ヒューマンインタフェース 97−7
音声言語情報処理 40−7
（２００２．２．１）

マルチモーダル音声認識のための顔検出

村井 和昌，中村 哲
ATR 音声言語コミュニケーション研究所

あらまし

低 SNR 環境の音声認識において、画像情報を併用することにより認識率が向上することが知られている。更に、話者の顔向きや位置など、音声情報からは得られない情報を検出することも可能である。 そのためには、実時間で画像中から顔や、調音器官の部位を同定したり、顔向きを推定する必要がある。ここでは、オンラインマルチモーダル音声認識のための顔検出アルゴリズムの１例について報告する。

Face Detection for Multimodal Speech Recognition

Kazumasa Murai, Satoshi Nakamura

Abstract

It is well known that the fusion of visual modality to audio modality improves recognition performance in low audio SNR environment speech recognition system. Visual modality also provides inaudible information like speaker's facial orientation, and location of a mouth. To acquire these information, we have to estimate the face and speech organs within face in real time. We will describe an implementation of our face detection algorithm for the multimodal speech recognition.

1. はじめに

人と人とのコミュニケーションでは、聴覚に加えて視覚を利用していることが知られている[1]。また、充分な読話の訓練を受けた人が発話者を直視した場合には発話内容を認識できることが知られている。直視に代えてビデオ録画された 2 次元動画像からも発話内容が認識できることが確認されており、ビデオ画像には発話内容を認識するために必要な情報が含まれていると考えられる。一方、このことを鑑みて、音声とビデオ画像を融合することにより音声認識の頑強さを得る研究[2][3]が進められている。

また、人と人が直接対話する際には、発話内容ばかりでなく、発話中であることや聞き手の特定のための顔向きや視線など、ノンバーバルな表現が活用されている。これらは音声だけから聞き取ることは困難であるが、視覚的には容易に観察でき、円滑なコミュニケーションに寄与している。我々は、発話に伴う口唇の動きやノンバーバルな表現を含む画像情報に基づいて音声認識に有用な情報を得ることを検討している。画像から得られる情報は、読話と同様の発話内容に加え、口位置を検出することによる音源の同定、本研究で言及する顔の向き検出などが考えられる(図１)。

筆者らは画像や、音声と画像を併用して発話を検出する方法[2][4][5][6]を提案し、音声と画像を併用することにより、騒音下では音声単独や画像単独よりも認識率が向上することを示した。これらの効用を得るためには音声と同様に実時間で頑強に

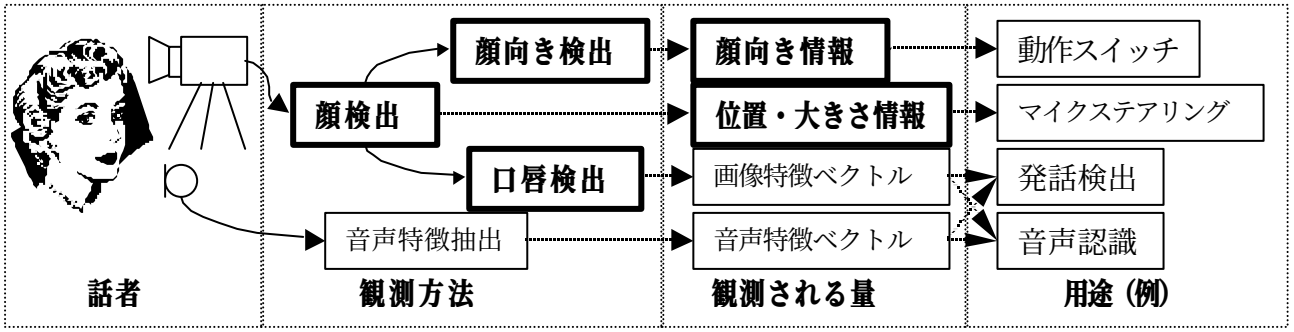

図1. マルチモーダル音声認識

顔を検出する必要がある。そこで、色情報や顔の特徴を用いた検出方法を提案した。しかし、話者に応じて肌色を設定する必要があること[2][4]や、背景や着衣の色や形に依存する[5][6]などの課題がある。 そこで、不特定話者・実時間で背景や照明に依存しにくい顔と顔向きを検出したので、手法およびその応用(図 1 中の太枠部分)を報告する。

以下、第 2 章でマルチモーダル音声認識のための顔検出の要件を挙げ、第 3 章で本研究で用いた顔検出と顔向き検出の手法について解説する。第 4 章で提案の顔検出方法の評価を行う。

2. 音声認識のための顔検出

顔の位置や向き、顔の器官の検出、顔の同定など、顔検出に関して多数の研究者が研究を進めている[2〜9]。画像単独の研究では、話者、照明や撮影条件に対する頑強さ、検出精度などが要件とされることが多い。マルチモーダル音声認識では、それらに加え、用途に応じて以下の要件を満たす必要がある：

・音声情報と画像情報との同期
・実時間での検出
・顔向きなどのノンバーバル情報の獲得

音声情報と画像情報の同期は一般的な手法により実現できるため、本研究では、実時間の検出と顔向きの検出について言及する。

オンラインの認識の場合、音声・画像とも実時間で検出し、認識しなければならない。 多くの音声認識システムは実時間処理を前提としているが、画像からの検出や認識は処理時間がかかるため、オフラインのものが多いが、オンラインの検出に向けて、高速化に関する研究[7][8]も進められている。 これらのアルゴリズムでは、主に色情報により検出を行うもの[8]、主に空間的な特徴を用いるもの[7]、色情報により枝刈を行い、探索空間を狭めてから空間的な特徴により詳細に探索を行うもの[3]などが挙げられる。

一般に、肌色と唇色の色度は、人種ごとに狭い範囲に分布していると考えられている。そこで 25 人の日本人男性の正面顔を含む RGB 静止画像をインターネットから収集・調査した。この際、明らかに画像を加工しているものと眼鏡を着用している画像は対象から除いた。各々の両眼の角膜(黒目)中心を目視により同定し、その中点の画素の RGB 値を、sRGB[10]色画像と仮定して CIE L*a*b*色座標系に変換した結果(a*b*平面)を図 2 中○に示す。比較のため、10 人分の上唇のほぼ中央の画素を同様に調査した結果を図中×で示す。調査の結果、a*b*空間における肌色同士の色差は、明度(L*値)を無視して最大 41.8 となり、唇の色領域とも重複することが判った。色領域が重複することからも類推することができるが、実際の画像中の肌色は広い範囲に分布している。

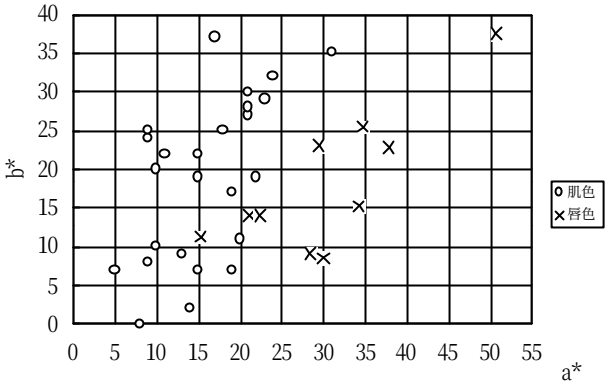

図２　顔面の肌色と唇色の分布

そこで、筆者らは話者ごとに肌色の代表色を 1〜3 色、肌色ではない代表色を 0〜1 色指定する方法により、色のみで顔領域を検出する方法を提案した[2]。この方法は、画素ごとに判定が可能であるため、検出時間が高速であり、また、肌色と肌色でない代表色を任意に設定できることから、背景色と肌色が近い場合でも、顔領域を得ることができる。しかし、色を指定するための手動操作が必要であることが課題である。

Yang[8]らは、話者や照明による肌色の影響を防ぐため、動的に肌色領域を変更する方法を提案し、実時間での顔検出を実現した。

色に基づく顔検出は、画素毎に判定できるため、実時間の検出には適するが、本質的に、背景が肌色であると検出できないという課題がある。

福井らは、空間的な特徴に基づく高速な検出方法を提案している[7]。空間的な特徴により認識する方法として一般的なテンプレートマッチングが広く用いられているが、探索空間が広いため、実時間での検出は困難である。そこで、同心円状の分散フィルター(後述)により、瞳、鼻孔、口端などの特徴点の候補を抽出し、この候補だけに対してパターンマッチングを行うことにより、0.1 秒での検出を実現した。分散フィルターは特徴点の明暗差の影響を受けず、かつ、ノイズの影響を受けにくいという特徴がある。また、計算が比較的シンプルで、ハードウェアにも適する。

3.色と空間的な特徴を併用した顔検出

本研究では、色と空間的な特徴を併用し、更に動画像の特徴を前提とした顔検出方法を提案する。

3.1 顔検出問題の設定

実際の動画像中の顔は、位置や大きさ、回転などが想定されるが、ここでは、 実時間で検出することを勘案して、画像中に正立した正面顔が高々1 つ存在する状況を想定している。画面中の位置と大きさ、垂直軸方向の回転だけを考慮し、正面顔画像を検出することを前提とした。画像中から顔を探索し、得られた探索結果に対し、正面顔と回転した顔とのマッチングを行い、顔向きを推定する。

顔の探索は、正面顔関数 *f(x, y, size)*を定義し、この関数の最大値を与える *x, y, size* を求める。ここで*(x,y)*は左右の黒目の中点の画像中の位置、size は左右の黒目の間の距離とする。

位置と大きさの張る空間から正面顔関数 *f* の最大値を与える点を探索すれば、顔を検出することができる。しかし、D1 規格(横 720 縦 480 画素)のテレビ画像を想定した場合、探索空間は $4x10^7$ を超え、実時間での評価は困難である。また、縦横とも 1/2 の画素数にダウンサンプリングした場合でも $5x10^6$ あるため探索は現実的ではない(35≦size≦70 に限定し 98.1 秒/フレーム、Athlon 1.4GHz、以下の計算時間も同環境で計測)。そこで、色と空間的な特徴により枝刈し、更に動画の特性を用いて高速化した。この計算手順を図 3 に示す。

3.2 両眼中点の色による枝刈

前述したように、肌色は広範囲に分布しているが、画像中には肌色ではない色が占める割合が多い場合もある。肌色ではないことが明らかな部分には顔があることは想定されないため、探索範囲を枝刈することができる。

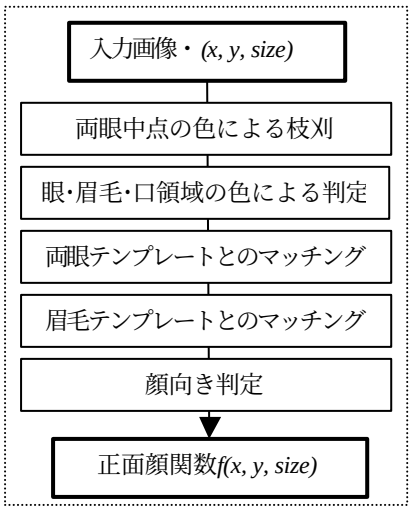

図3. 顔検出

正面顔では、両眼の中点が肌色であると想定できる。実用上は眼鏡のブリッジ部分となることもあるため、両眼の中点と、上下方向にそれぞれ両眼距離の 5%づつの画素を調べ、肌色の画素が存在しない場合には枝刈した 6 人の正面顔の動画(眼鏡なし 5 名、あり 1 名、計 58,166 フレーム)を認識したところ、背景色に依存するものの、高々11 画素を確認するだけで、探索空間の 36～96%を枝刈できることが確かめられた。また、誤って正解を枝刈りするケースは見られなかった。

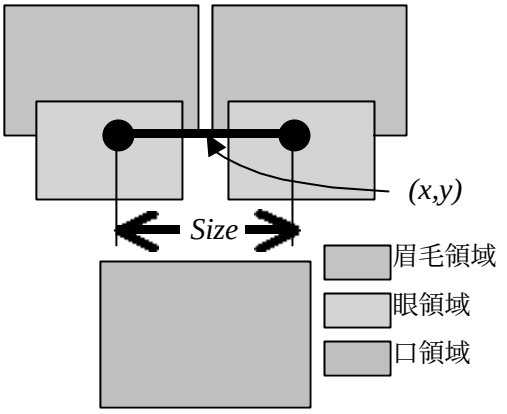

図４顔の各器官の領域

実装上の速度を向上するため、肌色の判定は、上述した CIE L*a*b*空間ではなく、実測した RGB 値(各色 0～255)に基づき、次の範囲とした:

$$\{(r,g,b) \mid 150<r \wedge 0.3r<g<0.9r \wedge 0.3r<b<0.8r \wedge b<g\}$$

以下では、「肌色」はこの範囲の色とする。枝刈の対象となった場合、正面顔関数を－∞とした。肌色の範囲を適切に変更すれば人種などの要因に対応できると考えられる。

3.3 眼･眉毛･口領域の色による判定

肌色の部位は前節で述べた両眼の中点に加えて、正面の顔面に分布する。前述の 25 名の正面顔では図 4 に示す目･眉毛･口の領域の 70%以上が肌色である。この際、前述 6 名の動画データを判定したところ、88%の画素が肌色であると判定された。図 4 の領域は文献[9]に基づいている。この領域中の肌色が少ない場合には顔ではないと判定することとととした。両目･両眉毛･口の領域中、肌色の画素の割合により判定するが、高速化と過度の枝刈を防ぐため、以下の基準としている:

- テスト画素:縦横とも 4 画素おきに１画素
- 枝刈の基準:テスト画素の 50%以下が肌色

これにより探索空間の 27%を枝刈すると同時に、誤検出を削減できることが確かめられた。正解を誤って枝刈りするケースは 0.3%であった。

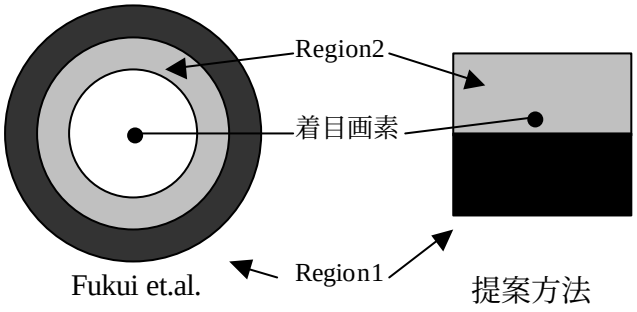

図5. 分散フィルターの形状

3.4 テンプレートマッチング

前章で述べたように、肌色の色域は広く、対応する部位も、着衣や頭髪の状況により変動する。そこで、空間的な特徴のテンプレートマッチングにより目と眉毛を同定する。テンプレートは被験者 14 名をディジタルスチルカメラで撮影した高解像度の静止画から作成した。

3.4.1 空間フィルターによる特徴抽出

入力画像は照明などにより明度が大幅に変動するため、明度の影響を受けにくいフィルターを用いて画像の空間的な特徴を得た。福井らは図 5 右に示す同心円上の 2 つの領域の分散によりフィルターを定義し、目頭・目尻・口角などの端点候補を選定している。この算出は自乗和演算に簡約することができるが、フィルターの形状が同心円であるため、着目画素のフィルター値を算出するためには、その都度、フィルター内の全画素を演算対象とする必要がある。本研究では、眉毛・目・口など水平方向の成分が多いことに着目し、福井らの分散フィルターを変形し、図 5 左に示す 2 つの矩形からなるフィルターを構成することにより水平成分を検出した。いずれのフィルターとも、対象とする領域を region1 と region2 に分け、夫々に属する画素のグレースケール値に基づき、次式により着目画素のフィルター値を定める。

$$h = \frac{sb^2}{sT^2}$$

$$sb^2 = n_1(P_1 - P_m)^2 + n_2(P_2 - P_m)^2$$

$$sT^2 = \sum_{i=1}^{n_1+n_2}(p_i - P_m)^2$$

ここで、n_1: region1 内の画素数、n_2: region2 内の画素数、P_1,P_2,P_m はそれぞれ、region1 内のグレースケール平均値、region2 内のグレースケール平均値、region1∪region2 内のグレースケール平均値、p_i は各画素のグレースケール値である。本研究の各 region は矩形であるため、各領域の和や平方和を求める際には、隣接した矩形の結果との差分のみの計算により高速に計算できる。このフィルターは分散の比により定まるため、ここでは分散フィルターと呼ぶこととする。

テンプレートは、両黒目中心の位置を、距離が 128 画素(*size*=128)となるように目視で合わせた 14 名分の静止画顔画像に基づいている。それぞれの画像に分散フィルターを適用し、眼領域、眉毛領域ごとに相関が最大となるように位置を合わせた上で、それらの画像の画素ごとの平均値を求めた。正面顔のほか、顔向きを検出する目的で水平方向 10°20°45°から撮影した静止画顔画像のテンプレートも作成した。また、正面画像(図 6)は鏡像を含め合計 28 画像の平均値としている。異なる大きさの画像を認識するテンプレートとして、15≦size≦128 となるように画像を縮小し、同様の手順で合計 1596 種類のテンプレートを作成した。

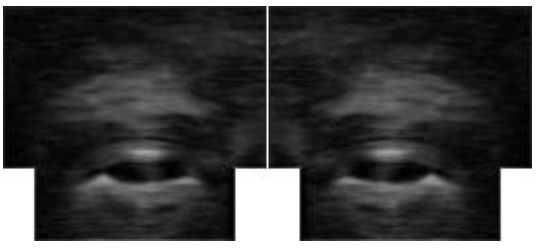

図6テンプレート画像（正面,眼・眉毛領域）

3.4.2 テンプレートの適用と顔の検出

テンプレートを適用するために、認識対象の画像に前述の分散フィルターを適用する。次に、両

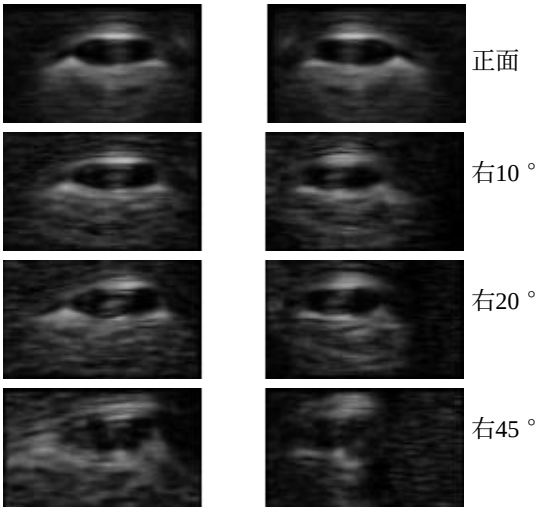

図7. 顔向き検出テンプレート

眼領域のテンプレートとの相関係数を算出する。相関係数が所定の値(実験では 0.3)以下の場合には、眼ではないと判定し枝刈する。この枝刈により、認識時間を 58%とすることができた。両眼の相関が所定の値を超える場合、眉毛領域の相関係数を求め、両眼と眉毛領域の相関係数の平均値を正面顔関数値とした。本研究では、正面顔関数値が最大値を与える*(x,y,size)*を探索することにより顔を検出しているため、画像中の顔が正面を向いていない場合でも､顔として検出する場合がある。そこで、検出された眼領域について、正面顔、左右 10°,20°,45°のテンプレートとの相関係数を算出し、相関係数が最大となる角度を顔向きと推定する。

3.5 画像の特性に基づいた高速化

前節までの枝刈とアルゴリズムの改善により、検出速度は 1.8～25s/フレームとなった。更に、探索の際の *x,y,size* のステップを、極大値近傍のたらいの形を勘案し、夫々4,1,2 とし、次いで *x,y,size* の 26 近傍(各変数とも－1,0,+1)を極大値が得られるまで順次探索することにより、0.22～2.6 秒/フレームとすることができた。実際のビデオファイル(Microsoft avi 形式)や、オンラインのディジタルビデオ信号は圧縮されており、デコードする必要がある。このデコードに要する時間は 0.006 秒/フレーム、また、分散フィルターの計算時間は 0.01 秒/フレームであった。

2.6 動画の特性に基づいた高速化

一般的に、動画像では連続するフレーム間の相関は高く、画像中での話者の顔の位置や大きさがフレーム間で大幅に動くことが想定されないため、顔の検出に成功した場合、それ以降の検出は、全フレームの位置と大きさを元に、その近傍のみを探索することにより検索時間を短縮することができる。本研究で用いた 6 人の正面顔の動画像とテンプレートを用いた実験に基づき、探索範囲を限定している：

1. 最初のフレームの場合と、直前のフレームで顔が検出できなかった場合には上記の手順で探索を行う
2. 直前のフレームで顔が検出された場合、その位置と大きさを初期値として正面顔関数の極大値が見つかるまで繰り返し 26 近傍を探索する
3. 2 の探索の結果、極大値が 0.5 以上の場合、検出結果とする
4. 極大値が 0.5 未満 0.4 以上の場合、*size* の探索範囲を－3～＋3 に限定して上述の探索を行う
5. 極大値が 0.4 未満の場合は最初のフレームと同様に上記の手順で探索を行う

以上の探索の結果、極大値(最大値)が 0.5 以上の場合、そのときの*(x,y,size)*を検出値とし、これに対して方向推定を行う。この方法により、フレーム間での探索時間はデコード、フィルターを含め 0.028 秒/フレームとなった。ここで、上記の正面顔関数の閾値(0.4, 0.5)は画像とテンプレートの構成方法により異なると予想される。

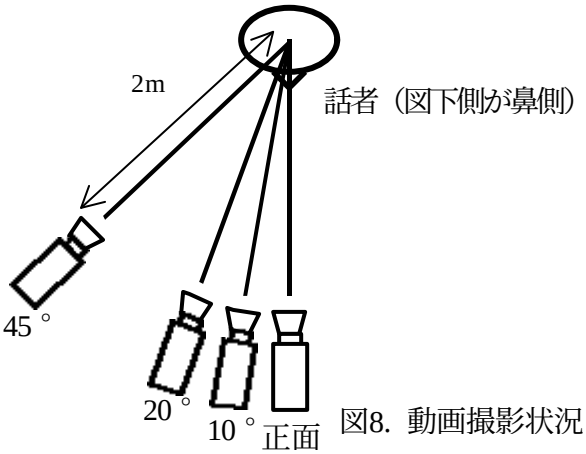

図8. 動画撮影状況

4.検出実験と結果

前述の 6 人の正面顔の動画(眼鏡なし 5 名､あり 1 名、計 58,166 フレーム)について顔画像の検出を行った(表 1 中の正面動画) 。この際、眼･眉毛･口領域の色による判定の有無についても比較を行った。また、このうち 3 名については、図 8 に示すように正面動画と同時に 10°,20°45°に設置した 4 台のビデオカメラで収録した。この結果をそれぞれ表 1 の 10°動画、20°動画、45°動画に示す。

表 1　顔検出率結果

検出条件	正面動画	10°動画	20°動画	45°動画
眼･眉・口領域判定なし	97.4%	62.2%	21.0%	0.2%
眼･眉・口領域判定あり	98.9%	27.0%	11.3%	0.2%

表 1 より、正面顔については高い精度で検出が可能であることがわかる。また、眼･眉毛･口領域の色による判定を行うことにより検出率が向上することが確認された。正面動画での誤検出は短時間の瞬きによるものが主であった。この判定を行い、顔が検出された画像について、顔の方向を推定した結果を表 2 に示す。

表 2　方向推定結果

顔向き検出結果	正面動画	10 ° 動画	20 ° 動画	45 ° 動画
正面	73.9%	34.1%	11.5%	0.0%
10 度	20.7%	46.3%	41.6%	0.0%
20 度	5.4%	19.6%	46.9%	0.0%
45 度	0.0%	0.0%	0.0%	100.0%

表２より、顔向きが正解となる率が最も高いことがわかる。また、正面～20 ° と、45 ° の 2 つのクラスとすると、クラス間の誤推定はなく、発話のスイッチとして用いることができる。

枝刈・判定による検出時間を図 9 に示す。色枝刈は両眼中点の色による枝刈を行った後の検出時間、領域色判定は眼･眉毛･口領域の色による判定を行った後の検出時間、両眼照合は両眼のテンプレートマッチングによる枝刈を行った後の検出時間、探索の間引き後の上限(◆)、下限(■)、 前フレームの値を初期値として近傍のみを探索した場合(●)の検出時間をそれぞれ示す。また、画像のデコード時間(codec)及び分散フィルターの検出時間との合計を直線で示す。図からも明らかなように、枝刈と判定、動画の特性の応用により検出時間が大幅に短縮していることが判る。

5．結論

実時間の顔検出方法を検討した結果、色による枝刈と空間的な特徴を併用することにより、実時間に適用可能な顔検出方法を得た。また、範囲を 0 ° ～20 ° に取れば顔向きによる音声認識の動作スイッチとして応用が可能である。テンプレートは少数の高解像度の静止画から作製したが、認識対象の動画像との画質のマッチング、テンプレートを作成する際のデータ量などを評価し、それに基づく性能の改善が考えられる。

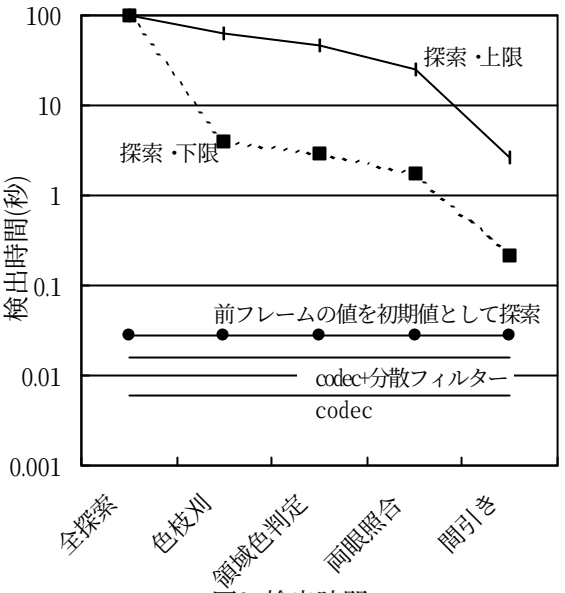

図9. 検出時間

参考文献

[1] http://www.theshop.net/campbell/mcgurk.htm

[2] 村井和昌，Reiner Gruhn，中村　哲,"口周囲画像による発話の検出",情報処理学会 2000 年秋期全国大会予稿集

[3] G.Potamianos,C.Neti,G.Iyengar,E.Helmuth,"Large-Vocabulary Audio-Visual Speech Recognition by Machines and Humans", Proc. Eurospeech 2001, pp1027-1030

[4] 村井和昌,野間啓介，熊谷建一,松井知子,中村哲,"頑強な発話検出",第 2 回音声言語シンポジウム予稿集,情報処理学会, 2000

[5] 村井和昌, 中村哲,"話者の顔画像を併用した発話検出", 日本音響学会 2001 年秋期研究発表会予稿集,pp23-24

[6] Kazumasa Murai, Kennichi Kumatani and Satoshi Nakamura, "A Robust End Point Detection by Speaker's Facial Motion", *Proc. HSC2001*, pp199-202, 2001

[7] Kazuhiro Fukui, Osamu Yamaguchi, "Facial Feature Point Extraction Method Based on Combination of Shape Extraction and Pattern Matching", D-II Vol.J80-DII No.8 pp.2170-2177, IEICE, 1997

[8] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel. "Real-time face and facial feature tracking and applications", Proc AVSP'98, pp79-84, 1998

[9] 原島ほか "感性擬人化エージェントのための顔情報処理システムの開発研究成果報告書",イメージ情報科学研究所, 1998

[10] http://www.srgb.com