf{V}_{\mathrm{k}}$は n×k 行列である．SVD により**A** を分解して得た$\mathbf{U}_{\mathrm{k}}$ の行ベクトルと$\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$列ベクトルは，それぞれ **A** の行要素と列要素にする情報を含有している．$\mathbf{U}_{\mathrm{k}}$ の行ベクトル同士，もしくは$\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$の列ベクトどうしを比較することにより，行要素同士または列要素同士の関連性を調べることができる．提案手法では，タスクを推定するため，列要素の情報が含まれている式(6) の $\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$ を利用する．

対象タスクが，どのタスクかを判断するには，タスク間の類似度を定義しなければならない．本論文では，推定対象タスクと既知タスクのベクトルとのコサイン尺度を用いる．既知タスクのベクトルは，$\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$ の各列のベクトルから得ることができる． しかし，$\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$ の列ベクトルと推定対象のタスクベクトル **t** では次元数が異なるため，これを (7) 式で$\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$と同じ空間に写像する．

$$\hat{\mathbf{t}} = \mathbf{S}_{\mathrm{k}}^{-1}\mathbf{U}_{\mathrm{k}}^{\mathrm{T}}\mathbf{t} \cdots\cdots (7)$$

ここで，$\hat{\mathrm{t}}$ は写像されたタスクベクトルを表す．この変換後に，$\hat{\mathbf{t}}$ と $\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$ の列ベクトルとの類似度 Sim をコサイン尺度によって式(8)のように計算する．

$$\mathrm{Sim}(n) = \frac{(\hat{\mathbf{t}} \cdot \mathbf{v}_{\mathrm{n}})}{|\hat{\mathbf{t}}|\,|v_n|} \cdots\cdots (8)$$

ここで，v_nは $\mathrm{V}_{\mathrm{k}}^{\mathrm{T}}$の n 列目のベクトル，(・)は内積演算である．最も類似度が高くなる $\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$ の列ベクトルに対応するタスクを推定結果とする．

予め画像のみの分類をおこなってから単語で分類するような手法も考えられる．しかし，そのような場合には，画像認識誤りが発生した場合に回復する事ができない．提案手法の様に両方を同時に扱うことでそれぞれの認識誤りを補う事が可能になると考えられる．

2.5　評価実験と考察

2.5.1　実験条件

提案するタスク推定の性能を調べるために評価実験を行った. 被験者は成人男性 3 名一組の 4 グループ，計 12 名である．被験者には，タスク遂行中，自由に発話してもらった．実験対象のタスクは，机上で行うトランプを用いる三種類のタスクと将棋駒を用いる二種類のタスク，計五種類とした．このうち，トランプを用いるタスクは，「ポーカー」,「大富豪」,「ブラックジャック」，将棋駒を用いるタスクは,「まわり将棋」と「詰将棋」である．タスクベクトルのオブジェクトの正規化頻度は画像認識をおこなった結果から得た．画像オブジェクトはそれぞれ 10 枚の画像から学習した．各タスクからランダムに 100 フレームを選びオブジェクトの認識率を測ったところ 77.4%であった．認識誤りのほとんどは，2 枚のトランプが重なったために 1 枚と認識してしまうといったもので，トランプを将棋駒と取り違えるような誤りは，4.5%であった．このほとんどはトランプが画面外に見切れているために起こったものである. 今回実験タスクでは，オブジェクトを間違えることにはさほど影響はないため，許容できる誤り率といえる．また，オブジェクトに被験者の腕が重なっていると認識率が低下するため，腕が写った画像を除去した．この画像の除去には，腕の写っている画像の画素の変化量が他のフレームに比べ大きくなることを利用している．また、オブジェクトがない部分のフレームも計量から外している.

一方，発話単語の頻度は，発話音声を人手によって書き起こした発話テキストから得ている. 予備実験から，タスク行列の近似に用いる SVD の次元数を 5 とした. 評価は，各組から 1 種類のタスクを評価データとして選び, 残りをタスク行列作成に用いる一つ抜き法によって行った．実験では，タスク開始からの経過時間に対するタスク正解率と類似度の推移を調べる．タスク正解率 C は，以下の式で求める.

$$C = \frac{n}{N} \times 100\ (\%) \quad \text{..... (9)}$$

ここで，n は正解タスク数，N は全タスク数である.

図 7 に実験風景を，図 8 に実験で使用したコーパスの一例を示す. 左図は発話文のコーパスであり,「発話開始-分:秒　発話文章」の形式となっている．右図はオブジェクトの個数のコーパスであり,「タスク開始からの経過ミリ秒，トランプの個数，将棋駒の個数」という形式になっている．発話文については話者の区別を行なっていない.

なお，タスクベクトルのサイズは 1017 であった．これは，全タスクでの，ユニークな単語数+ 2 オブジェクトに相当する.

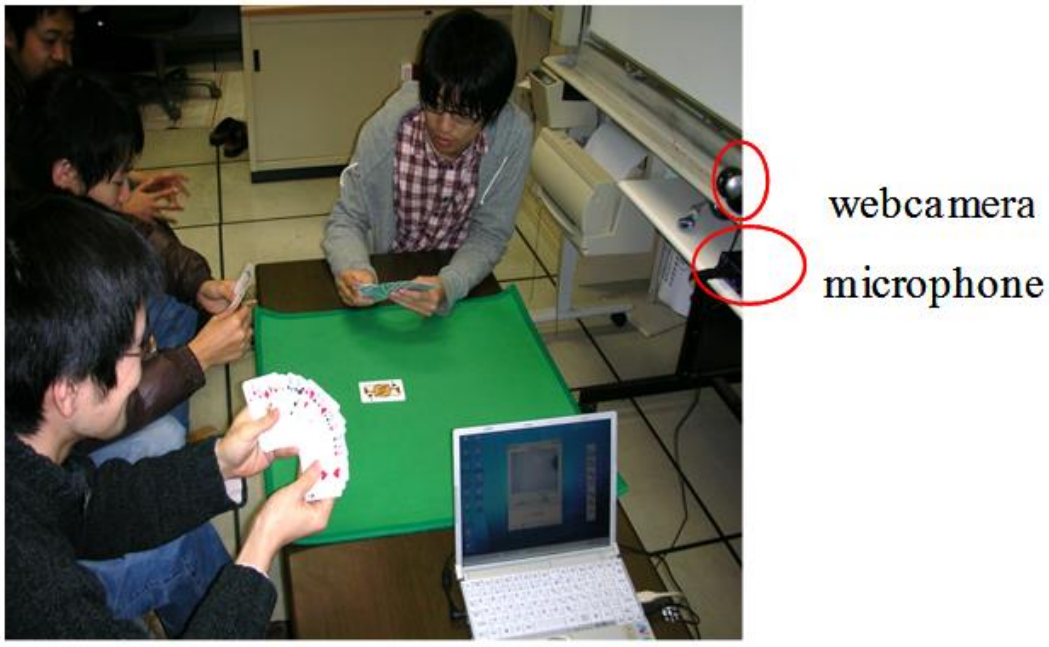


図 7 マルチモーダルコーパス収集シーン

0:01 あ、僕が配ります 0:02 そうか 0:03 じゃ、いきます 0:04 イカサマせんように 0:05 あ、はい 0:06 出来ないですけどね 0:08 そういって2枚配ったりするんです 0:10 え 0:10 セカンドディールと	,トランプ, 将棋駒, 2733, 1, 0 , 2800, 1, 0, 2933, 1, 0, 3000, 1, 0, 3133, 1, 0, 3200, 1, 0, 3333, 1, 0, 3400, 1, 0,

図 8 単語・オブジェクトコーパス

2.5.2 実験結果

図 9 は, タスク行列に単語のみを用いた場合とオブジェクトと単語を用いた場合のタスク正解率を時間進行に沿って示している．図 10 は各時刻の全てのタスクにおけるユニークな発話単語数を示す．これより，タスク遂行中の発話単語数は概ねタスク遂行時間に比例して増えると言える．表 1 にタスク行列に LSA を適した後の左特異ベクトルの要素を絶対値で昇順に並べたものを示す．各セルの上段は要素名を，下段にベクトルの要素値を表す. 要素名の先頭に``Obj_”と付いているものは画像のオブジェクトを示しており，それ以外のものは発話単語を示している．各タスクの類似度を 4 グループのデータで平均したグラフを図 11〜15 に示す．グラフは横軸がタスク開始からの経過時間（秒），縦軸が 3 グループで平均した類似度を表している．各図では，タスク行列に発話単語のみを用いた場合を (a) に，タスク行列にオブジェクトと単語を用いた場合を (b) に示している．表 2，3 にタスク行列にオブジェクトと単語を用いた場合のタスク開始から 20 秒，及び，100 秒の時点での混同行列を示す.

なお 2.5.1 に説明したように，今回は音声から書き起こしたテキストを使用したが，Julius[41] を用いたキーワードスポッティングによる発話単語抽出に基づくタスク推定時の正解率を図 22 に示す．図から，Julius を用いた場合のタスク推定率は 20%程度であることがわかる．この結果は，タスク推定における音声認識精度の重要性を示しており，現行の音声認識システムは未だ大幅な性能改良を必要としている．

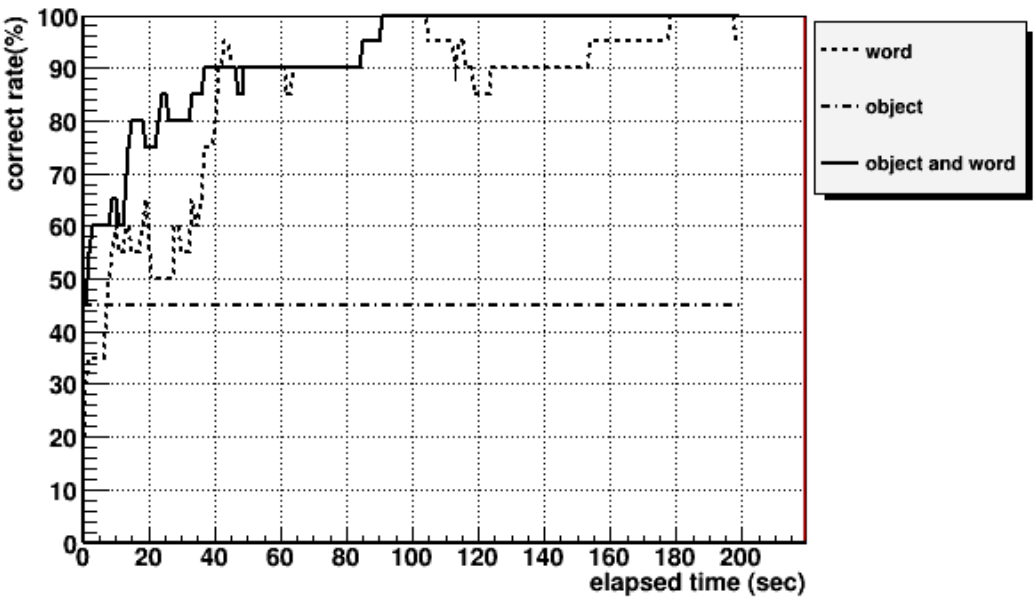


図 9 タスク正解率の時間変化

（単語のみ、オブジェクトのみ、両方を使った場合）

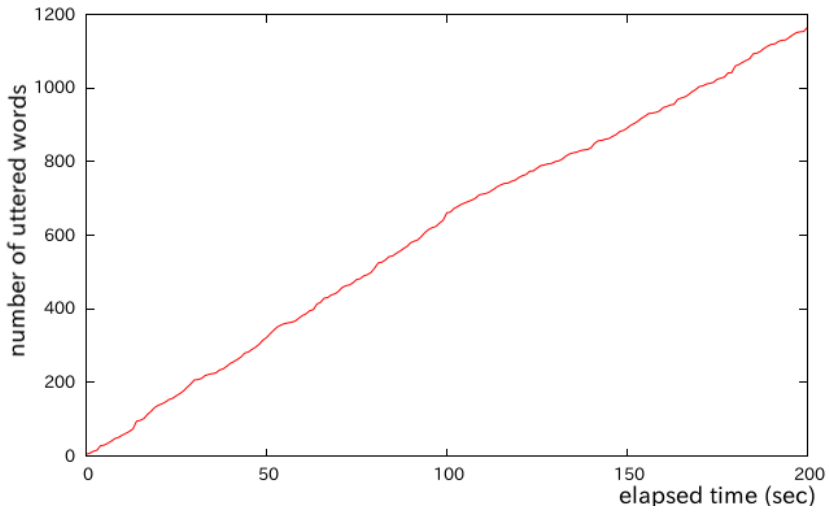


図 10 発話単語数の推移（全タスク）

表 1 絶対値で昇順に並べた左特異行列 U_k

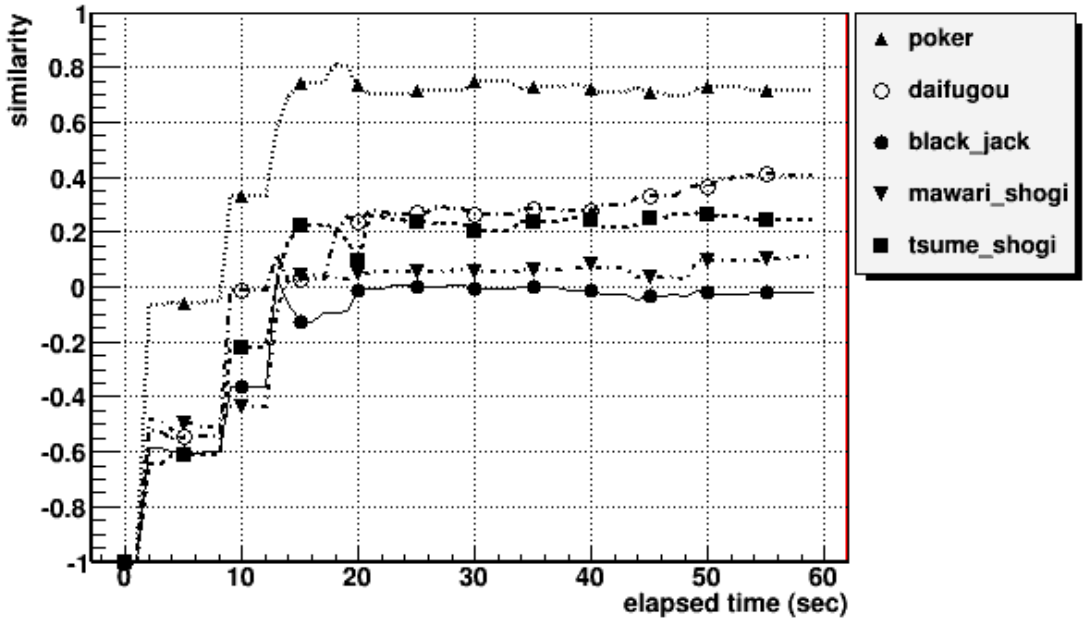
rank	1st left singular vector	2nd left singular vector	3rd left singular vector	4th left singular vector	5th left singular vector
1	Obj_トランプ -0.983	Obj_将棋駒 0.983	ヒット 0.642	1 -0.386	パス -0.346
2	Obj_将棋駒 -0.158	Obj_トランプ -0.163	1 0.410	取る 0.357	ワン 0.332
3	ん -0.0405	ん 0.0445	スタンド 0.264	ヒット 0.320	ペア 0.328
4	1 -0.0301	1 0.0221	パス -0.153	ん 0.283	いく 0.248
5	ヒット -0.0300	取る 0.0221	さん 0.148	0 -0.237	ロイヤル 0.246
6	てる -0.0249	0 0.0149	枚 -0.118	逃げる 0.175	ストレート 0.171
7	枚 -0.0201	飛車 0.0144	出す -0.116	手 0.137	ー -0.159
8	ー -0.019 7	なる 0.0129	2 0.112	こっち 0.136	8 -0.139
9	やる -0.0190	手 0.0126	なん 0.105	3 -0.126	俺 0.132
10	さん -0.0146	ー 0.0123	ペア -0.102	5 -0.114	三 -0.130

表 2　タスク開始 20 秒後の混同行列

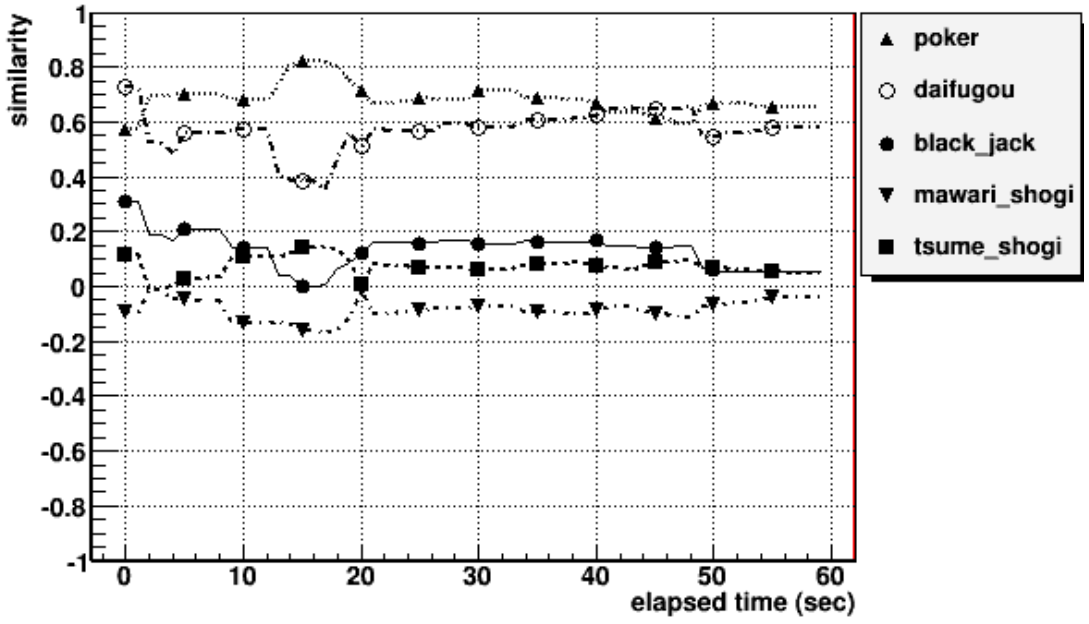
		タスク認識結果				
		ポーカー	大富豪	ブラックジャック	まわり将棋	詰将棋
正解	ポーカー	3	1	0	0	0
	大富豪	1	3	0	0	0
	ブラックジャック	2	0	0	2	0
	まわり将棋	1	0	0	3	0
	詰将棋	0	0	0	0	4

表 3　タスク開始から 100 秒後の混同行列

		タスク認識結果				
		ポーカー	大富豪	ブラックジャク	まわり将棋	詰将棋
正解	ポーカー	4	0	0	0	0
	大富豪	0	4	0	0	0
	ブラックジャック	0	0	4	0	0
	まわり将棋	0	0	0	4	0
	詰将棋	0	0	0	0	4

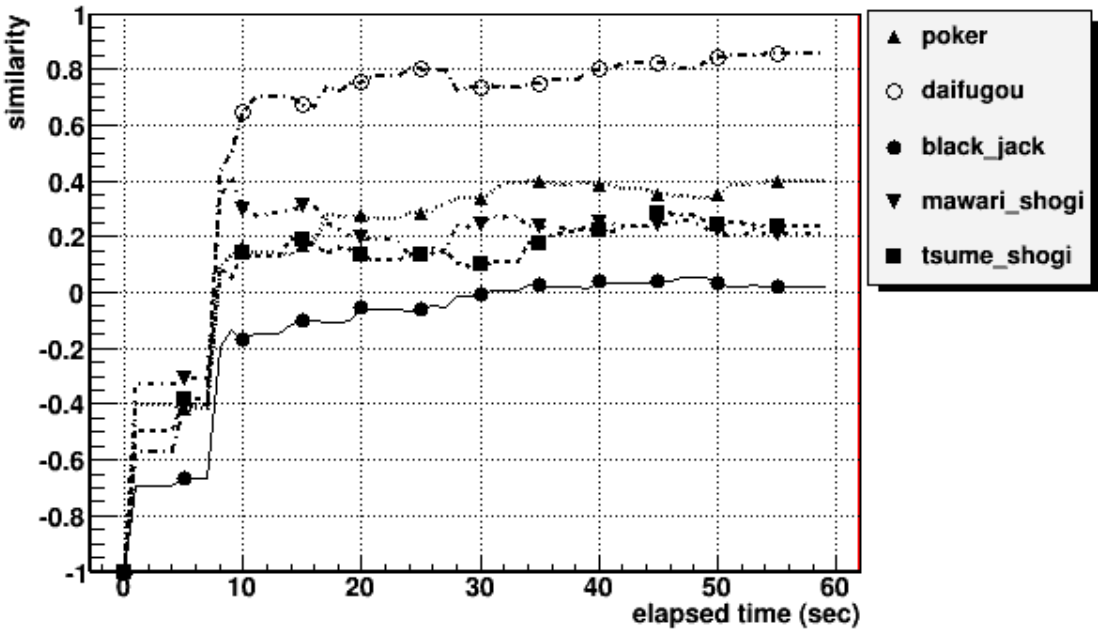


(a)　　発話単語のみ

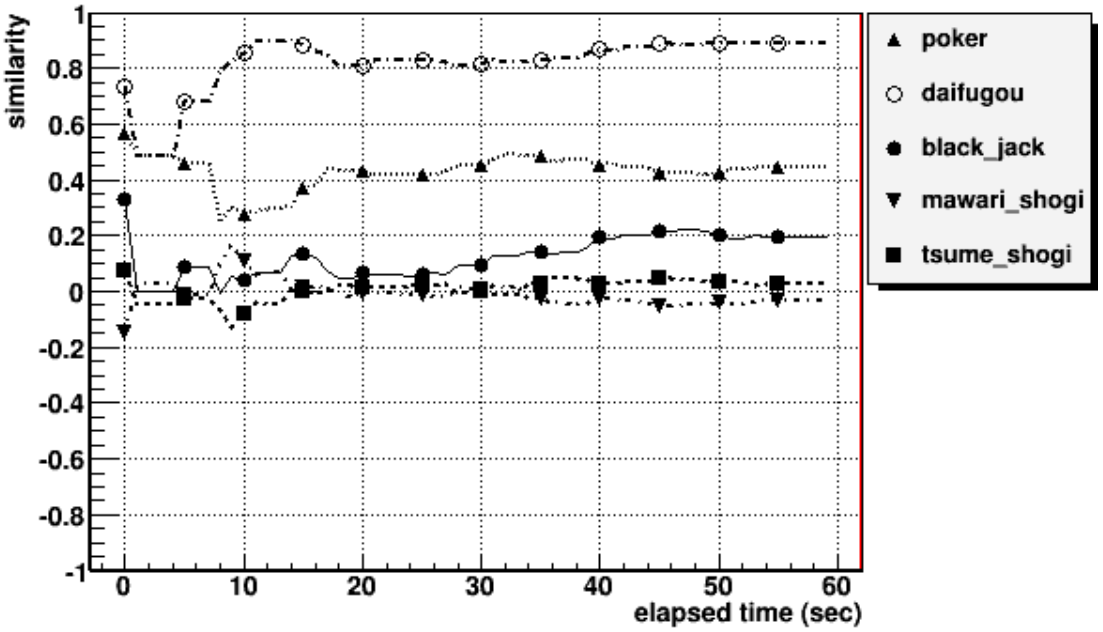


(b)　　発話単語と画像オブジェクト

図 11 平均タスク類似度の遷移　ポーカーの場合

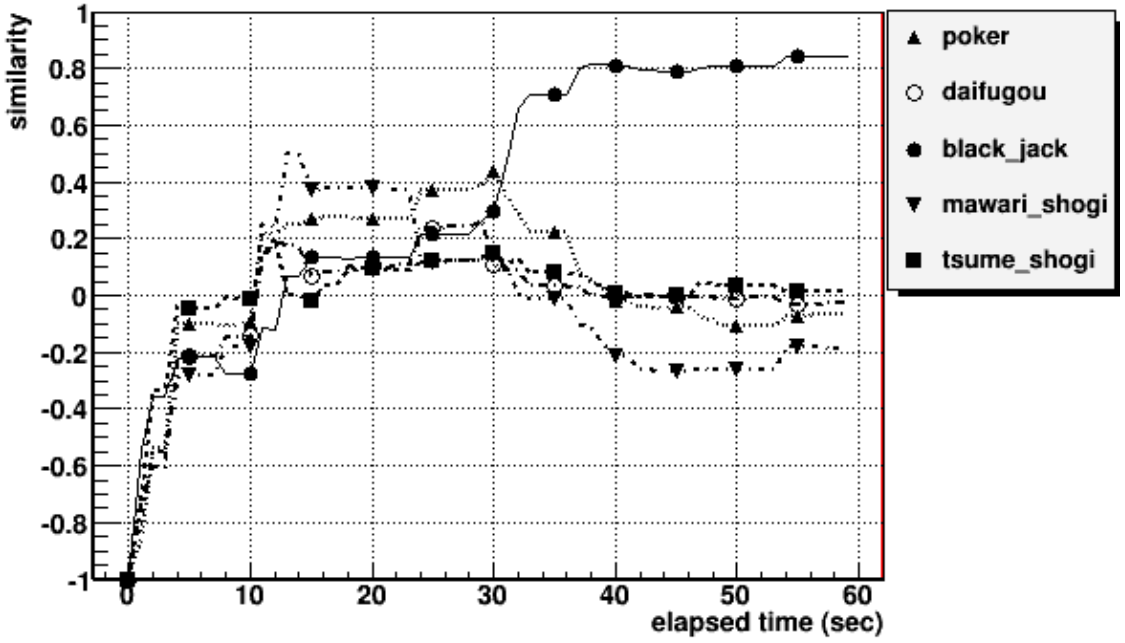


(a)　　発話単語のみ

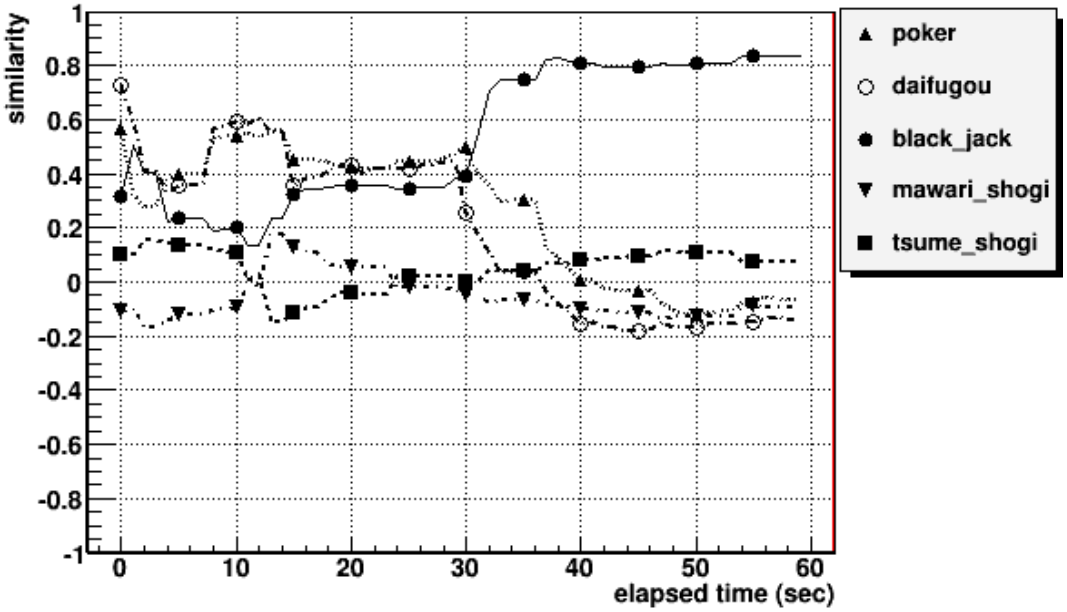


(b)　　発話単語と画像オブジェクト

図 12 平均タスク類似度の遷移　大富豪の場合

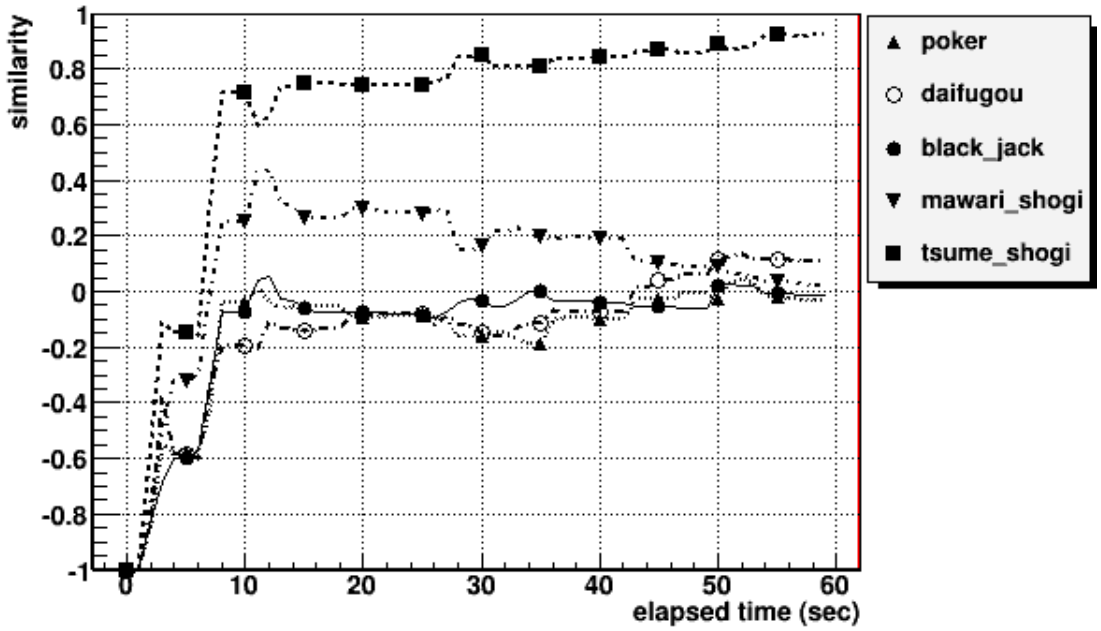


(a)　　発話単語のみ

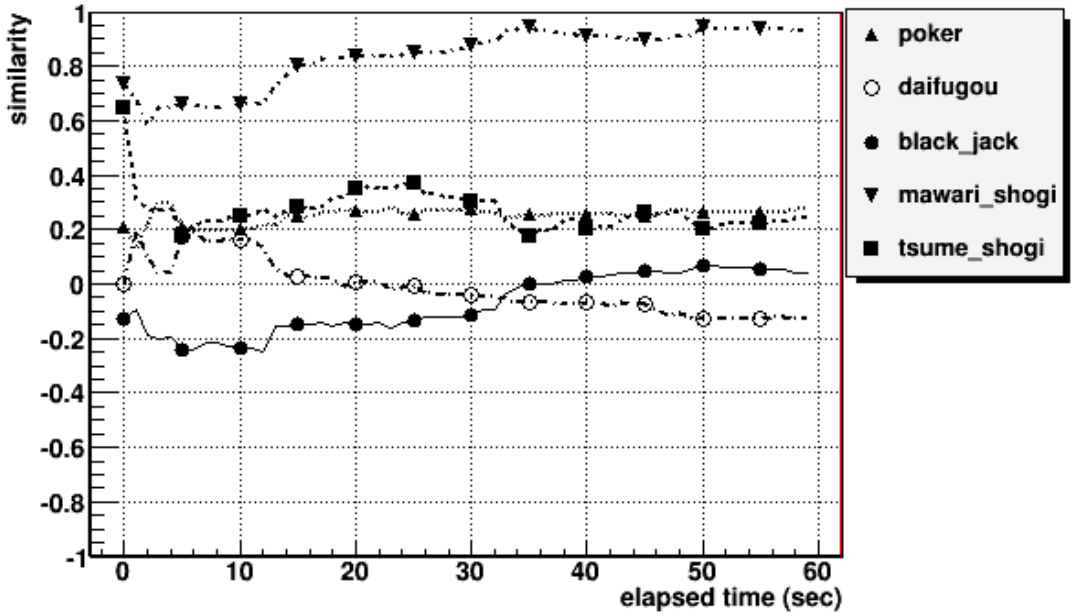


(b)　　発話単語と画像オブジェクト

図 13 平均タスク類似度の推移　ブラックジャックの場合

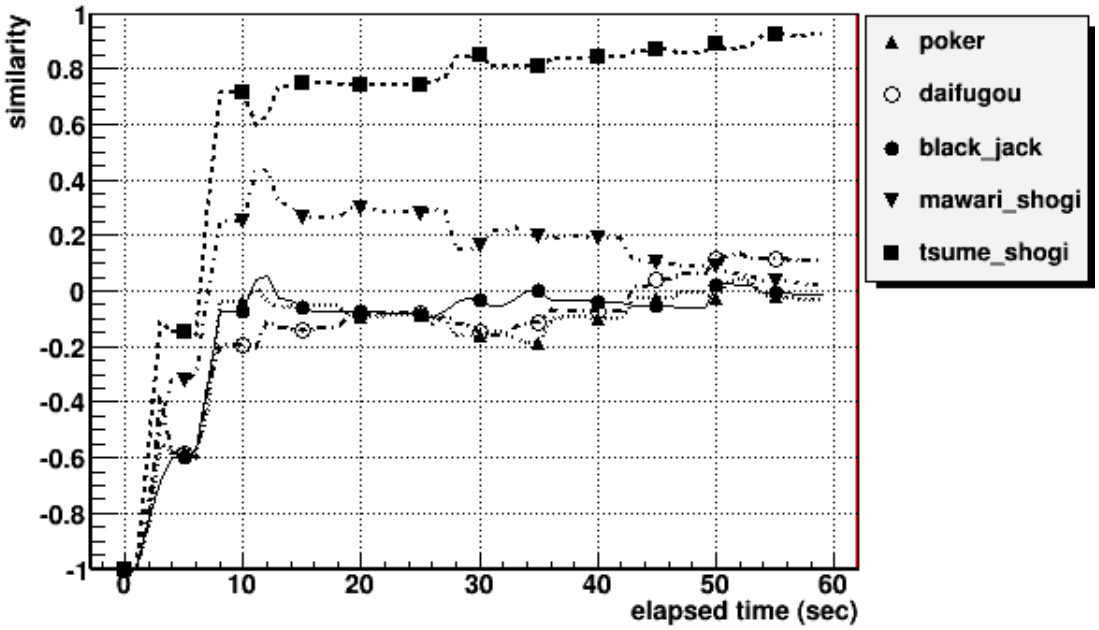


(a)　　発話単語のみ

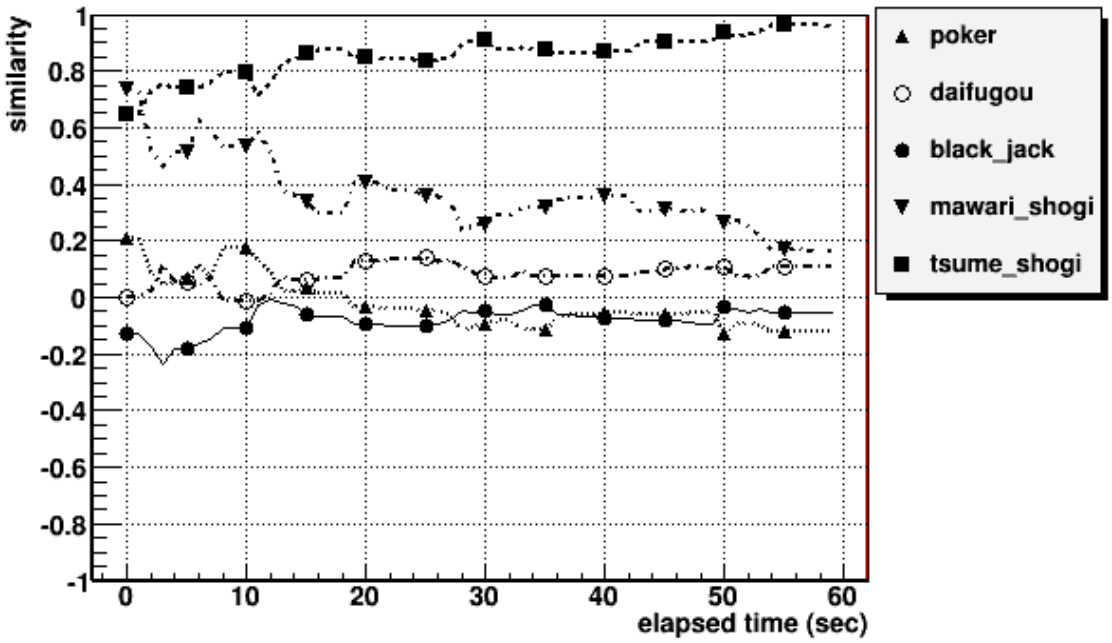


(b)　　発話単語と画像オブジェクト

図 14 平均タスク類似度の遷移　まわり将棋の場合



(a)　　発話単語のみ



(b)　　発話単語と画像オブジェクト

図 15 平均タスク類似度の遷移　詰め将棋の場合

2.5.3 TF-IDF[42] による重み付け

タスク推定率を向上させる方法として，タームに TF-IDF を用いた重み付けを行うことが考えられる．TF-IDF は文書中の単語の重み付け手法の一種である．ここで TF はタームの出現頻度を示し，IDF は逆文書頻度とも呼ばれる．TF-IDF を式 (10) に示す．

$$\mathrm{TF\text{-}IDF} = \mathrm{TF} \cdot \mathrm{IDF} \quad \cdots\cdots (10)$$

$$\mathrm{TF} = \frac{n_{ij}}{\sum_k n_{kj}}$$
$$\mathrm{IDF} = \log \frac{|\boldsymbol{D}|}{|\{d : d \ni t_i\}|}.$$

ここで，n_{ij}はターム i のタスク j での出現頻度，$|D|$は総タスク数，$|\{d : d \ni t_i\}|$はタームt_iを含むタスク数である．図 16 に重み付け無しと TF-IDF による重み付けを行った場合の平均タスク正解率を示す．また，図 17~21 に TF-IDF による重み付けを行った場合の各タスクの平均類似度の遷移を示す．

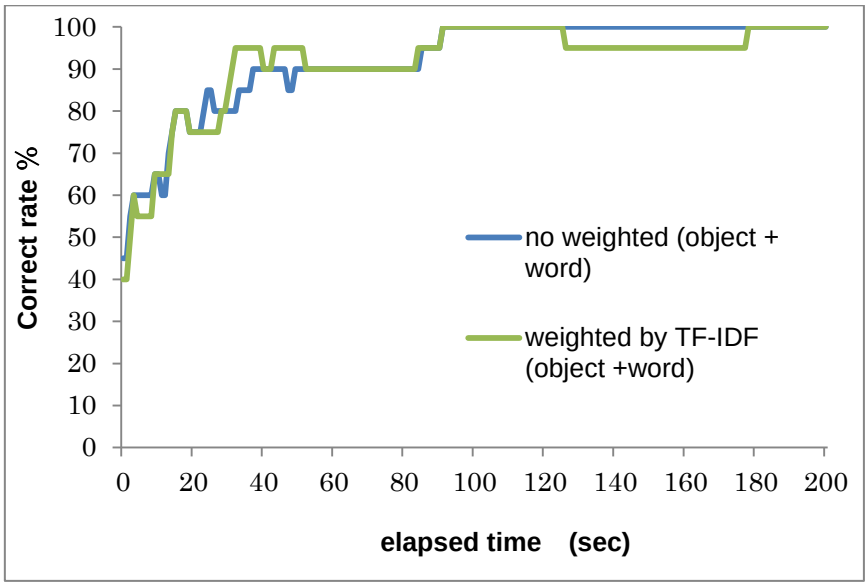


図 16　平均タスク推定率．重み付け無しと TF-IDF 重み付けの比較

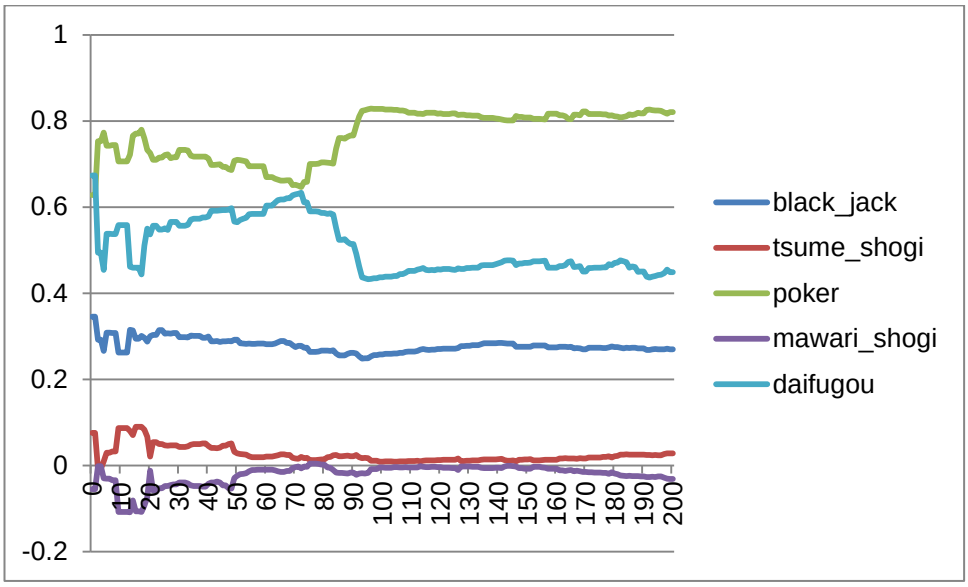


図 17 TF-IDF 重み付けを行った際の平均類似度の遷移 (ポーカータスク)

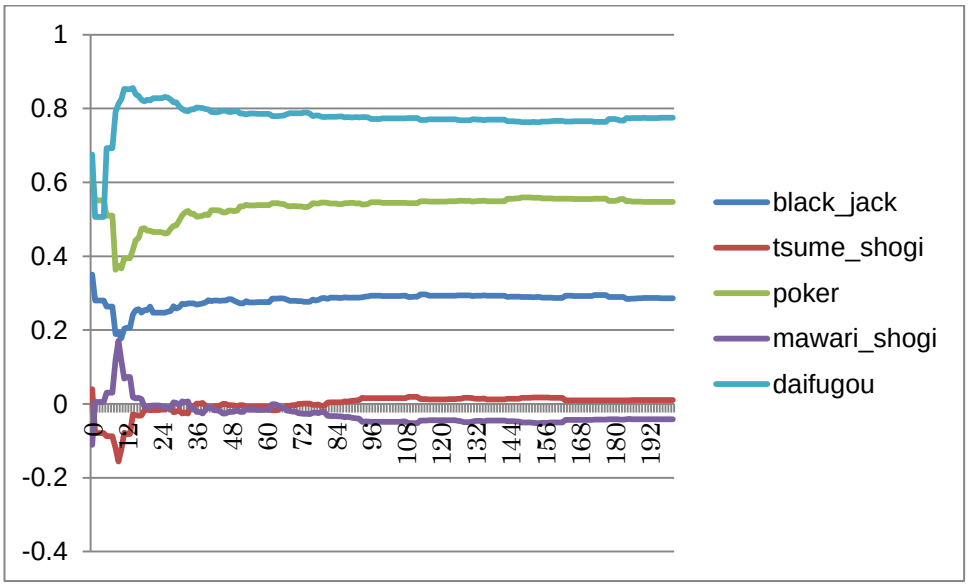


図 18 TF-IDF 重み付けを行った際の平均類似度の遷移 (大富豪タスク)

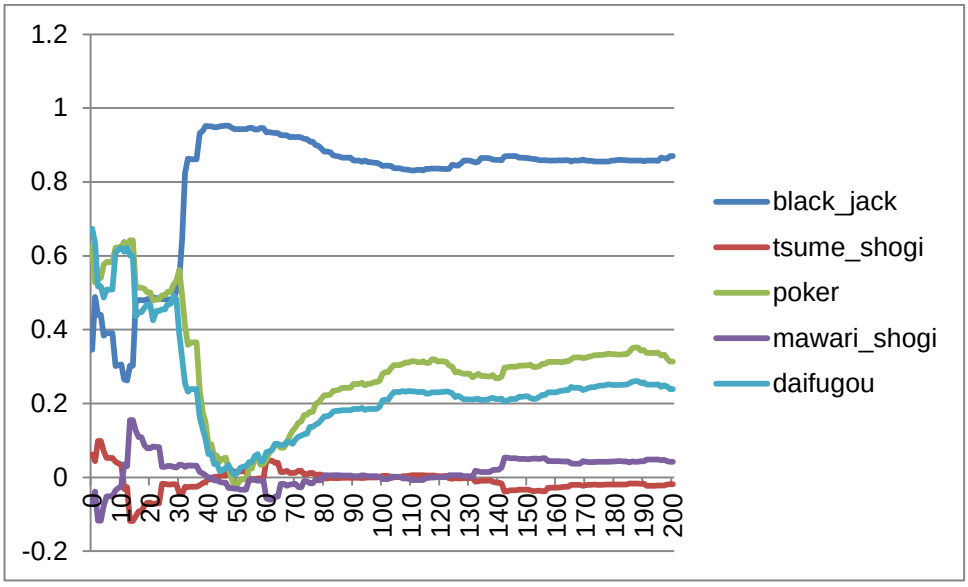


図 19　TF-IDF 重み付けを行った際の平均類似度の遷移 (ブラックジャックタスク)

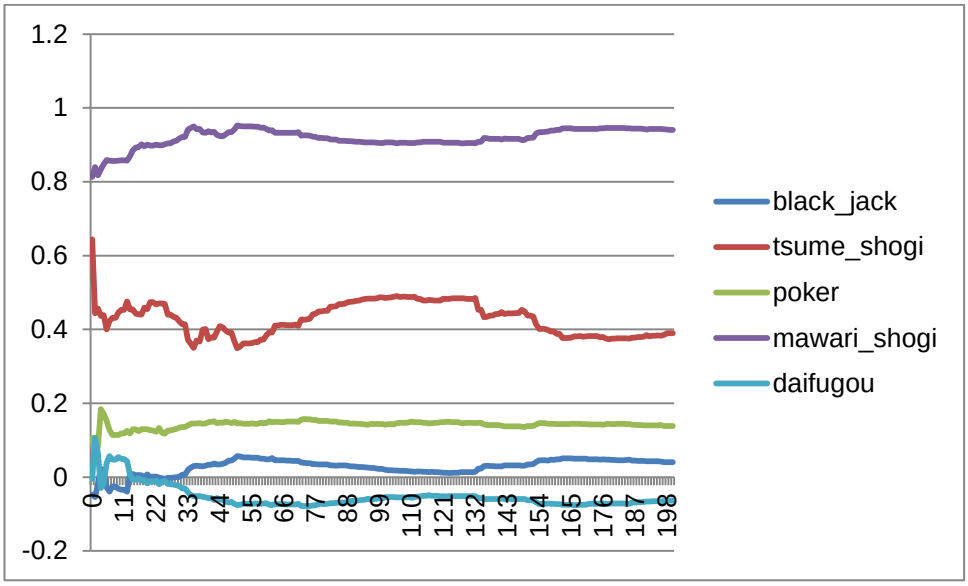


図 20 TF-IDF 重み付けを行った際の平均類似度の遷移 (回り将棋タスク)

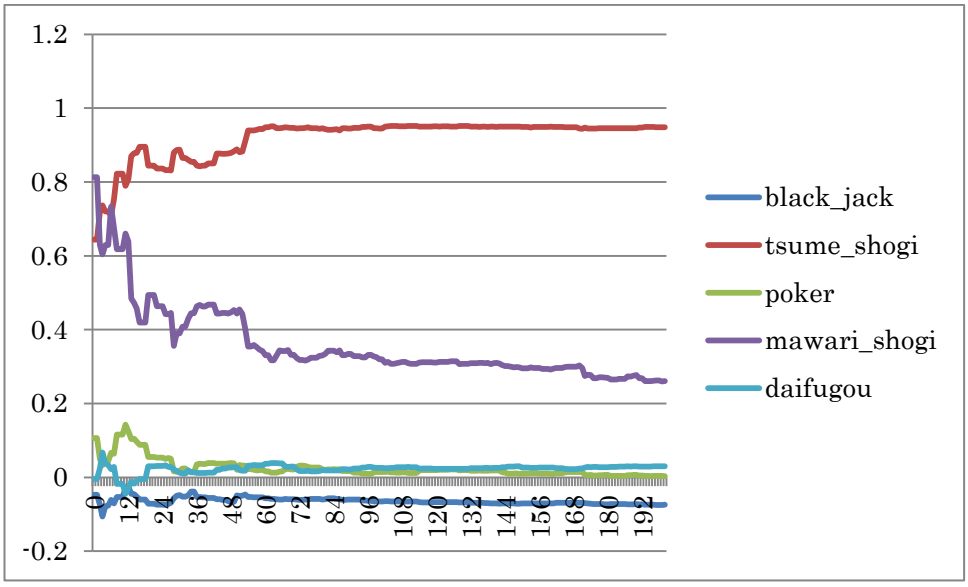


図 21 TF-IDF 重み付けを行った際の平均類似度の遷移 (詰将棋タスク)

2.5.4 考察

2.5.4.1 LSA によるタスク推定の有効性

表 1 の各左特異ベクトルの要素から，LSA によって各タスクが概ね分類されていることがわかる．第一左特異ベクトルには，各タスクに共通する要素が現れている．第二左特異ベクトルの上から二つに「トランプ」と「将棋駒」という画像オブジェクトが正負逆で現れていることから，この軸でトランプを用いるタスクと将棋駒を用いるタスクが分離される．この軸の正の方向には，「取る」，「飛車」など詰将棋に関連する発話単語も見られる．第三左特異ベクトルにはブラックジャックに関連する単語が現れている．第四特異ベクトルはまわり将棋で多く発話された「1」，「3」などの数字が負の方向に，「取る」，「逃げる」，など詰将棋に関する単語が正の方向に現れていることから，まわり将棋と詰将棋を分類するように働くと推察される．このように LSA から抽出されたタスク固有部分空間はタスク分類に有効なことを示している．

第五特異ベクトルには，「パス」に負値がついており，また，ポーカーに関連する単語が正値で上位に現れていることからポーカーと大富豪を分離するように働いていると推察される．

2.5.4.2 タスク推定の正解率

図 9 のタスク推定の正解率のグラフから，発話単語のみを用いる場合とオブジェクトと発話単語両方を用いる方法とでは，両方を用いるほうが正解率の立ち上がりが良いことがわかる．これは，タスク開始から 20 秒の時点で顕著である．また，タスク開始から 40 秒 手法がある程度のタスク推定性能を持つと言える．

単語のみを用いた場合で，最終的に正解率が 100%にならないのは，まわり将棋をポーカーと取り違えているためである．表 2 の混同行列から，20 秒の時点で特にブラックジャックがうまく推定できていないことがわかる．これは，開始から 40 秒ほどの間ブラックジャックに強く関連する単語があまり発話されなかったためである．このことは，図 13 (b) のブラックジャックの類似度のグラフからも，40 秒程度経過する時点まで，ポーカーや大富豪の類似度のほうがブラックジャックよりも高いことからもわかる．

表 2 の混同行列において，タスク開始 20 秒後のブラックジャックとまわり将棋を間違えてしまうのは，両タスクで共通して数字が発話されることによるものである．

2.5.4.3 画像オブジェクトを用いることの有効性

類似度の平均のグラフを見ると最終的に対応するタスクの類似度が一番高くなっていることがわかる．また，単語のみを用いた場合よりもオブジェクトをともに用いたほうが，図 11～15 の類似度のグラフで立ち上がりがよくなっていることがわかる．これは，単語と共にオブジェクトを用いることにより，先述した第 2 左特異ベクトルのような軸が形成され，他のタスクとの分類が行ないやすくなったためと考えられる．

画像のみを用いる場合で正解率が横ばいになるのは，画素の変化量が多いフレームを計算から除いているためと考えられる．これにより，各タスクでオブジェクトの平均数が近くなり，タスクの推定に有効に働いたのであろう．

図 11(a)と(b)を比較すると，オブジェクトを用いることによってポーカーと大富豪間の距離が小さくなるという悪影響が見られた．この原因として，(1) 両者がともにトランプをもちいるタスクであること，(2) LSA を行った結果，両者が近い空間に圧縮されたこと，の 2 点が考えられる．今回の実験では，この 2 つのタスクの類似度の接近はタスク推定の正解率には悪影響を与えていないが，このような接近が起きないように改善する必要がある．

また音声キーワード検出を用いた全タスクの平均推定率を図 22 に表す．この実験か

ら, 現在の音声認識を用いた場合本手法はうまく機能しないことがわかった. このため, 本手法を実際のタスクに試すためには、湧き出し誤りを抑制する手法や、キーワード検出率の向上手法を検討する必要がある. 図 17 に単語脱落を人工的に発生させた場合のタスク推定率を示す. これから, 単語脱落率 40%程度であれば, ８５％程度のタスク推定率を示すことがわかる. また, 単語脱落率が 50%以上に増えると, タスク推定率は大きく減少していくことがわかる. なお今回は, 挿入誤り, 置換誤りについては調査していない.

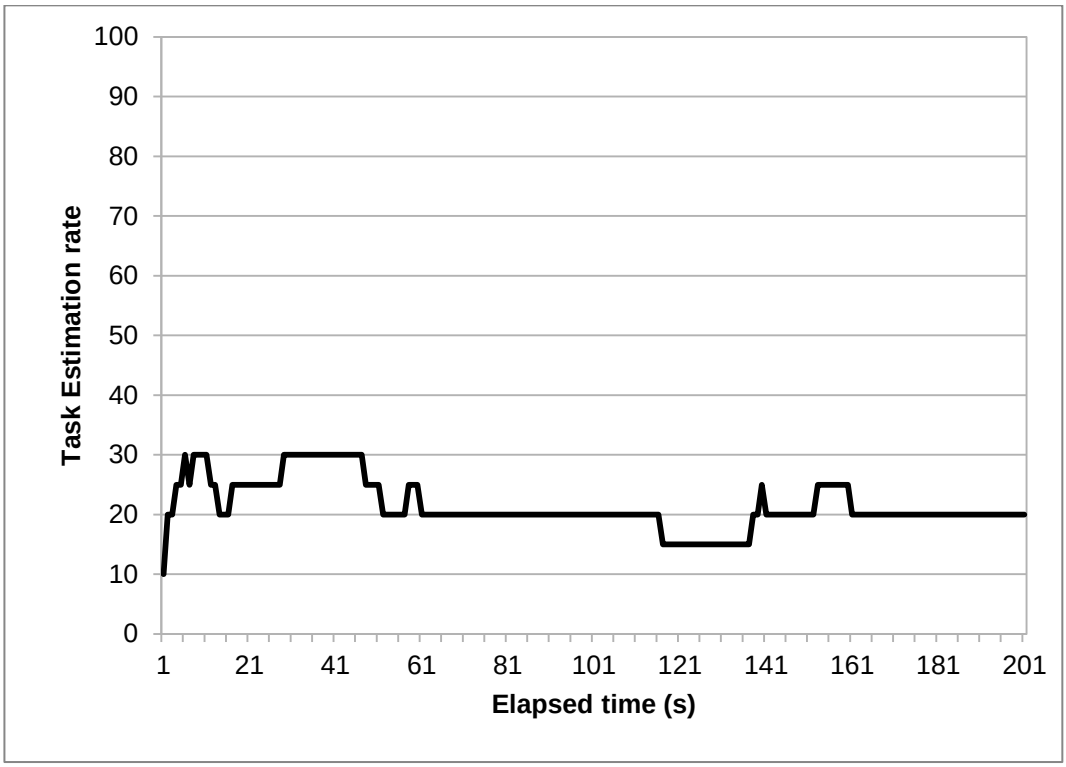


図 22 Julius による音声キーワード検出を用いたタスク推定率

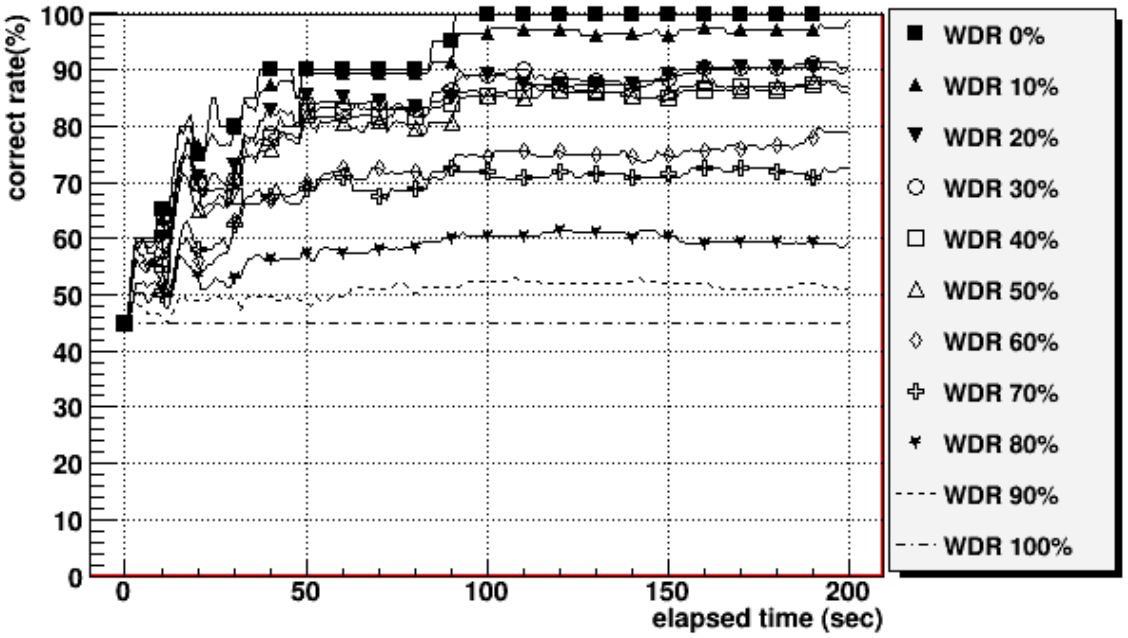


図 23 単語脱落を伴う場合のタスク正解率

2.5.4.4 扱えるタスク数の上限

タスク行列の SVD 計算は，タスク開始前に予め求めておく事が可能であり，タスク推定時の計算時間に影響を与えない．タスク数の増加に対するメモリ量の増加に関する問題も，タスク行列全てをメモリ上に展開しない方法や，ランクを低減させた近似解法が考案されている[43]．そのため，メモリに関する問題は回避できると考えられる．

また，タスク増加に対するタスクベクトルのサイズの増加については，ポーカー，大富豪，ブラックジャック，まわり将棋タスクに新たに詰将棋タスクを加えた場合の単語数の増加は 113 単語に過ぎない．ほとんどの単語はタスクに共通であり，タスク増加に対して単語数が爆発的に増加することは無いと考えられる．問題となるのは，タスクベクトルの潜在空間上への射影と類似度の計算時間である．現在のタスクベクトルの計算には 10ms 程度しかかかっておらず，タスクベクトルのサイズは数万語までなら実時間で処理できる．タスク増加に対する語彙の増加が 100 単語程度ならば，本手法は数百タスク程度の分類が実時間で可能である．タスク共通の単語が多いことを考えるとタスク数の増加に対する語彙の増加はもっと小さくなると考えられること，ベクトルプロセッサを用いて類似度の内積演算を並列化できることなどから，実際に取り扱えるタスクの上限は数百〜数千になると考えられる．

2.5.4.5 TF-IDF の効果

図 16 から，TF-IDF を用いた場合は重み付けを用いない場合とくらべてタスク推定率が安定しないことが見て取れる．また，TF-IDF を用いたことによる推定速度の向上も観測されない．

一方，図 17~21 の TF-IDF を用いた平均類似度のグラフでは，図 11~15 の重み付けをもちいない場合とくらべて，各タスク間の類似度の距離が大きくなっている．TF-IDF を用いることによって，平均的なタスク間の類似度を広げる効果があることがわかる．しかし，一部のタスクにおいて推定誤りを発生させている．

以上から，TF-IDF による重み付けには一定の効果があるものの，タスクによっては負の効果を与えることがわかる．

2.6 まとめ

タスク遂行中に現れた発話単語と画像オブジェクトから作成した行列をもとに LSA を行うことで，人が遂行中のタスクを推定する手法を提案した．実験では，机上で行うゲームタスク，ポーカー，大富豪，ブラックジャック，まわり将棋，詰将棋に対して本手法を適用し，タスク開始から 40 秒ほどで 90%の推定率が得られることを示した．提案手法は，タスク遂行中の発話単語と，使用したオブジェクトのみを用いており，動作の推定などの複雑な処理を行っていない点に利点がある．人が直接従事するタスクを推定することは，エージェントが人と対話する際に重要な情報となる．

また，TF-IDF を用いることによりタスク間類似度の分離がよくなるものの，正解率が途中で振動し 100%に届かなくなることを示した．

今後の課題としては，以下の三点があげられる．

(1) 机上以外にあるオブジェクトの認識
(2) タスク推定開始直後の性能改善
(3) 音声認識を用いた発話文の取得

(1) に関しては，認識手法の改良，あるいは，マーカーなどの使用があげられる．我々はオブジェクトを用いてタスクを遂行するときに，それを手に持ち操作したり，身につけたりすることが普通である．机上以外のオブジェクトを認識することで適応可能なタスクを増やすことができる．例えば，TV 番組の分類などに応用できるであろう．

(2) に関しては，LSA の改善が考えられる．例えば，テキスト分類の研究では，単語のベクトル空間の分類にサポートベクタマシンなど用いることによって，高い分類性能が得られることが示されている[43]．また，部分空間法の改良によっても分類性能の向上が見込めるであろう．

(3)の問題に関しては，現在の音声認識技術を用いて，今回の実験に用いた精度の書き下し文を得ることは難しい．これは，タスク遂行中の発話音声には，複数話者，割り込み，雑音などが多く含まれていること，全てのタスクに対応する辞書を用意するのが難しいことなどの問題があるためである．図 22 の Julius の音声実験結果は，本論文の目標であるタスク推定実験には，性能面から利用することができなかった．音声スポッティング能力が高い認識エンジンの開発が望まれる．

一方，筆者らの研究グループでは高い認識性能が期待できるワンモデル音声認識・合成方式について検討を進めてきた．この方式は，3.1.3 で述べるように，従来方式(MFCC/HMM)と比べて，高い認識能力を示す．またこの方式では，音声認識と合成を一つのモデルで行うため，将来，合成による認識(音声認識結果から音声を再合成して入力音声と比較する有力な認識方式)の実現が期待できる．次章ではワンモデル音声認識・合成方式の特に合成技術について検討する．

3 章　調音運動に基づくワンモデル音声認識・合成

２章のマルチモーダルタスク推定手法において，キーワードの抽出率が重要であることを示した．そこで，より精度の高い音声認識を実現するために，調音運動に基づくワンモデル音声認識・合成方式を提案する．提案システムの概要を図 24 に示す．上半分が音声認識部で下半分が音声合成部である．これらは共通の音響モデルである Articulatory Movement HMMs を利用する．

認識部（＜SR Engine>）では，二段のニューラルネットワークで構成された AF 抽出器（Articulatory Feature Extractor）によって調音特徴が抽出され，その後 Articulatory Feature HMMs（AF-HMM）によって音素列へと変換される．合成部（＜SS Engine>）では入力されたテキスト情報から音素列を得て，これを上述の AF-HMM を用いて調音特徴列へと変換する．変換された調音特徴列はニューラルネットワーク（AF-LSP Converter）を用いて LSP 系列へと変換される． AF-LSP Converter は 線スペクトル対(Line Spectrum Pair; LSP)解析器(LSP Analyzer) によって学習される．この LSP 系列とフィルタの駆動音源となる残差信号（residual signal）を声帯(Vocal Tract; VT)振動と音源モデル(VT and Source Model) から得て，LSP 合成器にかけることにより合成音声を得る．先行研究において，ワンモデル音声合成・認識の認識部について，すでに研究を行っている[44,45]．本章では最初にワンモデル音声認識・合成で用いる要素技術である HMM と調音特徴について説明した後に，ワンモデル音声認識・合成方式の要の技術の一つである音声合成の詳細を述べる．

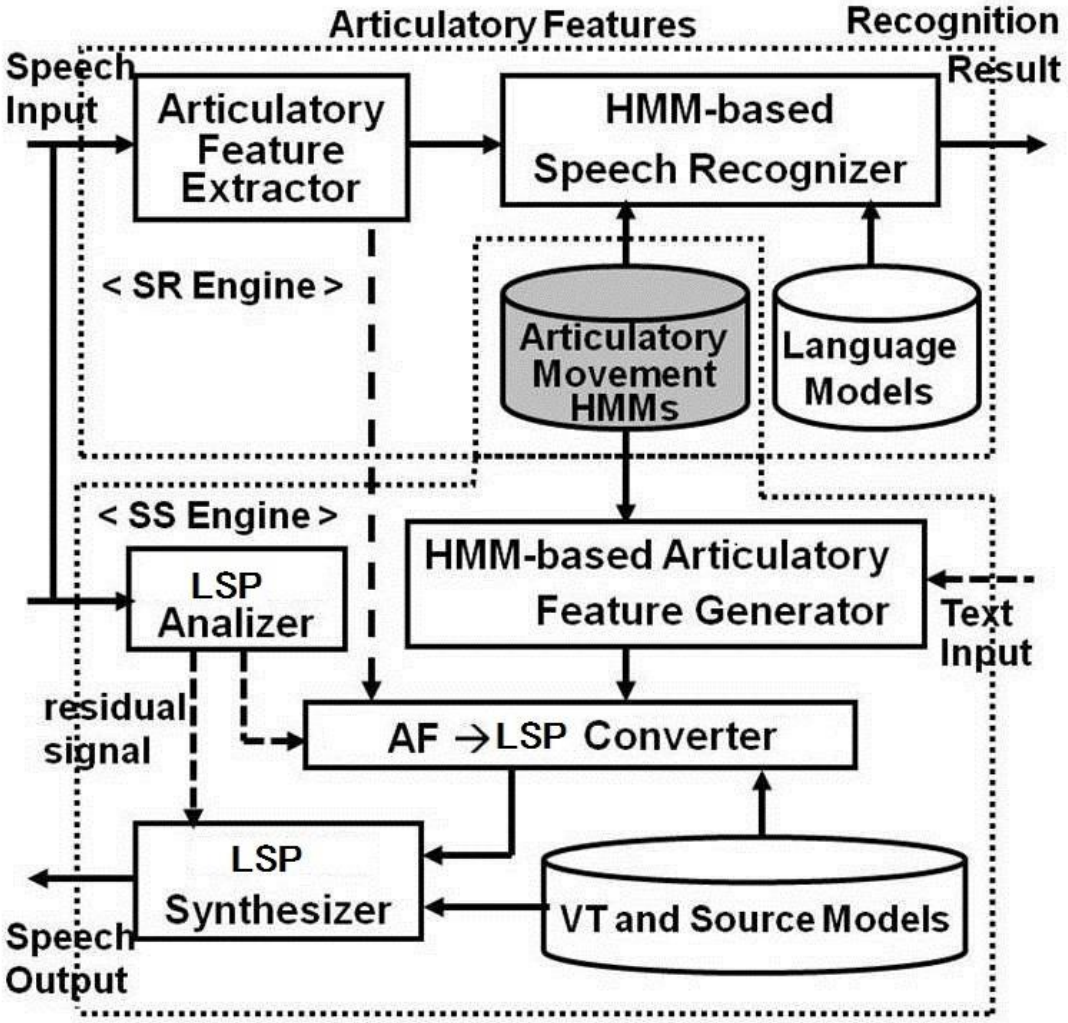


図 24 ワンモデル音声認識・合成システム

3.1 ワンモデル音声認識・合成で用いる要素技術

3.1.1 隠れマルコフモデル[46]

音声に含まれる音素の継続長は平均して数 10msec のため, 10msec をフレーム周期とする標準的な音声分析では，音素が複数フレームにまたがる．また，音素継続時間は前後のコンテキストにより，同一のコンテストでも発話により変動する．したがって音素をモデルとして表現するには，様々な長さの時系列信号生成に対応する必要がある．そのため，一般的な音声認識や一部の音声合成方式では，隠れマルコフモデル（Hidden Markov Model: HMM）を音素表現のモデルに用いることが多い．音声認識・合成の分野ではこの HMM を音響モデル(Acoustic Model)と呼ぶ．音響モデルとしては通常，図 25 に示す left to right HMM が利用される. HMM は音素ごとに作成される. 各 HMM は複数の状態を持ち，状態 i から状態 j への遷移確率 $a_{i,j}$ に基づき遷移する．また，各状態は出力確率 b(x) に基づき信号を生成する．ここで，x は入力特徴量を示す．状態音声認識では, 入力された特徴パラメータ系列に対する出力確率が最も高い HMM 系列に相当する音素列が認識結果となる．音声合成では，与えられる音素列に対応して連結した HMM の各状態から生成される出力信号系列を音声合成器に入力して合成音声を得る.

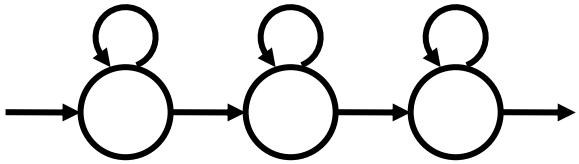


図 25 隠れマルコフモデルの例

以下に HMM の動作概要を説明する. 連続分布 HMM のパラメータ集合を λ とし, λ から次式に示す出力確率最大で長さ T の出力ベクトル系列 $\mathbf{X} = \{\mathrm{x}_1, \mathrm{x}_2, \cdots x_t, \cdots \mathrm{x}_\mathrm{T}\}$を求めることを考える．すなわち Viterbi パス上の確率

$$\overline{\mathrm{P}}(\mathbf{X}|\lambda, \mathrm{T}) = \max_{\mathrm{Q}} \mathrm{P}(\mathbf{Q}, \mathbf{X}|\lambda, \mathrm{T}) \quad \text{.....(11)}$$

を$\mathbf{Q}$に関して最大化することを考える．ただし$\mathbf{Q}$ は状態の遷移系列である.

$$\mathbf{X}_{\max} = \mathrm{argmax}_{\mathbf{Q}} \mathrm{P}(\mathbf{X}|\lambda) \quad \text{..... (12)}$$

ここで，最尤パスに相当するHMM状態系列 $\mathbf{Q}$: $\{\mathbf{q} = [\mathrm{q}_1, \cdots, \mathrm{q}_\mathrm{T}]^\mathrm{T}\}$ を通る尤度関数を次式で近似する.

$$\mathrm{P}(\mathbf{X}|\lambda, \mathrm{T}) = \mathrm{argmax}_{\mathbf{x}} \mathrm{P}(\mathbf{X}|\mathbf{Q}, \lambda, \mathrm{T}) \cong \mathrm{argmax}_{\mathbf{x}} \left\{ \max_{\mathbf{Q}} P(\mathbf{X}, \boldsymbol{Q}|\lambda, T) \right\} \text{. . . . . (13)}$$

さらに,

$$\mathrm{P}(\mathbf{X}, \mathbf{Q}|\lambda, \mathrm{T}) = \mathrm{P}(\mathbf{X}|\mathrm{Q}, \lambda, \mathrm{T}) \cdot \mathrm{P}(\mathbf{Q}|\lambda, \mathrm{T}) \quad \text{..... (14)}$$

とする．$\mathbf{Q}$を$\mathrm{P}(\mathbf{Q}|\lambda, \mathrm{T})$ によって定めた後 $\mathbf{X}$ を定めることにすれば，式 (11)の最適化問

題は次に示す，式(15), (16) のように近似する．

$$\mathbf{Q}_{\max} = \operatorname{argmax}_{\mathbf{Q}} \; P(\mathbf{Q}|\lambda, T) \quad \text{..... (15)}$$

$$\mathbf{X}_{\max} = \operatorname{argmax}_{\mathbf{x}} \; P(\mathbf{X}|Q_{\max}, \lambda) \quad \text{..... (16)}$$

式 (15) は状態継続長モデルを与えることで解くことができるが，本論文では簡単化のため事前に与えられるものとする．また式 (16) については，調音特徴が疎な特徴であることを利用して，各状態は単一ガウス分布を持つと仮定すると，$\mathbf{X}_{\max}$ は平均ベクトルで与えることができる．

3.1.2 調音特徴

調音特徴(Articulatory Feature: AF)[47, 48, 49] は，調音方法と調音部位から話者不変な音素属性を規定する．著者らの提案する AF は，音素間の調音上の対立を中心にした弁別的特徴(Distinctive Phonetic Features: DPF[49, 50])を基にして設計されている．AF とは，調音様式（母音性，連続性，摩擦性，舌端性，破裂性，連続性，鼻音性，半母音性）と調音部位（高舌性，低舌性，nil(高/低)，前方性，後方性，nil(前/後)），音源（有声性，非有生性）の情報からなる特徴である．ここで前方性，後方性は舌の最も盛り上がる位置を示している．AF は 2 値ではなく 0〜1 の実数値のスコアを取る．1 に近いほどその特徴を持つことを示している．なお，本論文では音響モデルを，AF を入力とする連続 HMM で表現している．本論文で使用する AF は，表 4 に示す特徴から音素を規定している．表では，DPF の一般的表現に基づき，1.0 を“+"，0.0 を“−" としている. AF 抽出器は中間値を含む実数値を出力する. 表の AF の基となった DPF は，音声認識・合成での利用を前提にしたものではないため，例えば音素により“+”の数が偏っているなど，音素分類には適さなかった．このため，表では各音素で"+" 特徴の数が 5 前後になるよう，中間的な特徴（nil（前/後），nil（高/低））を加えて調整している[50].

一般的な音声認識・合成システムは，音声の特徴パラメータとして短時間パワースペクトル情報に基づく MFCC（Mel-Frequency Cepstrum Coefficient）などが用いられる．これに対して，筆者らの方式は，調音様式と発声システムを分離できるため，音声を少量の音声試料で合成できる可能性がある．音声の生成が顎，唇，舌などから構成される力学系の運動に拘束されることを考えると, 特に調音結合の問題については AF など調音次元の特徴量を用いることが，解決への近道と考えられる．

多層ニューラルネットワーク（Multi-Layer Neural network; MLN ）は前後の音素環境の違いを学習しており，時刻 t で抽出された AF はその前後 (t-3，及び，t+3) を含む 3 フレーム分に相当する値が抽出される．音声認識で広く利用される MFCC などのスペクトルパラメータと比べ，AF は分散がごく小さいため，HMM の混合数が少ない場合にも高い性能が得られる．評価実験では 1 混合の monophone-HMM の場合で 80%を越える音素正解率を実現している[47]．本論文では，AF 抽出過程に先行研究の中で提案してきた，3 段の MLN を用いた高精度な AF 抽出方法を用いる．

音声波形から AF を抽出する処理の流れを図 26 に示す．AF 抽出器は，以下のモジュールからなっている．それぞれ，

(1) 局所特徴(Local Feature; LF) 抽出器[51]
(2) AF-LF 変換器
  - MLN-I により LF から AF へのマッピングを学習する．
(3) AF 系列整形部
  (3-1) MLN-II により注目フレーム t と t±3 のフレームのコンテキストを利用して AF 整形を行う．
  (3-2) MLN-III により⊿AF，⊿⊿AF を入力とした整形を行う．

からなる. MLN は注目フレーム t の前後（t-3, t+3） 3 フレームの音素環境の違いを学習しており，抽出される AF は合計 45 (15×3) 次元のベクトルを構成する．

音声認識で広く利用される MFCC などのスペクトルパラメータと比べ，AF は分散がごく小さいため，HMM の混合数が少ない場合にも高い性能が得られる．抽出された AF の例を図 27 に示す．発話は ATR 音素バランス文に含まれる「人工衛星…」の部分である．細線が理想値，太線が抽出結果を示す．図から比較的高い抽出結果が得られて

いる事が分かる.

表 4 15 次元調音特徴表

AF	a	i	u	e	o	N	w	y	j	my	ky	dy	by	gy	ny	hy	ry	py	p	t	ts	ch	b	d	g	z	m	n	s	sh	h	f	r
Vowel	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
High Tongue	−	+	+	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	+	−	−	+	−	−	−	−	+	−	+	−
LowTongue	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−
Nil(H/L)	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−	+	+	−	+	−	+	+	+	−	−	+
Front	−	−	−	−	−	−	−	−	−	+	−	+	+	−	+	−	+	+	+	+	+	+	+	−	+	−	+	+	+	−	+	+
Back	+	−	+	−	−	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−
Nil(F/B)	−	+	−	+	+	+	−	+	−	−	+	−	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−
Coronal	−	−	−	−	−	−	−	−	+	−	−	+	−	−	+	−	+	−	−	+	+	+	−	+	+	−	+	−	+	+	−	−	+
Plosion	−	−	−	−	−	−	−	−	−	−	+	+	+	+	−	−	−	+	+	+	−	−	+	+	+	−	−	−	−	−	−	−	−
Affiricate	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	+	−	−	−	−	−	−	−
Continuous	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	+	+	+	+	−
Voiced	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	−	−	−	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
UnVoiced	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	+	+	+	+	+	−	−	−	−	−	−	+	+	+	+	−
Nasal	−	−	−	−	−	+	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	−	−
Semi-Vowel	−	−	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+

(+:A phoneme has perticular AF, −: A phoneme does'nt have perticular AF)

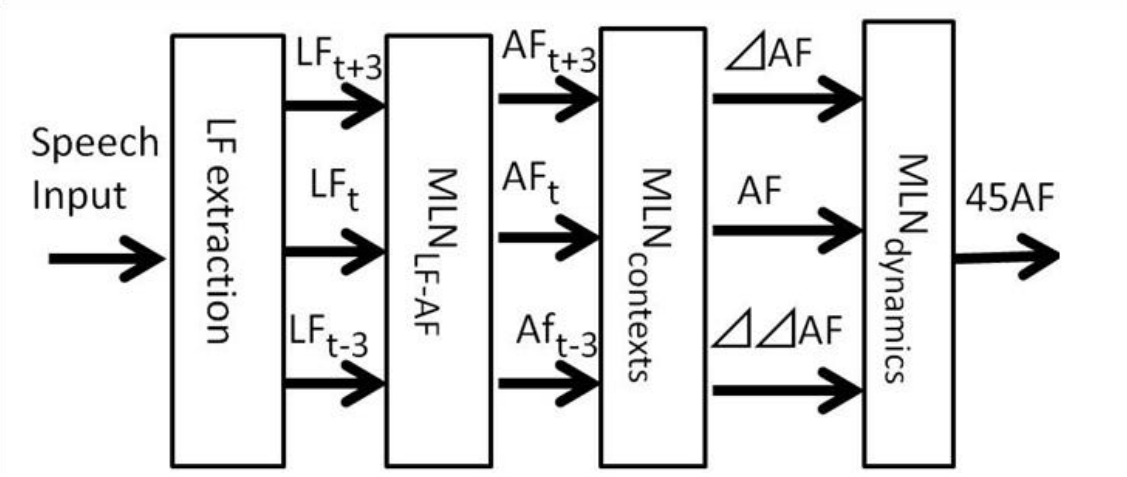


図 26 調音特徴抽出の流れ

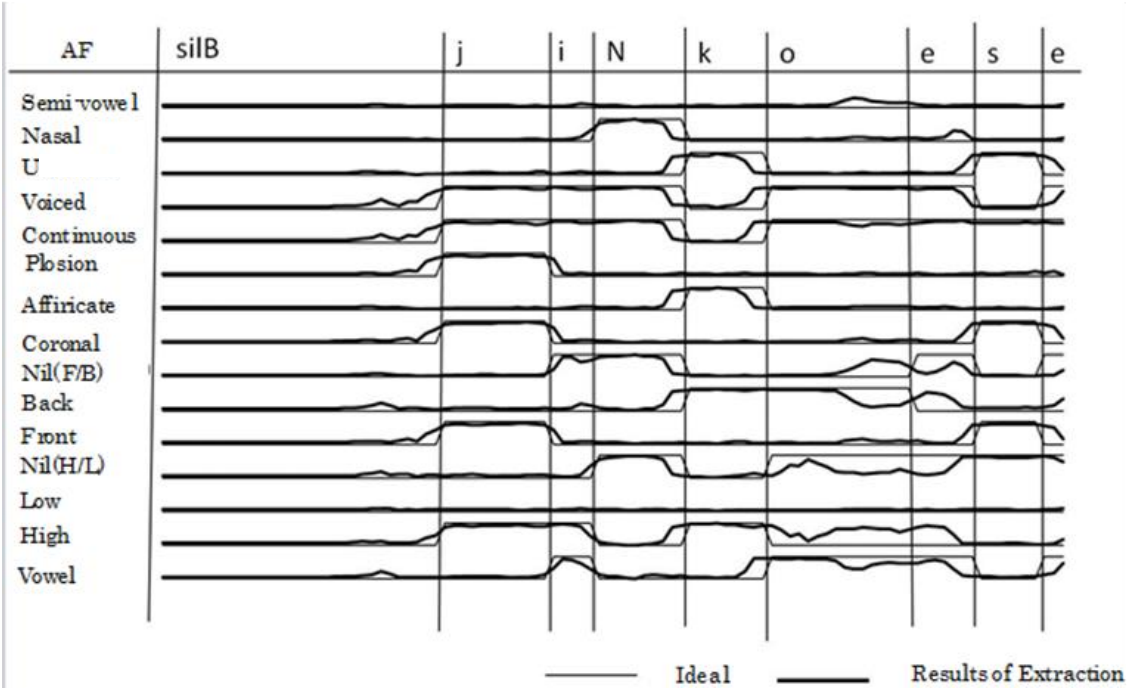


図 27 AF 抽出結果の例　発話文：「人工衛星…」

3.1.3 調音特徴に基づく音声認識

我々はこれまで調音特徴による音声認識法を提案し，精度の高い音声認識エンジンを研究している．以下にこの概要を説明する．このワンモデルの音声認識エンジンは，図 24 に示すように，入力音声を AF 系列に変換する AF 抽出器と，調音運動を表現した AF-HMM から成る．

以下では音素認識率の評価を示す．音声コーパスには JNAS を使用した[52]．HMM は 5 ステート 3 ループの標準的な left-to right 型を使用した．単音(mono-phone)単位で，混合数を 1, 2, 4, 8, 16 とし，学習に使用した話者は 1 名→2 名→ 4 名→ 8 名→33 名と増加させながら，音素認識性能を調べた結果を図 28 に示す．

調音特徴は登録人数に関係なく，また混合数にも無関係である．これに対して MFCC は，登録人数を増やし，同時に正規分布の数を増やすほど向上する．この結果から，調音特徴は話者不変のパラメータであることが示唆される．

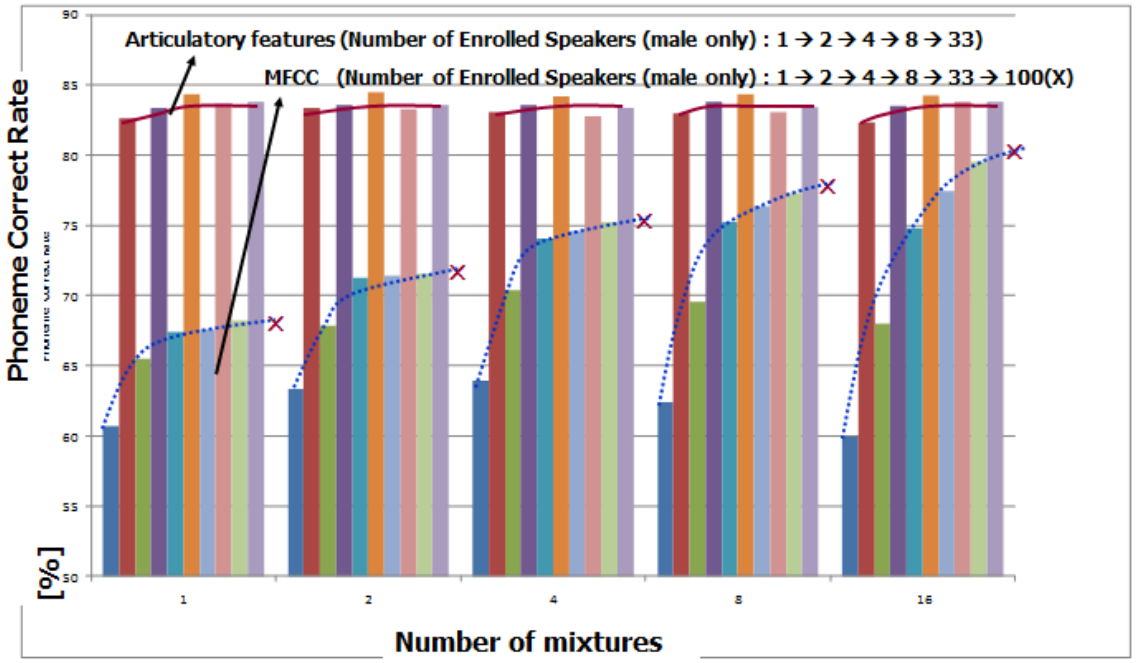


図 28 AF と MFCC の音素認識率の比較

3.2 調音運動 HMM に基づく音声合成システム

この節では音声認識と同じ枠組み（ワンモデル）で，対話 システムの応答部の主要技術である．音声合成システムの研究結果を述べる．

3.2.1 調音特徴に基づく HMM 音声合成システム

HMM から生成される調音特徴系列 AF と，さらに LSP 系列への変換と合成までの処理の流れを図 29 に示す．以下に音声合成システムの処理手順を示す．

(1) HMM を，音素列と状態継続長に従い連結する．
(2) HMM の各状態から AF（平均ベクトル）を出力する．
(3) 得られた AF を$\mathrm{MLN}_{\mathrm{AF-LSP}}$ へ与え，話者固有の LSP 係数列を出力する．この際，注目フレーム t と$\mathrm{t}\pm 3$ の AF を話者固有の$\mathrm{MLN}_{\mathrm{AF-LSP}}$ に与えること

で調音結合の問題に対処する．

（4）LSP 合成器を MLN からの LSP 係数と，残差音源符号帳からの信号で駆動し，合成音声を得る．

(1)では，システムに入力された音素列と状態継続長に従い，学習済みの AF-HMM を連結する．(2) では，式(13)を用いて，音素系列に対応する連結 AF-HMM から AF を生成する．(3)では，図 24 と図 26 の AF-LSP Converter に従い，LSP 系列を得る．この AF-LSP Converter は話者固有のものである．(4) では，3 章で説明する CELP 方式を用いた駆動音源の改良手法に基づいて，得られた駆動音源と(3) の LSP フィルタ係数を LSP 合成フィルタに与えて合成音声を得る．

3.2.2 線スペクトル対 (LSP) による声道音響パラメータのモデル化[53]

声道音響パラメータとして，線スペクトル対（LSP: Line Spectral Pairs）が知られている．LSP 分析は線形予測分析と等価な音声分析法である．線形予測法では，音声の生成モデルにおいて，式 (17) に示す全極型のフィルタで声道調音等価フィルタを模擬する．このモデルにおいて，分母の多項式に現れる係数 a_i を線形予測係数と呼ぶ．

$$\mathrm{H}(z) = \frac{1}{\mathrm{A_p}(z)} = \frac{1}{1+a_1 z^{-1}+a_2 z^{-2}+\cdots+a_p z^{-\mathrm{p}}} \quad \cdots\cdots (17)$$

線形予測分析の前向き予測誤差$\mathrm{F_M(z)}$と後ろ向き予測誤差$\mathrm{B_M(z)}$について以下の式が成り立つことが知られている．

$$\mathrm{F_{M+1}}(z) = \mathrm{F_M}(z) + \gamma_{\mathrm{M+1}}\,\mathrm{B_M}(z)z^{-1} \quad \cdots\cdots (18)$$

ここで，$\gamma_{\mathrm{M+1}}$ を 1 とした時の$\mathrm{F_{M+1}}(z)$を$\mathrm{P}(z)$，$\gamma_{\mathrm{M+1}}$ を-1 としたときの$\mathrm{F_{M+1}}(z)$を$\mathrm{Q}(z)$とすると，

$$\mathrm{P}(z) = \sum_{\mathrm{m=0}}^{\mathrm{M}} a_{\mathrm{M}}(\mathrm{m})z^{-\mathrm{m}} + \sum_{\mathrm{m=0}}^{\mathrm{M}} a_{\mathrm{M}}(\mathrm{M}-\mathrm{m})z^{-(\mathrm{m+1})} \quad \cdots\cdots (19)$$

$$\mathrm{Q}(z) = \sum_{\mathrm{m=0}}^{\mathrm{M}} a_{\mathrm{M}}(\mathrm{m})z^{-\mathrm{m}} - \sum_{\mathrm{m=0}}^{\mathrm{M}} a_{\mathrm{M}}(\mathrm{M}-\mathrm{m})z^{-(\mathrm{m+1})} \quad \cdots\cdots (20)$$

となる．ただし，$\mathrm{a_M(m)}$は線形予測係数である．ここで M を偶数とすると，$\mathrm{P}(z)$と$\mathrm{Q}(z)$は次のように因数分解できることが知られている．

$$\mathrm{P}(z) = (1+z^{-1})\prod_{\mathrm{m=1}}^{\mathrm{M/2}}(\exp[\mathrm{j}\omega(\mathrm{m})] - z^{-1})(\exp[-\mathrm{j}\omega(\mathrm{m})] - z^{-1}) \quad \cdots\cdots (21)$$

$$\mathrm{Q}(z) = (1-z^{-1})\prod_{\mathrm{m=1}}^{\mathrm{M/2}}(\exp[\mathrm{j}\theta(\mathrm{m})] - z^{-1})(\exp[-\mathrm{j}\theta(\mathrm{m})] - z^{-1}) \quad \cdots\cdots (22)$$

よって，$\mathrm{P}(z)$の根$-\exp[\mathrm{j}\omega(\mathrm{m})]$, $\exp[-\mathrm{j}\omega(\mathrm{m})]$, $\mathrm{Q}(z)$の根は，1, $\exp[\mathrm{j}\theta(\mathrm{m})]$, $\exp[-\mathrm{j}\theta(\mathrm{m})]$となる．つまり，複素平面における単位円上に$\mathrm{P}(z)$, $\mathrm{Q}(z)$のすべての根が存在する．この$\mathrm{P}(z)$, $\mathrm{Q}(z)$ の根を与える$\omega(\mathrm{m})$と$\theta(\mathrm{m})$のことを LSP 係数と呼ぶ．

$\mathrm{P}(z)$，$\mathrm{Q}(z)$から以下の式によって，$\mathrm{A_p}(z)$を再構成できる．

$$A_p(z) = \frac{P(z)+Q(z)}{2} \quad \text{..... (23)}$$

LSP パラメータから線形予測による声道調音等価フィルタと等価なフィルタを構成できる．LSP パラメータを用いて構成した音声合成フィルタ（声道調音等価フィルタ）を，図 30 に示す．

これまでに述べたように，LSP パラメータは，線形予測係数と等価なパラメータと言って良い．しかし，LSP パラメータは，線形予測係数やと比べて，量子化特性に優れ，少ないビット割り当てでも合成音声の劣化を招きにくい．また，局所的な微細な動きを除けば，LSP パラメータは発話中の音韻の推移に従って滑らかに変化する特性も持つ．

調音特徴に基づく連続 HMM を音声合成に利用する場合，HNM を MLN から直接抽出した多量のデータで学習すると，HMM から生成される AF 系列(平均ベクトル)が理想値とずれてしまうことがある．これは AF 抽出器が誤識別を起こした場合，平均ベクトルが誤った方向へ引っ張られるためである．

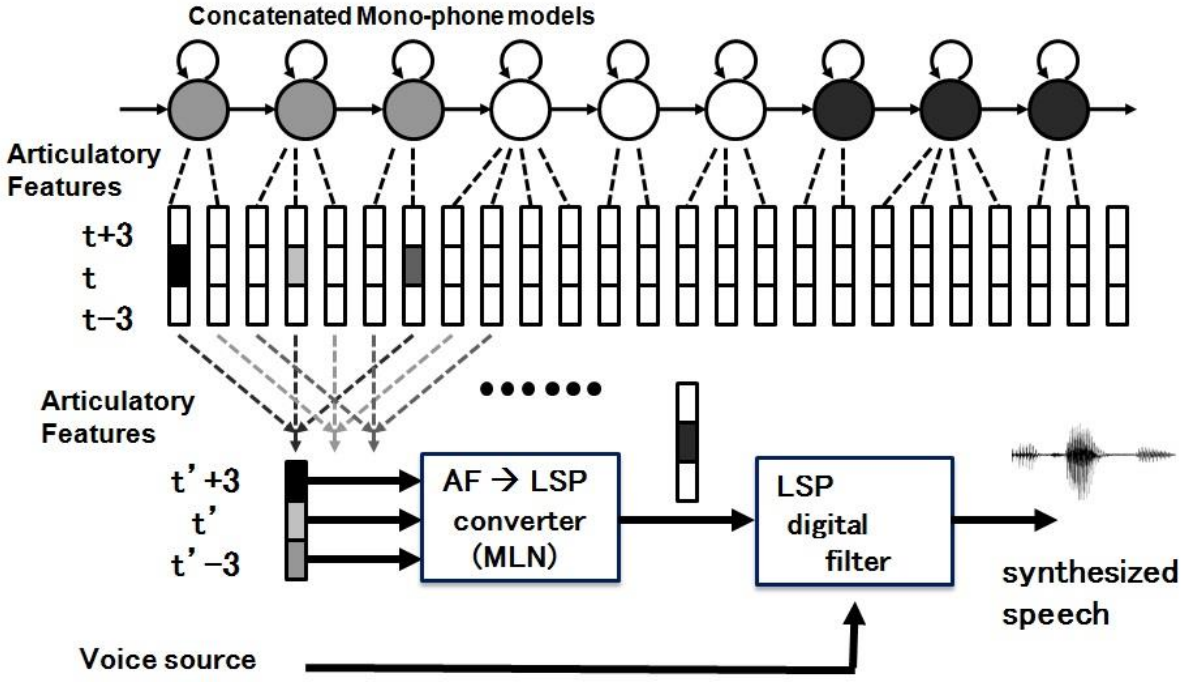


図 29 調音運動 HMM による音声合成

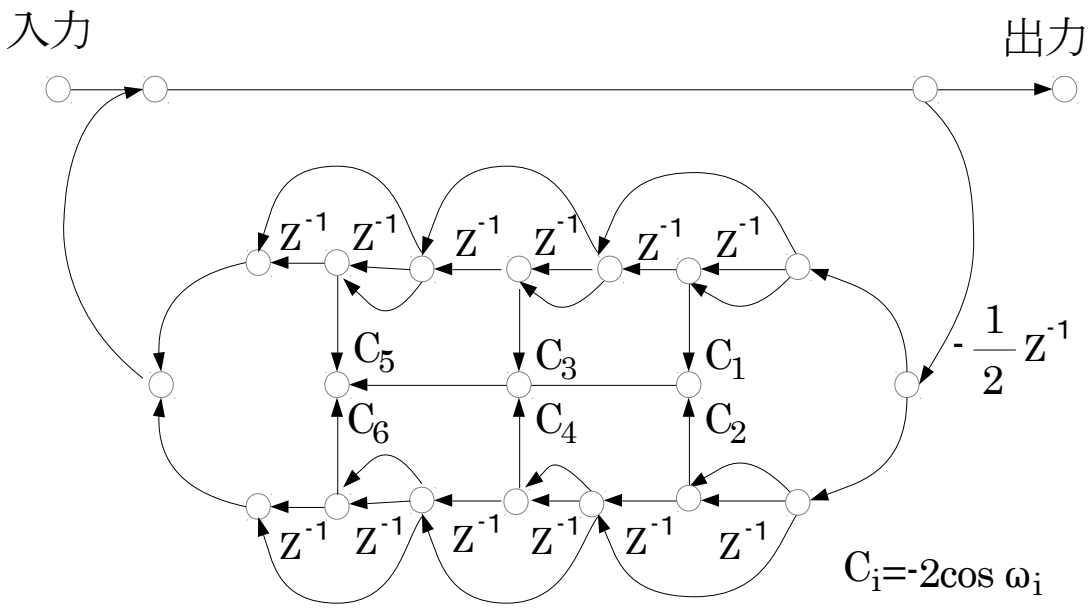


図 30 LSP 合成フィルタ(6 次の場合)

3.2.3 調音特徴－声道パラメータ変換器の改良

話者に依存しない AF 系列から話者固有の音声信号を得るため，AF 系列を MLN を用いて，LSP パラメータに変換する．

HMM から取り出される AF は平均ベクトルであるため階段状に変化する．つまり，このまま MLN を構築して LSP パラメータを生成させると LSP パラメータも階段状の物が出力され，滑らかな合成音が生成できない．そこで，MLN には注目フレーム tに加えて，$t+3$，$t-3$のフレームの AF を入力することにより，滑らかな LSP パラメータの生成を実現する．

HMM から生成される AF 系列を LSP 係数列に変換する$\mathrm{MLN}_{\mathrm{AF-LSP}}$ の学習では，音声から抽出した AF を直接用いる方法（直接学習）と，音素と調音特徴の対応テーブルから AF を出力する HMM を学習する方法（AF－HMM 学習）が考えられる．直接学習では，学習時に音声から学習された$\mathrm{MLN}_{\mathrm{AF-LSP}}$ から出力された AF が与えられ，合成時には AF-HMM から出力された AF が与えられる．このため，学習時と合成時で$\mathrm{MLN}_{\mathrm{AF-LSP}}$ に与えられる AF 系列が異なる結果，合成される LSP 系列が劣化する可能性がある．一方，AF-HMM 学習の目標値は AF-HMM の出力となり，学習時と実施時で誤差の少ない LSP が得られることが期待できる．次節では直接学習と AF-HMM 学習を実験的に比較することによって，AF-HMM 学習の優位性を示す．

3.2.4 音声合成評価実験

以下では，直接学習と AF-HMM 学習の比較実験について述べる．初めに，全般的な音質の MOS 値による主観評価結果を示し，続いて，LSP 値の相関，及び，スペクトル歪みによる客観評価結果について述べる．

3.2.4.1 実験条件

音声合成評価に使用した HMM の仕様を表 5 に，AF-LSP パラメータ変換器の仕様を表 6 に示す．音声コーパスには ATR デジタル音声データベース[53]を用いた．このコーパスには無声化母音，撥音の異音，声門閉鎖部もラベリングされている．本実験では無声化母音を前音素の子音に，声門閉鎖部は後続の子音にまとめている．また，撥音は全て歯茎音として扱っている．ここで，分析窓は 25ms の Hanning 窓，分析周期は 10ms である．HMM の学習は，男声話者 MSH の ATR 音素バランス文 B セット計 50 文を使用した．AF-LSP 変換 MLN の学習には FANN[54]を用いた．学習文は，評価対象話者 MMY，MHT の音声から，評価に用いる 10 文を除いた 493 文である．MLN の学習には RPROP 法[55] を用いた．評価に用いる合成音声 10 文は ATR 音素バランス文のうち，話者 MMY と MHT の A01～A10 である．なお，今回の評価実験では，状態継続長とゲインを音声から直接抽出している．音声の有声・無声判定は原音声を用い，基本周波数の抽出には get f0s[56]を利用した．音源波形の生成には SPTK[57]の excite コマンドを利用した．

以下,「AF－HMM 学習」は HMM が出力した AF を入力して学習した MLN の出力 LSP 系列に対する評価を,「直接学習」は原音から抽出した AF 系列を入力して学習した MLN の出力 LSP 系列に対する評価を示している．

なお，本章では LSP フィルタの駆動音源には有声音にはパルス，無声音にはホワイトノイズを利用した．

表 5 実験に使用した HMM の仕様

HMM	monophone-HMM (38 音素), 7-state 5-loop, left-to-right
学習コーパス	ATR デジタル音声データベース（男声話者 MSH の音素バランス文 B セット, 50 文; ラベルデータから生成した AF）
特徴量	AF 15 次元× 3 フレーム（計 45 次元）

表 6 実験に使用した AF-LSP 変換 MLN の仕様

MLN	入力層 45, 中間層 450, 出力層 60(LSP 20 次元 3 フレーム)
学習コーパス	ATR デジタル音声データベース, ATR 音素バランス文（16bit, 16kHz） ・対象話者： MMY（男性, A01～A10 を除く 493 文） ・対象話者： MHT（男性, A01～A10 を除く 493 文）
特徴量	AF 15 次元× 3 フレーム（計 45 次元）

3.2.4.2 主観評価

MOS テストの結果を図 31 に示す．濃いグレーが話者 MMY，薄いグレーが話者 MHT の合成音に対する評価である．被験者は成人男性 10 名である．図 31 より，LSP 符号化音声の MOS 値は約 5 であった．AF-HMM 学習は MMY の場合約 3.5，MHT の場合約 3.3 であった．直接学習は，MMY の場合に約 1，MHT の場合に約 1.5 であった．AF-HMM 学習が直接学習よりも高い評価を得たが，LSP 符号化音声とは差が大きく，さらなる改良が必要である．

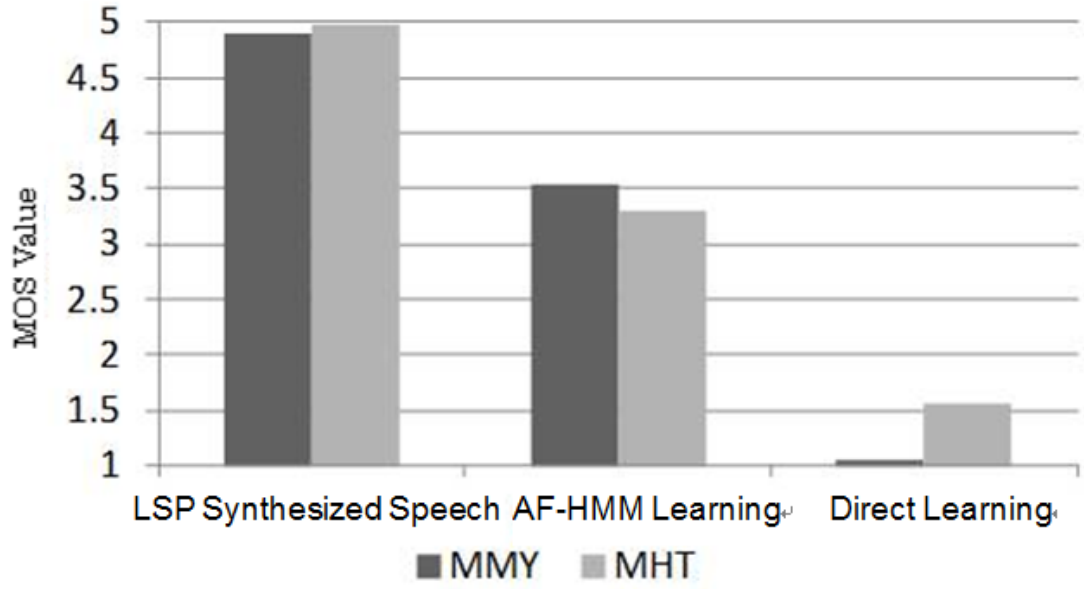


図 31 MOS テストの結果

3.2.4.3 客観評価

3.2.4.3.1 原音声から抽出した LSP との相関

図 32，33 に，話者 MMY と MHT の原音から抽出した LSP 係数列と$\mathrm{MLN}_{\mathrm{AF-LSP}}$から出力された LSP との相関を示す. ここでは A01～A10 の 10 文の相関値の平均を用いた. なお，無音部分は計算から除外している．両話者共，AF-HMM 学習法は LSP 係数 13 次前後まで 0.7 ～0.9 程度の強い相関を示している．しかし，高次になると徐々に低下している．これは高次になると MLN による平滑化が起こり，正しい値を得られなくなる為と考えられる．今回の実験では LSP 14 次以上は，5kHz 以上の帯域に相当するため，音声知覚にはあまり影響しないものと思われる．

AF-HMM 学習と直接学習を比較すると，AF-HMM 学習の方が直接学習より全体的に相関が高い．原音声学習の相関が低い部分は，MMY では低周波で顕著であり，MHT では LSP9 次から 11 次の部分で顕著である．MMY の直接学習の MOS 値が低い理由も，低周波での LSP の相関の低さが影響しているものと推定される．AF-HMM 学習ではこのような低下が起こらず，話者依存性もあまり見られない．図 34～36 に MMY の音素バランス文の音声について，1 次，13 次，17 次の際の LSP 系列の時間変化を示す．縦軸は各次数の LSP 周波数，横軸は時間を示す．音声は ATR 音素バランス文の A01 である．実線が原音から抽出した LSP，破線が AF-HMM 学習により得られた LSP 系列，点線が直接学習により得られた LSP 系列である．1 次，13 次では，AF-HMM 学習で得られた LSP は原音抽出の LSP にある程度追随している事がわかる. これに対して 17 次ではあまり追随できていないことが見て取れる．これは，図 32，33 の相関係数の結果と一致している．同様に直接学習はどの例でも原音声抽出 LSP に追随できていない.

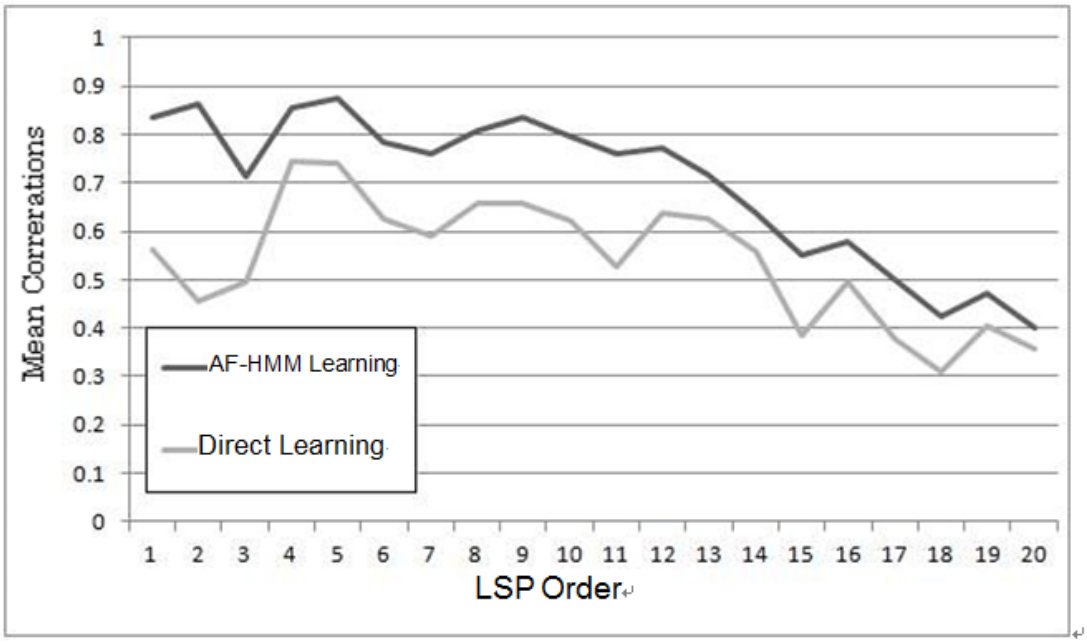


図 32 原音から抽出した LSP と AF―HMM 学習法，直接学習法で生成した LSP との相関係数の A01～A10 までの 10 文の平均．(MMY)

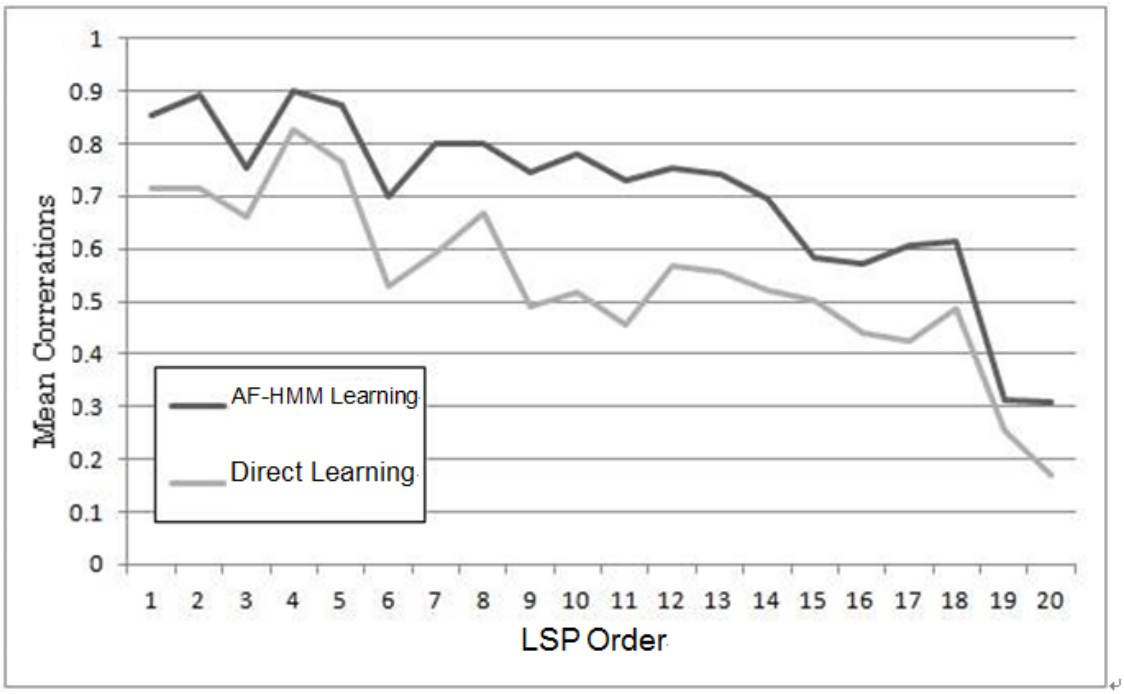


図 33 原音から抽出した LSP と AF―HMM 学習法，直接学習法で生成した LSP との相関係数の A01～A10 までの 10 文の平均．(MHT)

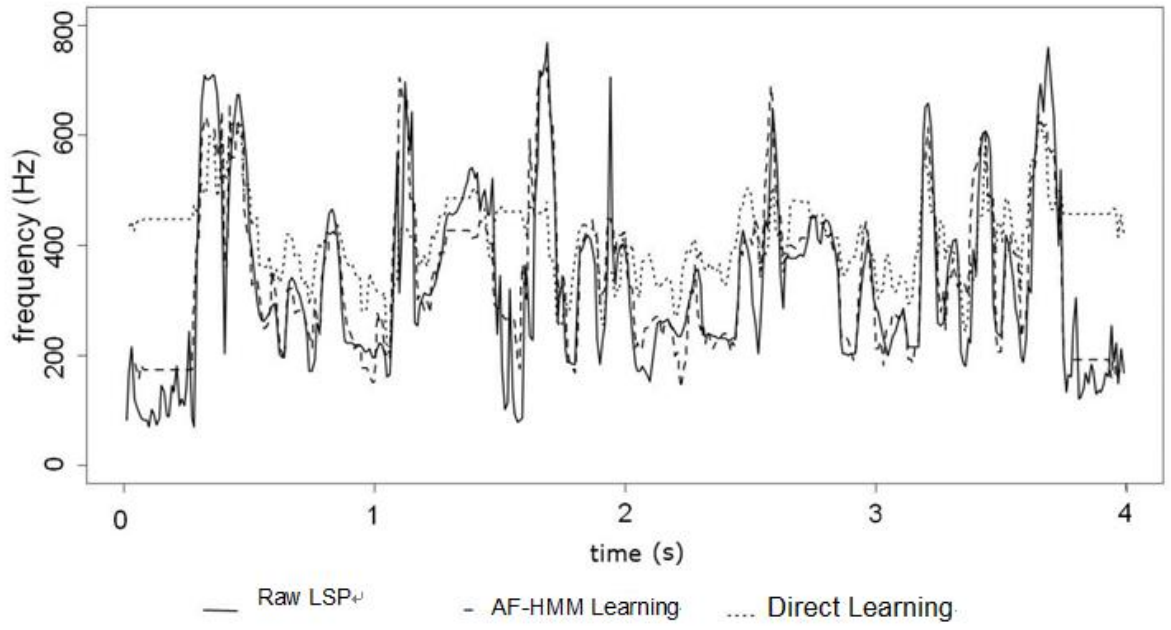


図 34 LSP 係数の時間変化(1 次) 話者:MMY, 音声:A01

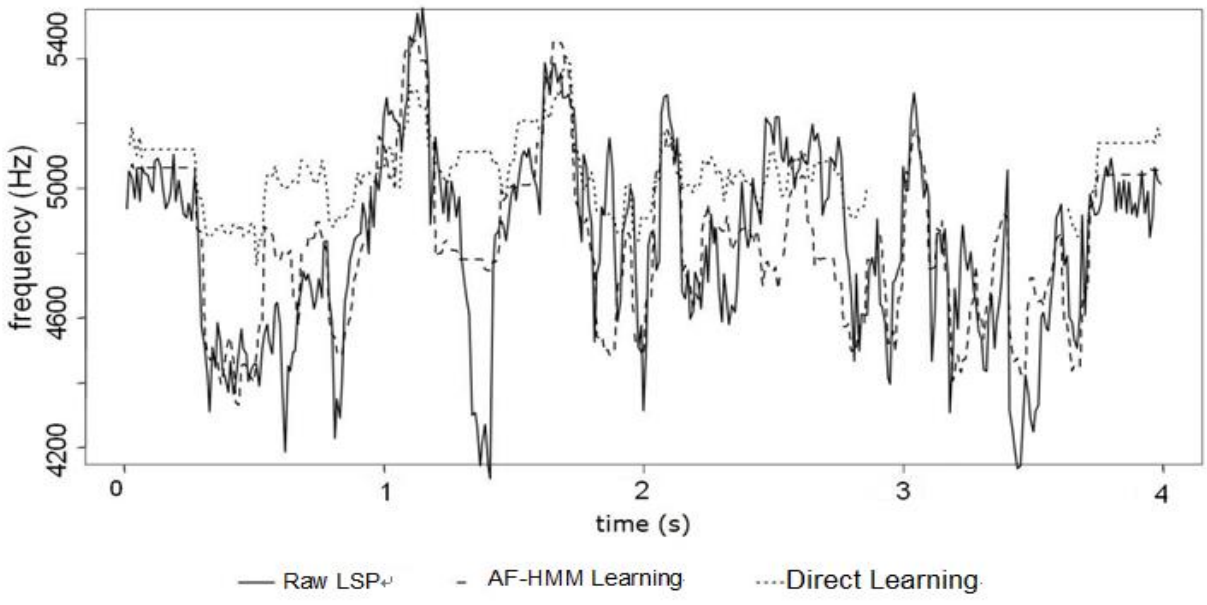


図 35 LSP 係数の時間変化(13 次). 話者:MMY, 音声:A01

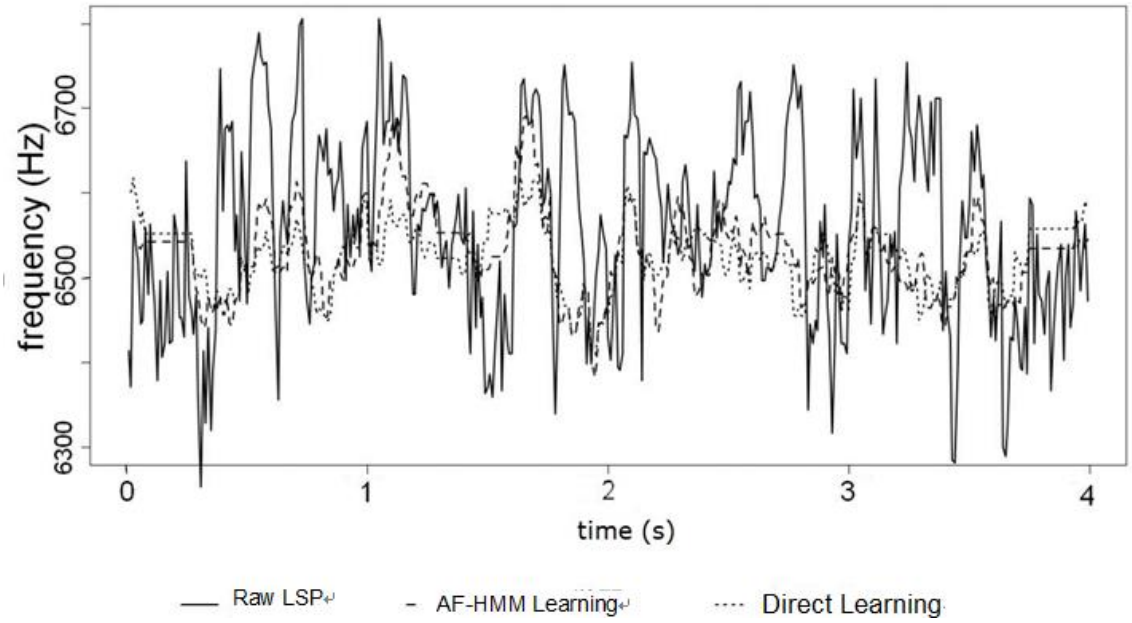


図 36　LSP 係数の時間変化(17 次). 話者:MMY, 音声:A01

3.2.4.3.2 スペクトル歪みの比較

生成した合成音声と原音声間のスペクトル歪み(SD) を式(24) から計算した．話者 MMY と MHT の結果をそれぞれ図 37，38 に示す．縦軸がスペクトル歪み(dB)を，横軸が各音声の平均スペクトル歪みを表している．横軸の最後の項目は全評価データに対する平均スペクトル歪みを示す．式(24) の$\mathbf{W}_{\mathrm{syn}}$ と$\mathbf{W}_{\mathrm{org}}$は，それぞれ合成音声，原音声の対数振幅スペクトル，L はフレーム番号，K はスペクトル番号である．計算では無音部分を評価から外している．

$$\mathrm{SD} = \sqrt{\frac{1}{\mathrm{L}}\sum_{\mathrm{l}=0}^{\mathrm{L}}\frac{1}{\mathrm{K}}\sum_{\mathrm{k}=0}^{\mathrm{K}}\left(\left|W_{syn} - \mathrm{W}_{\mathrm{org}}\right|\right)} \quad \text{..... (24)}$$

MMY の場合，10 音声の平均 SD 値は LSP 符号化音声は約 0.8dB，AF―HMM 学習では約 1.53dB，直接学習は約 2.26dB であった．また MHT では，LSP 符号化音声は 0.44dB，AF－HMM 学習では 1.35dB，直接学習では 1.60dB であった．AF―HMM 学習は，LSP 符号化音声と比較すると，歪みが大きく，改善の余地があると考えられる．AF-HMM 学習において，MHT と MMY であまり差は見られなかった．

図 39~41 にそれぞれ，話者 MMY の原音声，直接学習，AF-HMM 学習による合成音声のスペクトログラムを示す．発話内容は ATR 音素バランス文 A01 冒頭部分の「あらゆる現実を」である．図からスペクトルが滑らかに変化していく様子が見て取れる．HMM から出力される AF は平均ベクトルを用いているため階段状に変化する．しかし、合成音声のスペクトルは滑らかに変化することがわかる．これは，$\mathrm{MLN}_{\mathrm{AF-LSP}}$ に注目フレーム t と前後フレーム t±3 の AF を同時に入力したことで，$\mathrm{MLN}_{\mathrm{AF-LSP}}$が LSP の時間変化を学習しているためである．

しかし，フォルマントピークが過度につぶれており，これが主観的な音質の劣化に影響しているものと考えられる．

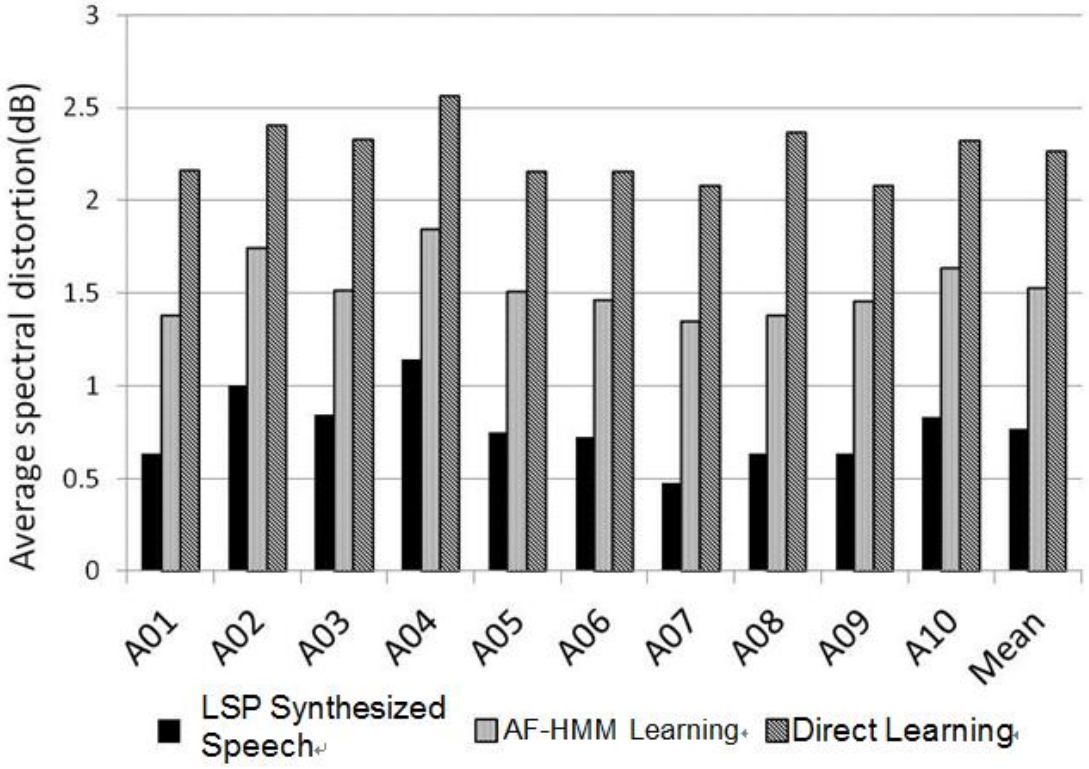


図 37 スペクトル歪み 話者: MMY

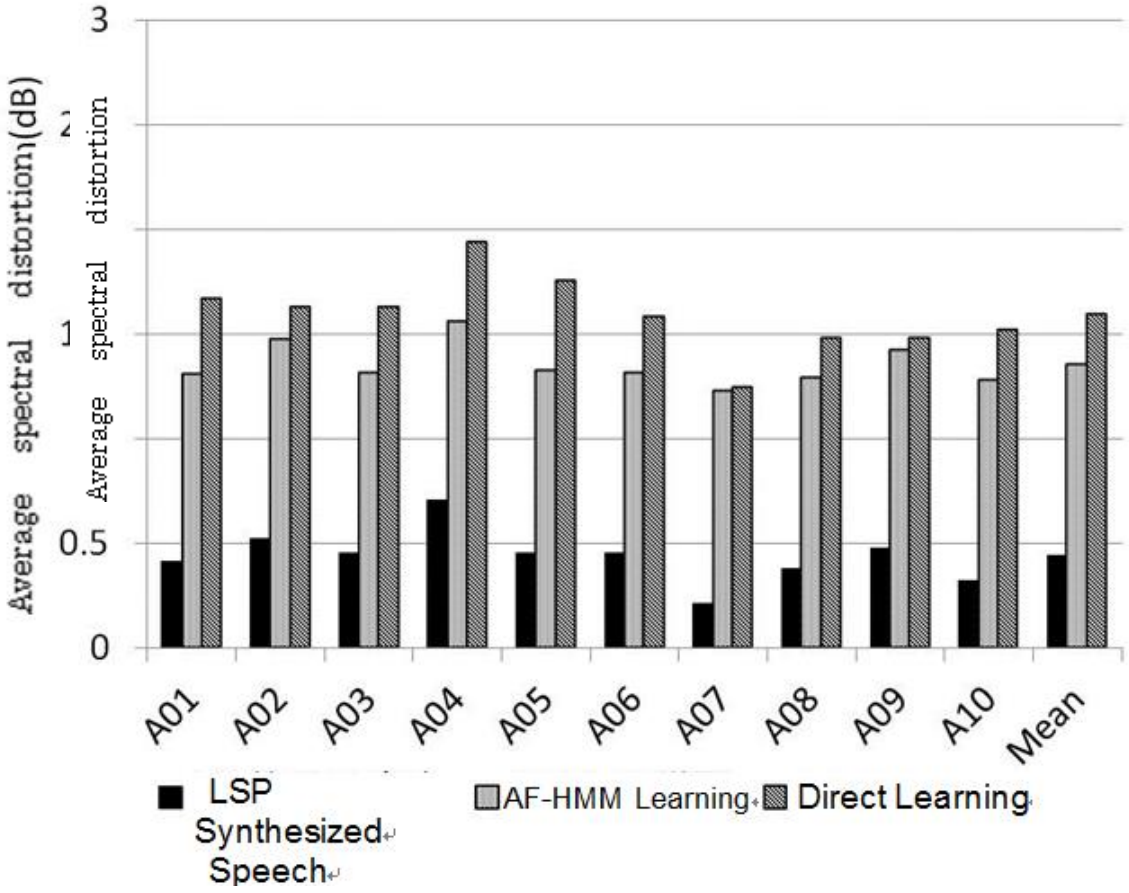


図 38 スペクトル歪み 話者: MHT

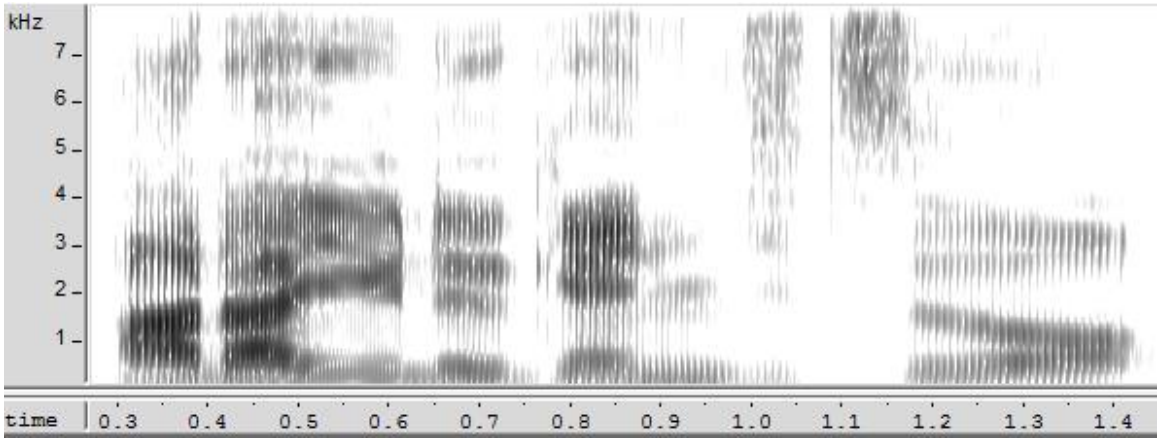


図 39 原音声のスペクトログラム
（話者: MMY，発話内容：「あらゆる現実を」）

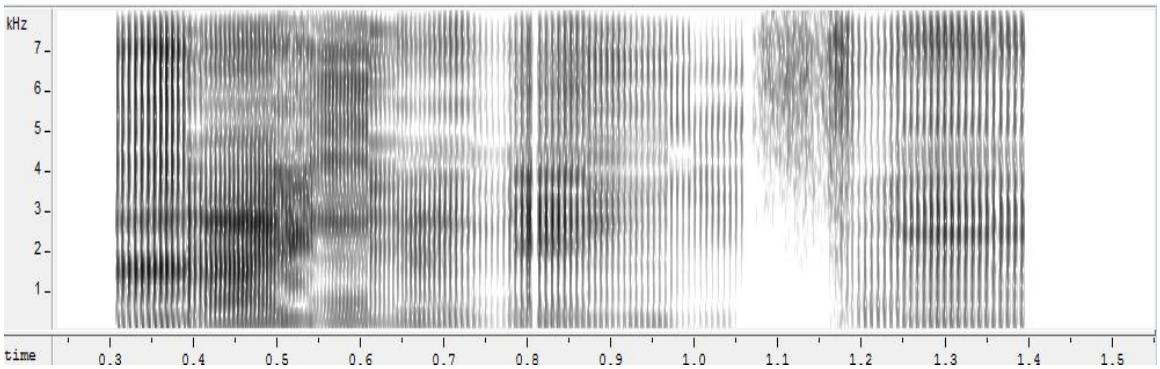


図 40 直接学習のスペクトログラム
（話者: MMY，発話内容 :「あらゆる現実を」）

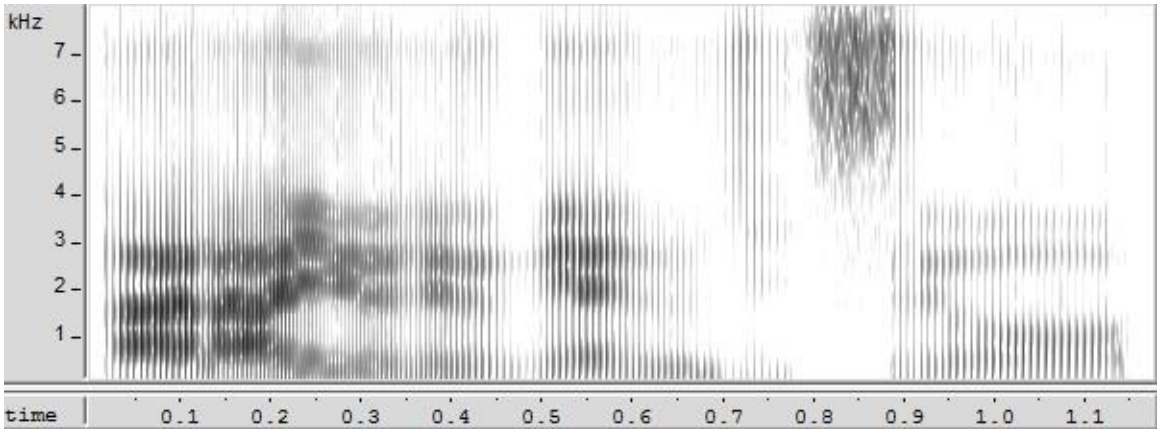


図 41 AF-HMM 学習による合成音のスペクトログラム
（話者：MMY，発話内容:「あらゆる現実を」）

3.3 まとめ

本章では，話者共通の調音特徴から，話者固有の AF-LSP 変換器を用いることによって，特定話者の音声が合成できることを示した．また，AF-LSP 変換器の学習に，AF 抽出機の出力を用いる直接学習よりも，話者固有の HMM から出力された AF 列を用いる AF-HMM 学習のほうが主観，客観評価ともに高い性能を示した．これは，AF-HMM 学習のほうが直接学習よりも，実施時の入力 AF 値に近いためである．

本手法の改良法としては，AF-LSP 変換器に深層学習（Deep Learning）を用いることが考えられる．深層学習は音声や画像等の分野で他手法を上回る性能を示しており，音質向上が期待できる．

4 章 駆動音源の改良

3 章で述べた HMM 音声合成方式では, デジタルフィルタの駆動音源にパルス列と白色雑音を用いるため，これによる音質劣化が問題となる．本章では，駆動音源に線形予測係数の残差信号を利用することで合成音声の品質向上を目指す.

すでに小池ら[58]は、残差駆動による HMM 音声合成法を提案し，少量の残差データベースから主観評価による明瞭性や総合評価で従来法を上回る結果を得ている．この手法は HMM 音声合成の枠組みの中で，音響特徴としてメルケプストラム系列を用い，残差波形を接続して音源を構築している．残差選択には分析時のケプストラム間の距離が最小になるような尺度を用いている．しかし，一般にフィルタとしての声道形状と声帯振動は相互に独立していると考えられており，必ずしももとの音声に近づくという保障は無い．そこで，学習用音声コーパスの中で，CELP(Code Excited Linear Predicition)符号化[59]を行ない，残差符号インデックスを HMM の各状態に登録することで、高品質化を図る．図 42 に音源改良手法のシステム図を示す．それぞれの部分について次節以降で詳しく述べる.

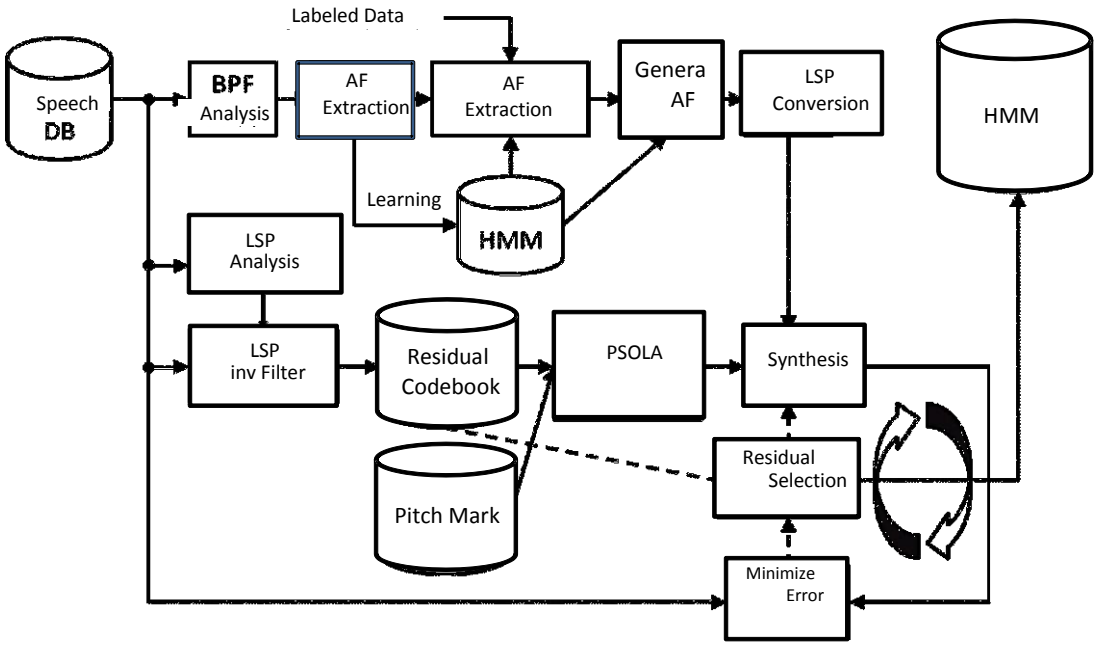


図 42 音源改良手法のブロック図

4.1 CELP 符号化[59]

CELP 符号化は，電話帯域の低ビットレート音声符号化の主流であり，人間の発声機構を音源成分とスペクトル包絡成分に分離してモデル化する vocoder 方式に属する．二つの成分を合成フィルタに供給して音声を生成する際，駆動音源成分を符号帳から探索し，入力波形に最も近いものを決定する．A-b-S（Analysis by Synthesis）法に基づく閉ループ探索を実装したことで，高音質音声符号化を実現している．

符号化の流れを図 43，及び，以下に示す．

1. 残差符号帳を構築する．
2. 符号化器において，入力音声を声道パラメータに変換を行い，残差符号帳内の残差素片の組み合わせで構成された音源と逐次合成を行う．
3. 音声波形レベルでの誤差が最小となる残差の組み合わせを選択する．
4. 選択された残差のインデックスと声道パラメータを復号器へ伝送することで，音声の再合成を行う．

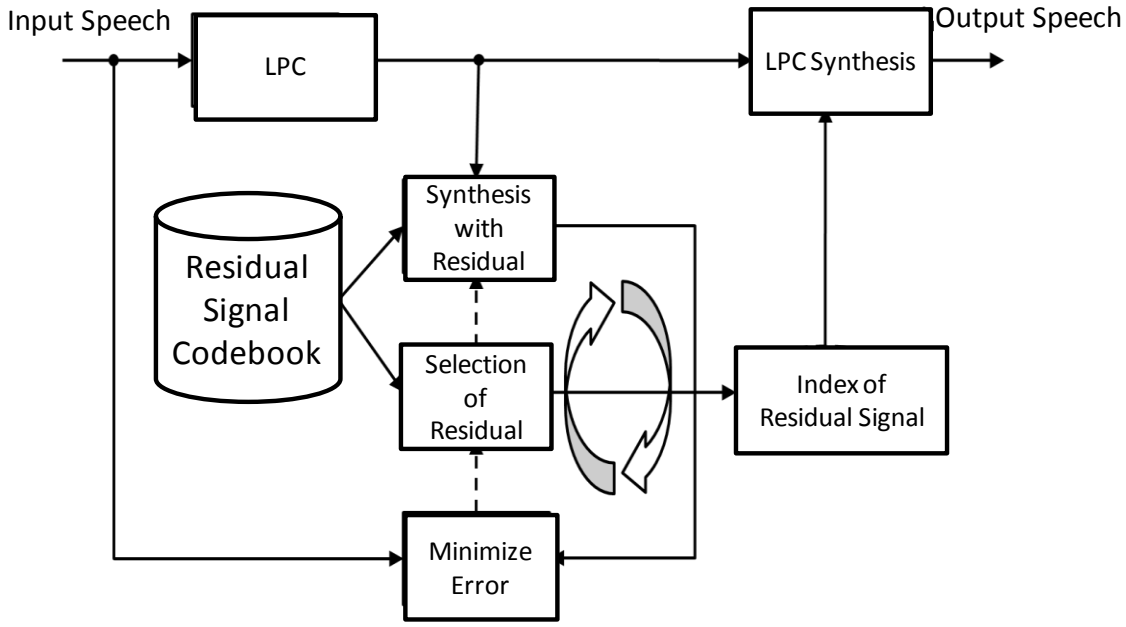


図 43 CELP 符号化の流れ

4.2 残差符号帳

駆動音源を構築する残差素片のサンプル数は，出来るだけ多いほど，高音質となる．しかし，前述の残差素片選択アルゴリズムをそのまま適用すると，計算量は膨大になり，現実的ではない．そこで大量の残差素片に対して，音韻環境クタスタリングと LBG（Linde-Buzo-Gray）アルゴリズム[60]によるクラスタリングを施すことで，残差素片の二分木を構築し，実際に誤差比較に用いる残差素片の数を減少させる手法を用いた．こ

こで，残差素片間の距離を時間領域で考えると，位相の違いと残差素片の窓幅の違いによる問題が発生してしまう．これらの問題を解決するため，残差素片を一定の窓幅（512点）で FFT 分析を行ない，周波数領域で残差素片間の距離を測った．

自然音声から残差波形を抽出し，有声区間のピッチマークに基づいた窓掛けにより，残差素片の集合である残差符号帳を作成する．ピッチマークは，線形予測分析によって抽出された残差波形に対して，有声区間のピッチパルスの位置を目視によって付与することで得た．

続いて，ピッチマークを付与した残差波形から，ピッチマークを中心に基本周期の約 2 倍の領域を抽出し，一つの残差素片とする．こうして得た残差素片をデータベース化し，残差符号帳を構築する（図 44）．

そして，残差符号帳の生成に音韻環境クラスタリングを導入することで，前後の音韻環境を考慮した合成音声生成を試みた．

音声信号の物理的特徴量を統計的に処理することで，最適な合成単位の生成を行う．なお，今回は，直前，及び，直後の音韻環境のみに注目したため，生成されるラベルは triphone のように扱うことが可能となる．

音韻環境クラスタリングの手順を以下に示す．

1. 該当するクラスタにおいて選択可能な直前または直後の音韻環境すべてに対して，その音韻環境を含むものと含まないものでクラスタを分割し，分割したふたつのクラスタにおける分離度を求める．
2. 分離度が最大となる音韻環境を選択し，その音韻環境を含むものと含まないものでクラスタを分割し，子クラスタとする．
3. クラスタの要素数が一定数以下になる，または 1. においてクラスタの分離が不可能になるまで 1. 及び 2. を繰り返す．

なお，音韻環境クラスタリングの導入にあたり，従来手法では有声子音と無声子音をまとめてひとつの初期クラスタとしていた．本論文では同一の音素に対して前後の音韻環境に注目するため，各音素に対してひとつずつ初期クラスタを生成した．

また，本論文では分離度が最大になる音韻環境を選択するにあたり，クラス内分散が最小となるものを選択する方法をとった．クラス内分散 G_w は式(25) によって導出する．ただし，N は参照しているクラスタの要素数，N_c は c 番目の分割クラスタの要素数，$G_{c,n}$ は c 番目の分割クラスタの n 番目の要素のパワースペクトル，$G_{c,av}$ は c 番目の分割クラスタの平均パワースペクトルである．

$$G_w = \frac{1}{N-2} \sum_{c=1}^{2} \left(\sum_{n=1}^{N_c} \| G_{c,n} - G_{c,av} \| \right) \quad \text{(25)}$$

しかし，音韻環境クラスタリングのみでは HMM への残差信号割り当てに適用するこ

とができない．そこで，音韻環境クラスタリングによって選択された各音韻に対して，従来手法と同様にLBGアルゴリズムによるクラスタリングを実行し，二分木を構築する．残差素片の二分木の構築手順を図 45 と以下に示す．その際に，音韻環境クラスタリングによる決定木の各末端のクラスタの要素数が少ない場合，要素数が一定数以上になるまで親クラスタを遡り，LBG アルゴリズムに適用するクラスタを選択する．

1. 得られた残差素片に対して FFT 分析を行い，振幅スペクトルを得る．
2. 得られた振幅スペクトルのユークリッド距離を残差素片間の距離尺度とし，LBG アルゴリズムを適用してクラスタリングを行い，その過程で二分木を構築する．
   （a）残差素片グループのセントロイド（重心）を計算し，重心に位置する残差素片（以下，重心素片とする）を二分木のルートノードとして登録する．
   （b）LBG アルゴリズムにより，残差素片グループを 2 つに分割する．
   （c）2 つの残差素片グループのセントロイドを計算し，各グループの重心素片を，分割前グループの重心素片の子ノードとして登録する．
   （d）全ての残差素片グループの中で，最も多くの残差素片を有しているグループに対して，LBG アルゴリズムを適用し，残差素片グループを 2 つに分割する．
   （e） 残差素片の数だけ，(a)～(d).の操作を繰り返す．

ここで、残差符号帳は 1 つではなく、図 46 に示すように各音素に対応した残差符号帳（母音 5 種＋撥音＋有声子音＋無声子音の計 8 種類）を持たせるようにした．これにより，割り当ての際に誤った音素の部分から抽出された残差素片が選択されることを防ぐことができる．

本論文では，精度の影響を鑑み，目視によるピッチマーク付与を行なった．しかし，基本的には，基本周期ずれた区間内の最大振幅の点を求めるに過ぎない．そのため，ピッチ抽出などを利用することによって，ピッチマーク抽出の自動化を行うことも可能であると考えられる．

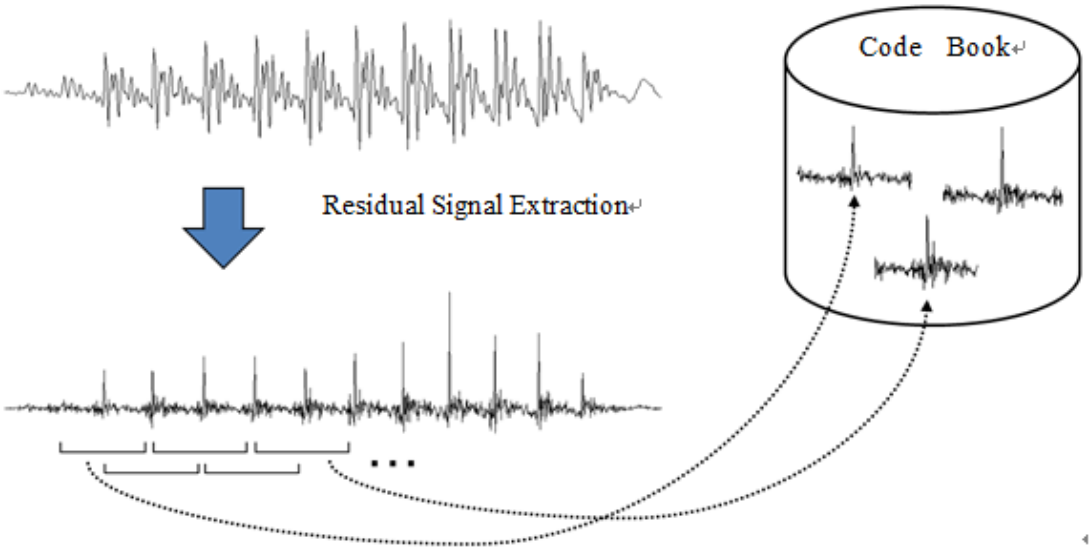


図 44 残差符号帳の作成方法

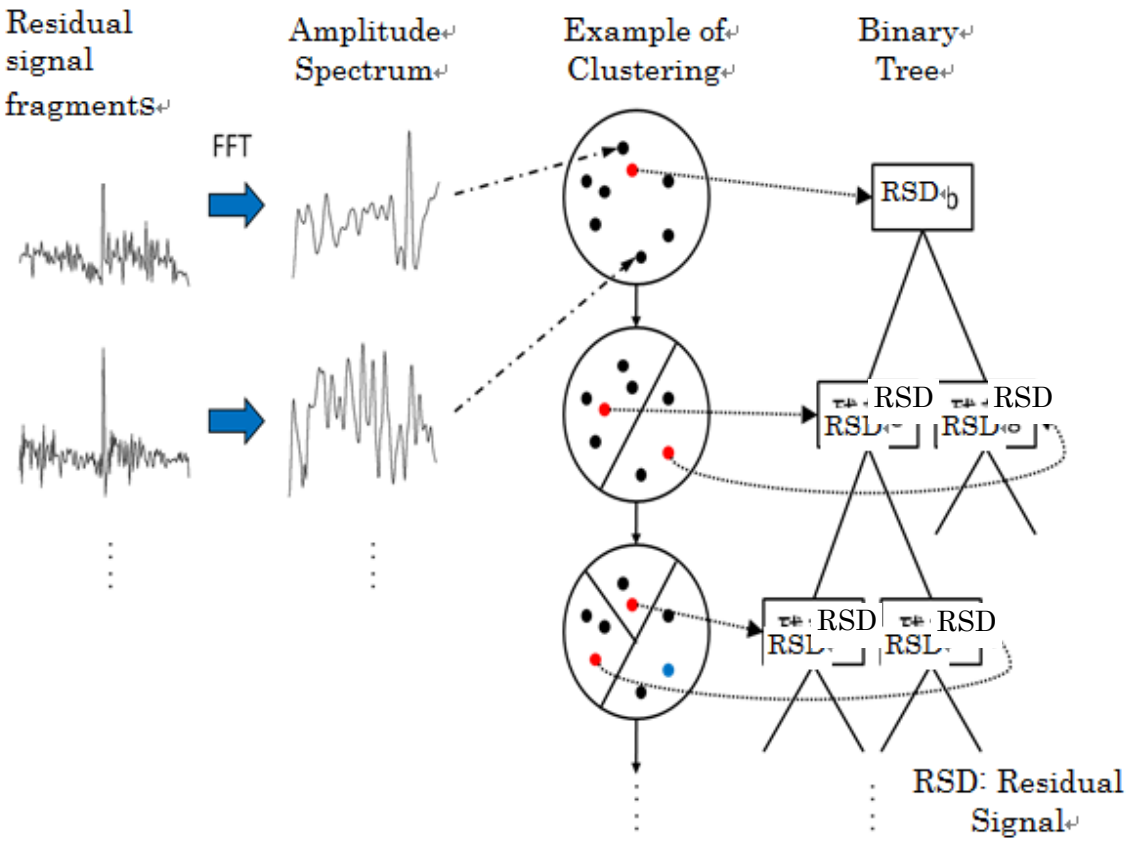


図 45 LBG クラスタリングによる

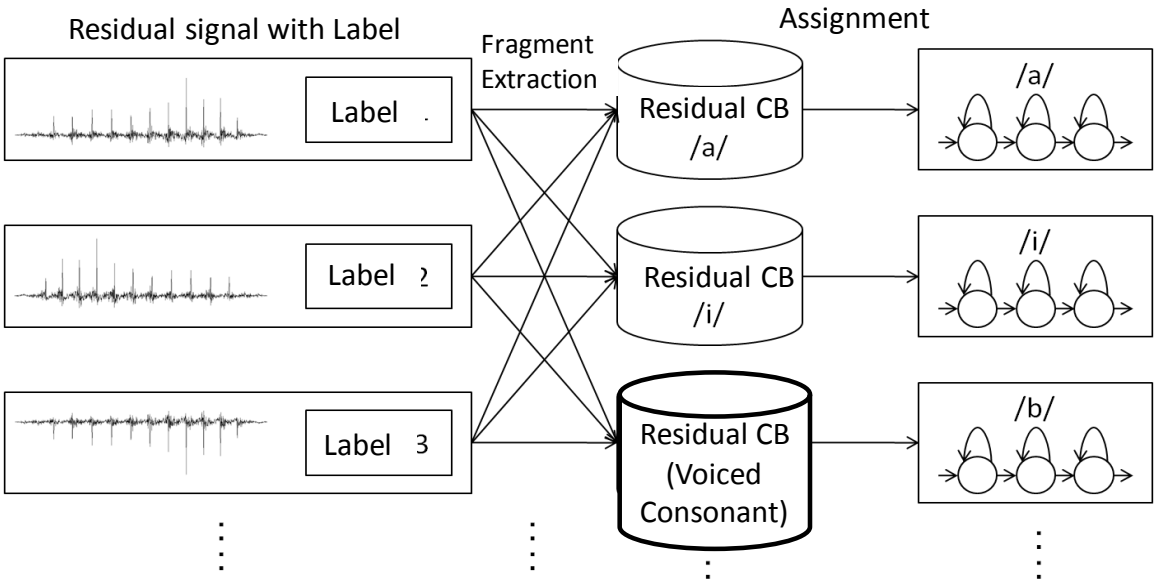


図 46 複数の残差符号帳

4.3 CELP 方式による駆動音源の改良

提案手法では，通常の音声合成 HMM を学習する過程に加え，学習データから抽出した残差素片を CELP 符号化に基づく閉ループ探索を適用して，すべての HMM の各状態に割り当てる過程が必要となる．その手順を図 47 及び，以下に示す．

1. 割り当てたい音素 HMM の各状態から得られた AF（平均ベクトル）から，LSP 係数を抽出する．
2. 得られた LSP 係数と残差符号帳内の残差素片を合成し，閉ループ学習による残差素片選択を行う．その際，予め付与したピッチマークを用いて，元の音声とピッチパルスの位置を合わせておく．
   a. 残差素片帳二分木のルートの子ノードに当たる 2 つの残差素片に対してそれぞれ音声合成を行い，誤差の小さい方を選択する．
   b. 選ばれた残差素片の子ノードに当たる 2 つの残差素片に対して同様の処理を行い，リーフノードに到達するまで繰り返す．
3. 最終的に，誤差が最小となる残差素片を HMM のある状態に割り当てる．

ここで，前後音素を考慮した残差素片選択を行い，各音素の HMM に，前後音素によって異なる最適な残差素片を複数持たせるようにした．これにより，滑らかな音源を実現することができる．

また、各音素の HMM の状態数（始終端状態含む）は，従来の 5 状態ではなく 7 状態とした．AF は，音素境界部で急激に変化するため，過渡部に当たる状態数を増やすことで，より滑らかな音源を実現することができる．

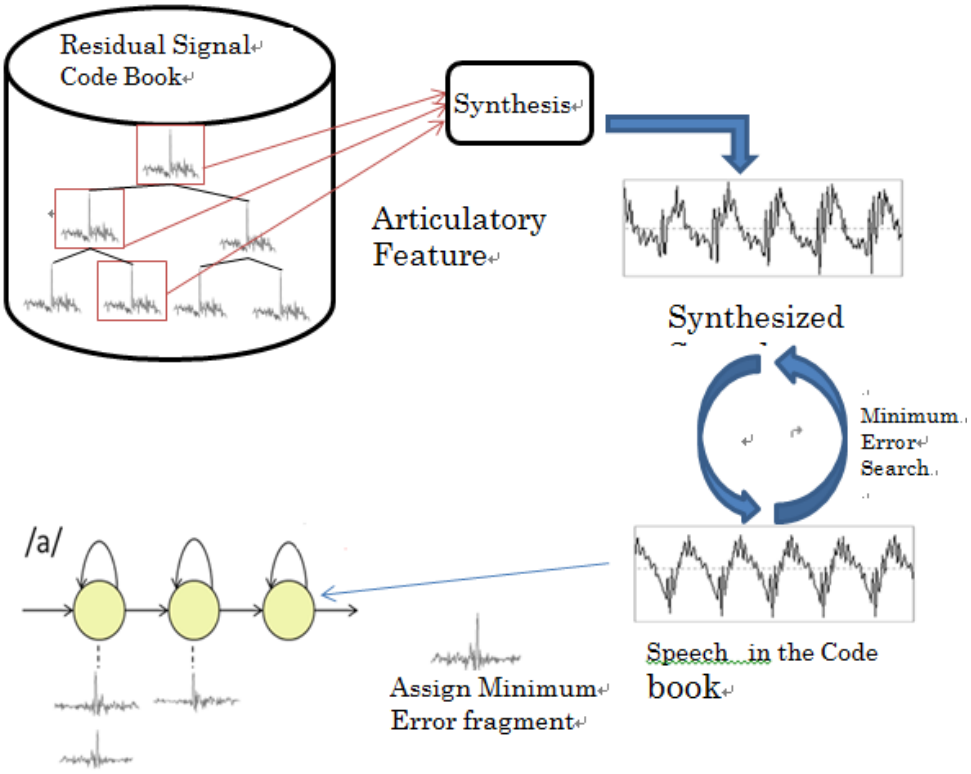


図 47 HMM への残差符号割り当て

4.4 PSOLA 法に基づく駆動音源の生成

本論文で使用する駆動音源は，HMM の各状態に割り当てた残差素片を PSOLA 法[61]によって重ね合わせ作成した残差信号を用いる． PSOLA とは，音声信号のスペクトルをある程度保ったまま ピッチや持続時間を変更可能な音声処理のためのデジタル信号処理手法であり，音声合成時には音高変化や継続時間の調整を行いやすいという特徴がある．

PSOLA では，図 48 のように音声波形を基本周期と同期した分析窓で、互いにオーバーラップした短い断片に分割して再配置する．

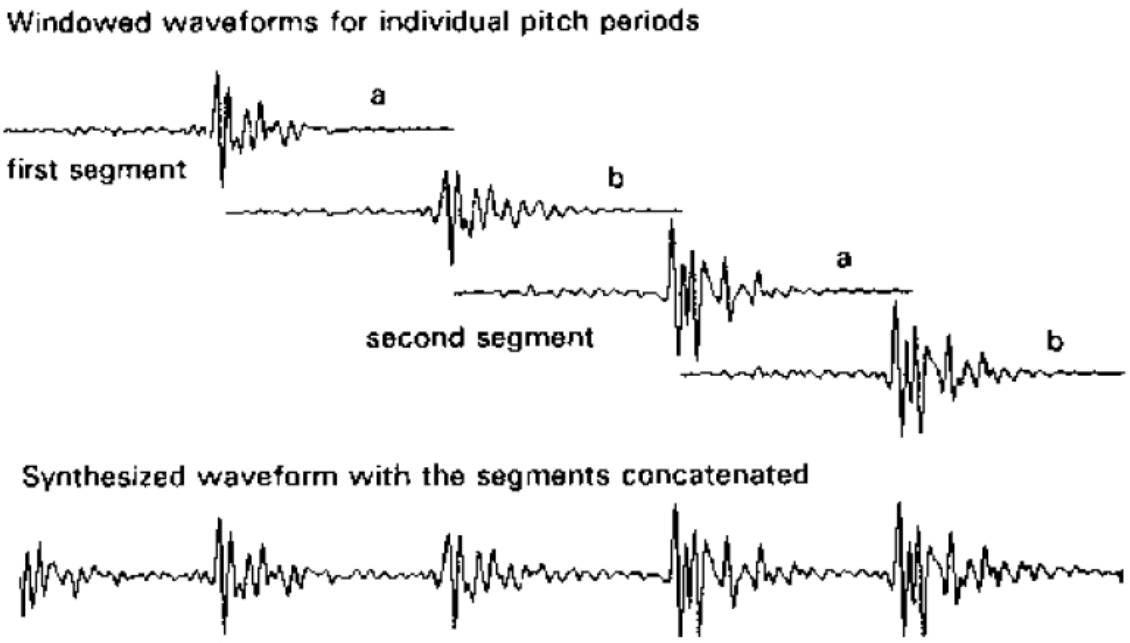


図 48 PSOLA

4.5 評価実験

今回の音声合成システムで使用したHMMの仕様を表7に，AF-LSPパラメータ変換器を表8に，CELP符号化で使用した残差符号帳を表9に示す．ここで，分析窓長は25msのHanning窓，分析周期は10msである．また，今回の実験では，基本周波数と状態継続長を，音声から直接抽出している．

表 7 実験で使用した HMM

HMM	monophone-HMM（38音素）, 7-state 5-loop, left-to-right
学習コーパス	JNAS（男性38名, 5000文; 16bit, 16kHz）
特徴量	AF 15次元×3フレーム（計45次元）

表 8 実験で使用した AF-LSP パラメータ変換器

MLN	3層（入力層45, 中間層450, 出力層42）
学習コーパス	ATR音素バランス文（16bit, 12kHz） ：MHT（男性, 493文）
入力	AF15次元×3フレーム（計45次元）
出力	LSPパラメータ14次元×3フレーム（計42次元）

表 9 実験で使用した残差符号帳

学習コーパス	ATR音素バランス文（16bit, 12kHz） ： MHT（男性, 150文）
残差素片数	7653個

図 49 に主観評価値（MOS : Mean Opinion Score） の値を示す．左から原音声，駆動音源にパルスとノイズを利用した物，符号化合成音，提案手法である．

本手法で合成した音声の明瞭さについて，5 段階（5：良い～1：悪い）の主観評価値を記述させた．被験者は 14 名である．

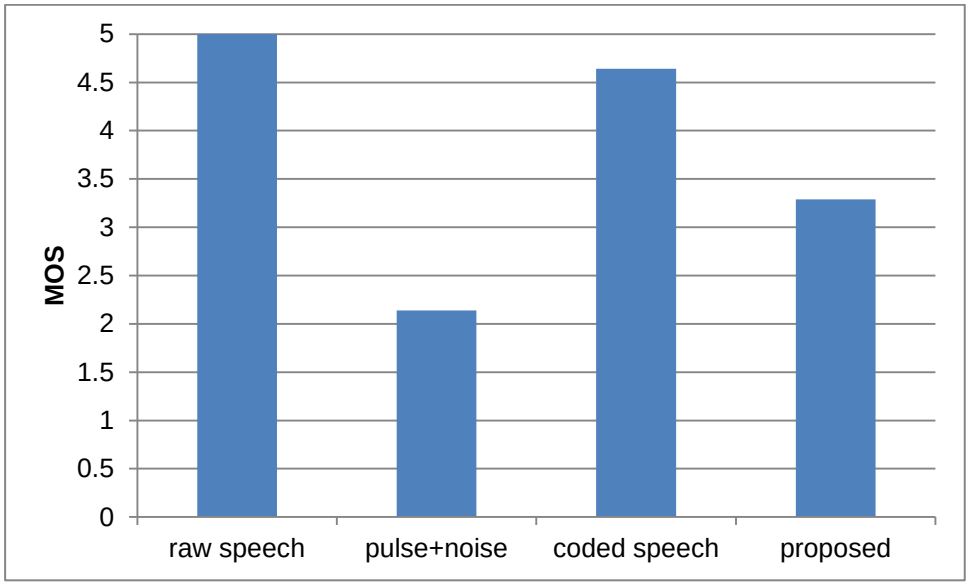


図 49 スペクトル歪みの改良による性能の比較

4.6 考察

図 49 の実験結果から，提案手法は従来の駆動音源にパルスとノイズを使用するものよりも高い値を示すことがわかる．しかし，符号化合成音とは差があり，さらなる改良が必要である．具体的には，駆動音源符号帳の作成方法の改良や，コンテキストを考慮した駆動音源の選択などが考えられる．LBG クラスタリングでは残差素片選択に音素コンテキストを考慮していない．駆動音源を選択する際に，トライフォンなどの音素の前後関係を考慮することによって性能が向上する可能性がある．

今後の課題として，通常の CELP 方式では，素片を選択する際に以前に連結した数フレームの素片を覚えておき，結合した際に歪みが小さくなる素片を当てはめている．現在の手法では，このような過去フレームの素片をとの関係性を考慮しておらず，それが聴感上の違和感を生じさせるために主観評価値が悪くなっている可能性がある．

本手法では F0 のモデル化を行っておらず，元となる音声波形から F0 の抽出を行っている，歌声合成などへの応用を考える場合は F0 が楽譜から抽出できるため問題とならないが，TTS（Text To Speech）を行う際には実装の必要がある．

ほかには利得符号帳の実装も課題である．合成音声の利得は F0 と同じように元音声を参照している．TTS を行う際に適切なイントネーションのモデル化が必要となる．

これらの改良によって音質の向上や歌声や TTS への応用が可能になる，

4.7 まとめ

本章では，音声から調音特徴系列を抽出して HMM を構成すると共に，HMM が生成する調音特徴系列を声道音響パラメータに変換することで音声を合成する方式について提案し，合成音の品質評価結果を述べた．今後は，AF-LSP 変換器のさらなる改良を行うと共に，駆動音源を実際の残差信号を用いて，音源符号帳から設計する方法を検討する．これらの改良を通して，ワンモデル音声認識合成器を作成しタスク推定法の性能改善を図りたい．

5章　結　論

本論文では，音声の合成と認識を共通のモデルで行うマルチモーダル音声対話システムのうち音声合成に関する研究と，ロボットが周囲で行われているタスクを推定するためのタスク推定に焦点をあてた研究を行った．音声合成については，人間の音声認識と合成が 1-model であるという仮説に基づいている．このような方式では，モデルの認識誤りを合成部で出力された音声との間で何らかの距離尺度を用いて修正できる可能性があり，将来には「合成による認識(Recognition by Synthesis: RbS)」を実現できる可能性がある．また，ロボットが人間と共生するために必要な，人間がそのときに従事しているタスクを推定するマルチモーダルタスク推定手法を提案した．これにより，人間が従事しているタスクをロボットが推定できるようになれば，具体的なタスクの支援を行うことができるようになると考えられる．以下，本論文の各章の成果についてまとめる．

２章ではマルチモーダルタスク推定手法を提案した．これは，将来のロボットが人間の行っている多様なタスク遂行をサポートする際に必須の技術である．タスク遂行中に現れた発話単語と画像オブジェクトから作成した行列をもとに LSA を行うことで，遂行中のタスクを推定する手法を提案した．実験では，机上で行うゲームタスク，ポーカー，大富豪，ブラックジャック，まわり将棋，詰将棋に対して本手法を適用し，タスク開始から 40 秒ほどで 90%の推定率が得られることを示した．提案手法は，タスク遂行中の発話単語と，使用したオブジェクトのみを用いており，動作の推定などの複雑な処理を行っていない点に利点がある．人が直接従事するタスクを推定することは，エージェントが人と対話する際に重要な情報となる．一方，現在の音声認識器の性能では，未だキーワードの抽出制度が不十分であるため，このような応用のためにはより高性能な音素認識器が必要であることを実験から示した．近年，Deep Learning の発展により，音素認識率は飛躍的に向上しており，タスク推定への応用も可能になることを期待する．

３章では，認識部と共通の調音特徴を用いた HMM モデルから音声合成が可能なことを示した．合成音声と原音声の相関値の平均の比較から，提案する AF-HMM 学習法は LSP 係数 13 次まで 0.7 ～0.9 程度の強い相関を示した．しかし，高次になると徐々に低下している．これは高次になると MLN による平滑化が起こり，正しい値を得られなくなる為と考えられる．今回の実験では LSP 14 次以上は，5kHz 以上の帯域に相当し，音声知覚にはあまり影響しないものと思われる．AF-HMM 学習では合成音の話者依存性もあまり見られなかった． 13 次以下の部分では，AF-HMM 学習で得られた LSP は原音抽出の LSP にある程度追随している事がわかった．これに対して 17 次ではあまり追随できていないことが見て取れ，改良の余地が残っている．これには，AF-LSP Converter に Deep Neural Network を用いる手法などが考えられる．また，本論文では基本周波数のモデル化について，考慮されておらず，既存の基本周波数モデル化手法との相性について調査する必要がある．

また，４章の駆動音源の改良では，CELP 方式による駆動音源の選択により，合成音声の主観評価値が向上することを示した．ここでは，音声から調音特徴系列を抽出して HMM を構成すると共に，HMM が生成する調音特徴系列を声道音響パラメータに変換することで音声を合成する方式について提案し，合成音の品質評価結果を述べた．今後は，駆動音源のクラスタリング手法などの改良をめざしたい．具体的には音素コンテキストを利用したクラスタリング手法などが考えられる．そのほか，前フレームのコンテキストを利用した駆動音源の選択，CELP 方式の利得符号帳の実装などにより性能の改善が見込まれる．

最後に，今後の課題であるが，本論文の研究成果を応用することにより，将来的にワ

ンモデル音声認識合成システムが実現できる可能性がある．このようなシステムでは，音声認識の誤りを合成音声と比較して修正する，あるいは合成音声の誤りや歪みを認識結果から調整するなど，認識と合成を協調させたモデルの修正が可能となろう．これは人間の言語獲得過程とも類似しており，ロボット研究における音声言語獲得手法としての応用が可能である．ほかにも，人間と共生して多様なタスク遂行を支援するロボットの実現には，様々な能力が要求される．例えば，タスクを実際に遂行するためのマニピュレータの制御，自然な音声対話による指示などが挙げられる．本論文では，その一つである，人間が従事しているタスクの推定手法を提案した．人間が従事しているタスクが推定できれば，ロボットはそのタスクに特化したサブモジュールを利用し，具体的なタスクの遂行支援が行えるようになると考えられる．合成による認識のためには，合成音声の品質のさらなる改良が必要になると考えられる．また，タスク推定では，より高速にタスクの推定ができる方式が必要となる．

本論文に述べた，タスク遂行中に表れる発話単語と画像オブジェクトを使って人間が遂行中のタスクを推定する技術，話者不変の AF-HMM と話者固有の$\mathrm{MLN}_{\mathrm{AF-LSP}}$ を導入して音声を合成する技術により，近い将来，複数タスクに対応可能なマルチモーダル音声対話システ ムの実現が可能となることを期待する．

謝　辞

本論文は豊橋技術科学大学大学院工学研究科電子・情報工学専攻において行った研究の成果をまとめたものである．

本研究がこのようなまとまりのあるものとなったのは，ひとえに，研究当初以来終始ご懇切なるご指導と温かいご配慮を頂いた，豊橋技術科学大学新田恒雄名誉教授の賜であり，ここに深く感謝申し上げます．また，ご多忙な時間を割いてご討論頂き，貴重なるご意見を賜った同工学研究科情報・知能工学系　増山繁教授，堀川順生教授、中川聖一特任教授、桂田浩一准教授に深く感謝の意を表します．そして，日頃から惜しみないご支援，ご助力を頂いた豊橋技術科学大学知識情報工学系の教職員の皆様，ならびに桂田研究室の諸氏に感謝申し上げます．

なお，本研究の一部は，文部科学省 21 世紀 COE プログラム「インテリジェントヒューマンセンシング」，及び文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の援助を受けたことを付記し，深く謝意を表します．

参考文献

[1] J.L. Miller and P.D. Eimas, “Internal structure of voicing categories in early infancy,” Percept. Psychophys., 58, 1157-1167, 1996..

[2] A. M. Liberman and I. G. Mattingley, “The motor theory of speech perception revised,” Cognition, Vol. 21, pp.1-36, 1985

[3] G. Miller, “The science of word,” Scientific American Library, 1991.

[4] S.M Wilson, A.P. Saygm, M.I. Sereno, and M. Iacoboni, “Listening to speech activates motor areasinvolved in speech production,” Nat. Neurosci., 7, 701-702, 2004.

[5] 小野田高幸，桂田浩一，新田恒雄，“調音運動 HMM 音声合成における調音特徴—声道パラメータ変換と音源の改良,” 情処学音声言語処理研報, 2010-SLP-84(30), 2010.

[6] S. King and P. Taylor, “Detection of phonologicafeatures in continuous speech using neural networks,” Comput. Speech Lang., vol.14, no.4, pp.333-345, 2000.

[7] E. Eide,“Distinctive features for use in an automatic speech recognition system,” Proc. Eurospeech 2001, vol.III, pp.1613-1616, 2001.

[8] K. Kirchhof, “Combining acoustic and articulatory feature information for robust speech recognition,” Speech Commun., vol.37, pp.303-319, 2002.

[9] 新田恒雄, 武井匠, 木村優志, 桂田浩一, “調音運動 HMM に基づくワンモデル音声認識合成,” 情処学音声言語情報処理研報, Vol.2009-SLP-77, No.4, pp.1-6,, 200.

[10] 木村優志，小野田高幸，入部百合絵，桂田浩一，新田恒雄，“調音運動に基づくワンモデル音声認識合成への CELP 適用,” 第 24 会人工知能学会全国大会, 1J1-OS13-2, 2010.

[11] S. Sivadas, and H. Hermansky, “Hierarchical tandem feature extraction,” ICASSP’02, vol.I, pp.809-812, 2002.

[12] T. Fukuda, W. Yamamoto, and T. Nitta “Distinctive phonetic feature extraction for robust speech recognition,” Proc. ICASSP'03, vol.II, pp.25-28, 2003.

[13] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speech synthesis from HMMs using dynamic features,” Proc. of ICASSP1996, pp.389-392, 1996.

[14] 板橋秀一編著, “音声の分析,” 著: 音声工学, 森北出版, 2005, pp. 第 4 章,.

[15] 菅村昇, 板倉文忠, “線スペクトル対音声分析合成方式による音声情報圧縮,” 信学論, 64-A, 8, pp.599-607, 1981.

[16] 岩橋直人, “人と機械の共有経験を基盤とする言語コミュニケーションの計算機構とシンボルグラウンディングの階層性,” 人工知能基本問題研究会, Vol. 61, pp. 25-32, 2005.

[17] 松坂要佐, 東條剛史, 小林哲則, “グループ会話に参与する対話ロボットの構築築.,” 信学技報. SP, 音声, Vol. 102, No. 417, pp. 39-40,, 2002.

[18] 神田崇行, 石黒浩, 小野哲雄, 今井倫太, 前田武志, 中津良平, “研究用プラットホームとしての日常活動型ロボット robovie" の開発,”電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol. 85, No. 4,pp. 380-389, 2002.

[19] 土井利忠, 藤田雅博, 下村秀樹 “脳・身体性・ロボット(インテリジェンス・ダイナミクス),” シュプリンガー・フェアラーク東京, 2005.

[20] Robocup@home., “http://www.ai.rug.nl/robocupathome/”.

[21] A. A. Alatan, A. N. Ali , W. Wolf, “Multi-modal dialog scene detection using hidden Maarkov models for content-based multimedia indexing,” Multimedia Tools and Applications, Volume 14 Issue 2, pp. 137-151, 2001.

[22] G. M. John, S. Gauch, S. Bouix , X. Zhu, “Real time video scene detection and classification,” Infomation processing and management, Volume 35, Issue 3, pp. 381-400, 1999.

[23] C.-W. Ngo, T.-F. Ma , H.-J. Zhang, “Video summarization and scene detection by graph modeling,” IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, Issue 2, pp. 296-305, 2005.

[24] N. Inamoto , H. Saito, “Intermediate View Generation of Soccer Scene from Multiple Videos,” 16th IEEE International Conference on Pattern Recognition, pp. 713 - 716, 2002.

[25] E. Andre, G. Herzog , T. Rist, “On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER,” In: Proc. of the 8th ECAI, pp.449-454, 1988.

[26] Y. Gong, T. Sin, C. H. Chuan, H. Zhang , M. Sakauchi, “Automatic Parsing of TV Soccer Programs,” Multimedia Computing and Systems, 1995., Proceedings of the International Conference on IEEE Multimedia Computing and Systems, pp.167-174, 1995.

[27] S.-C. Chen, M.-L. Chyu, C. Zhang, L. Luo , M. Chen, “Detection of Soccer Goal Shots using joint Multimedia features and classification rules,” Reules”, Proceedings of the Fourth International Workshop on Multimedia Data Mining, pp. 36-44, 2003.

[28] J. Wang, E. Chng , C. Xu, “Soccer replay detection using scene transition structure analysis,” IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05), Vol. 2, pp. 433-436, 2005.

[29] T. S, Dumais, “Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval,” Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval, 1990

[30] T. D. a. R. H. S. Deerwester, “Indexing by Latent Semantic Analysis,” jornal of the american society for information science, vol. 41, no. 6, pp.391-407, 1990.

[31] M. Cascia La, S. Sethi and S. Sclaroff, “Combining textual and visual cues for content-based image retrieval on the world wide web,” In: Proceedings of the IEEE Workshop on content-based access of image and video libraries. pp. 24-28, 1998.

[32] M. A. Nascimento and A. L. Oliveira, “An Empirical Comparison of Text Categorization Methods,” String Proecssing and Information Retrieval. Lecture Notes in Computer Science. pp.183-196, 2003.

[33] J. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” Proceedings of the IEEE, vol. 88-8, pp. 1279‑1296, 2000.

[34] T. Westerveld, “Image Retrieval: Content versus Context,” In Content-Based

Multimedia Information Access, RIAO 2000 Conference, pp.276-284, 2000.

[35] R. Zhao and W. I. Grosky, “Narrowing the semantic gap Improved text-based Web document retrieval using visual features,,” IEEE Transactions on Multimedia. Vol. 4-2, pp. 189-200., 2002.

[36] S. Morita, K. Yamazawa, M. Terazawa and N. Yokoya “Networked Remote Surveillance System Using Omnidirectional Image Sensors,” IEICE, J88-D-II, 5, pp. 864-875, 2005

[37] J. Sklansky, “Finding the Convex Hull of a Simple Polygon,” PRL, vol. 1, issue 2, pp 79-83, 1982.

[38] 画像処理ハンドブック編集委員会, “画像処理ハンドブック,” 昭晃堂, ISBM-4785690240, 1987.

[39] Matthew Brand “Fast Low-Rank Modifications of the Thin Singular Value Decomposition”. Linear Algebra and Its Applications 415, pp.20‐30, 2006.

[40] Thomas Landauer, P. W. Foltz, & D. Laham. “Introduction to Latent Semantic”. Discourse Processes 25, pp. 259‐284, 1998.

[41] http://julius.sourceforge.jp/

[42] Spärck Jones, K.. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation 28, pp. 11–21, (1972) .

[43] Nascimento, Mario A. and Oliveira, Arlindo L, "An Empirical Comparison of Text Categorization Methods," String Proecssing and Information Retrieval. Lecture Notes in Computer Science. pp.183-196. 2003.

[44] ムハマドヌルルフダ, 河嶋 宏明. 新田恒雄. “3 ステージ MLN と抑制／強調処理に基づく調音特徴抽出,” 情処音声言語処理研報, Vol.2008, No.123, pp.149-154,, 2008.

[45] M.N. Huda, K. Katsurada, and T. Nitta, “Phoneme recognition based on hybrid neural networks with inhibition/ enhancement of Distinctive Phonetic Feature (DPF) trajectories,” Proc. Interspeech'08, pp.1529-1532,, 2008.

[46] L.Rabiner, “An introduction to hidden Markov models,” ASSP Magazine, IEEE Vol.3 , Issue. 1, pp. 4–16, 1986

[47] M.N. Huda H. Kawashima, and T. Nitta,, “Distinctive Phonetic Feature (DPF) extraction based on MLNs and Inhibition/ Enhancement Network,” IEICE Trans. Inf. & Syst., Vol.E92-D, No. 4, pp.671-680, 2009

[48] M. N.Chomsky, “The sound pattern of English,” Harper and Row, 1968.

[49] 福田隆, 山本航, 新田恒雄, “弁別的特徴ベクトルを用いた音声認識に関する検討,” 音学講論, Vol. I, No. 1-9-1, pp. 1-2, 2002.

[50] T. Fukuda, “Orthogonalized Distinctive Phonetic Feature Extraction for Noise-Robust Automatic Speech Recognition,” IEICE TRANS. INF. & SYST., Vol.E87-D, 2004.

[51] 新田恒雄, 井上雄, 正井康之, 松浦博, “複合音響特徴平面に基づく音声認識のための局所特徴抽出法,” 信学論 D, Vol.J83-D2, No. 11, pp. 2341-2349, 2003.

[52] Kobayashi, T., Itahashi, S., Hayamizu, S. and Takezawa, T. "ASJ Continuous Speech Corpus for Research," Acoustic Society of Japan Trans. Vol.48, No.12, pp.888-893, 1992.
JNAS: Japanese Newspaper Article Sentences.

http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[52] “ATR デジタル音声データベース,” ATR-Promotion, Available: http://www.atr-p.com/sdb.html#DIGI.

[53] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals”, The Journal of the Acoustical Society of America, vol. 57, issue S1, p.S35, 1975.

[54] “Fast Artificial Neural Network Library,”. Available: http://leenissen.dk/fann/wp/.

[55] M. R. a. H. Braun.,“A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” Proc. IEEE International Conference on Neural Networks,, 1993.

[56] GalateaTalk, ” http://sourceforge.jp/projects/galateatalk/downloads/22206/get_f0s-0.1.tar.gz/.

[57] Speech Signal Processing Toolkit (SPTK),” http://sp-tk.sourceforge.net/.

[58] 小池 宗幸, 古井公司, 古井 貞熙, “HMM 音声合成における残差駆動による自然性の向上,” 日本音響学会 2003 年春期公園論文集 Vol., No., 1-6-10, pp.241-242, 2003.

[59] M. R. Schroeder, B. S. Atal, “Code-excited linear prediction (CELP): high-quality speech at very low bit rates,” ICASSP’85, vol.10, pp.937‐940, 1985.

[60] Y. Linde, A. Buzo, R. M. Gray, “An Algorithm for Vector Quantizer Design,” IEEE Transactions Communications, Volume:28, Issue:1, pp.84-95, 1980.

[61] Eric Moulines; Francis Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, Speech Communication 9, pp. 453–467, 1990.

研究業績目録

論　文（査読付き，筆頭）

1. 木村 優志，澤田 心大，入部 百合絵，桂田 浩一，新田 恒雄，“音声と画像シーンを用いた潜在意味解析に基づくタスク推定”，電気学会論文誌 C，Vo.132，No.9，pp. 1473-1480，(2012-9)（２章）
2. 木村優志，入部百合絵，桂田浩一，新田恒雄,“調音特徴―声道音響パラメータ変換を用いた調音特徴運動 HMM 音声合成”電子情報通信学会，Vol.J96-D，No.5，pp.1356-1364, May. 2013　(４章)

論　文（査読付き，共著）

3. 田口亮，木村優志，小玉智志，篠原修二，入部百合絵，桂田浩一，新田恒雄：　“幼児の学習バイアスを利用したエージェントによる語意学習の効率化,”人工知能学会論文誌，Vol.22，No.4，pp 444-453，2007 年 5 月

国際会議発表（筆頭）

4. Masashi Kimura，Takayuki Onoda，Yurie Iribe，Kouichi Katsurada and Tsuneo Nitta: ″One-Model Speech Recognition and Synthesis Based on Articulatory Movement HMMs″，Proc. of NCSP11，pp.392-395 (2011-3)　（３章）

国際会議発表（共著）

5. Ryo Taguchi，Masashi Kimura，Shuji Shinohara，Kouichi Katsurada and Tsuneo Nitta：Implementation of Bias Observed in Child Development into Concept Learning Agent ， The IASTED International Conference on Articial Intelligenceand Applications，pp.507-512 (502-123)，2006 年 2 月 15 日
6. Tsuneo Nitta，Takayuki Onoda，Masashi Kimura，Yurie Iribe，Kouichi Katsurada，“One-model speech recognition and synthesis based on articulatory movement HMMs”，INTERSPEECH'10，pp. 2970-2973，2010

口頭発表（筆頭）

7. 木村優志，作元佑輔，田口亮，桂田浩一，岩橋直人，新田恒雄：“共有信念に基づく発話場面の推定”，電子情報通信学会 2008 年総合大会講演論文集，DS-2-5 (2008-3)
8. 木村 優志，作元 佑輔，田口 亮，桂田 浩一，岩橋 直人，新田 恒雄：“人間 - ロボット間の共有信念に基づく発話場面の推定”，2008 年度人工知能学会全国大会論文集，3E3-07 (2008-6)
9. 木村 優志，作元 佑輔，田口 亮，桂田 浩一，岩橋 直人，新田 恒雄：“人間 - ロボット間の共有信念に基づく発話場面の推定”，2008 年度人工知能学会全国大会論文集，3E3-07 (2008-6)
10. 木村 優志，作元 佑輔，田口 亮，桂田 浩一，岩橋 直人，新田 恒雄：　“発話シーンの共有信念に基づく推定とその評価”，日本認知科学会第 25 回大会発表論文集，pp.318-319 (2008-9)
11. 木村 優志，澤田 心太，桂田 浩一，新田 恒雄：　“情景と音声言語の混在情報から得た部分空間に基づくタスク推定”，日本音響学会 2010 年春季研究発表会講演論文集，3-6-3 (2010-3)

12. 木村 優志，小野田 高幸，入部 百合絵，桂田 浩一，新田 恒雄： “調音運動に基づくワンモデル音声認識合成方式”， 電子情報通信学会技術研究報告，SP2011-41，pp. 1-6 (2011-7)

口頭発表（共著）

13. 溝口 勇太，田口 亮，木村 優志，篠原 修二，入部 百合絵，桂田 浩一，新田 恒雄：“相手モデルを利用した対話エージェントの教示戦略”，2007 年度人工知能学会全国大会論文集，2G5-1 (2007-6)
14. 溝口 勇太，田口 亮，木村 優志，土井岡 伴哉，桂田 浩一，新田 恒雄： “知的エージェント学習実験プラットフォームの構築”， 2008 年度人工知能学会全国大会論文集，3E3-08 (2008-6)
15. 武井 匠，木村 優志，桂田 浩一，新田 恒雄： “調音特徴に基づく 1-model 音声認識-合成”， 2009 年度人工知能学会全国大会論文集，1F2-OS7-6 (2009-6)
16. 澤田 心太，木村 優志，入部 百合絵，桂田 浩一，新田 恒雄： “情景と音声言語の混在情報から得た部分空間に基づくタスク推定”， 2009 年度人工知能学会全国大会論文集，1F2-OS7-7 (2009-6)
17. 溝口 勇太，木村 優志，桂田 浩一，新田 恒雄： “Q学習を用いた協調行動のための戦略獲得”， 2009 年度人工知能学会全国大会論文集，1F2-OS7-11 (2009-6).