音声言語情報処理 37−10
（2001．7．14）

画像と音声情報の併用による雑音に頑強な発話検出

村井 和昌，中村 哲
(株)ATR 音声言語通信研究所

あらまし

A Noise Robust Speech Detection Method by Audio Visual Information

Kazumasa Murai, Satoshi Nakamura
ATR Spoken Language Translation Research Laboratories

ABSTRACT

In this paper, we propose a method to detect speech by audio and visual modalities. It is well known that the accuracy of speech detection affects speech recognition accuracy. Because the detection by audio modality is intrinsically disturbed by audio noise, we have researched on the video modality speech detection. The method is not only robust to the audio noise, but also robust to the speaker's motion and other video modality disturbances. However, the accuracy of detection is less accurate because the duration of speech motion is intrinsically longer than the duration of speech. Thus, we propose a bimodal speech detection method. Proposed method is able to eliminate the false detection caused by audio noise. The experiment confirms that the proposed method improves the word accuracy not only in clean condition, but also in the noisy condition (SNR 10dB).

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,発話内容が認識できることが知られている.ビデオ録画された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,記録された2次元動画像にも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.

音声認識では主に音声情報に基づいて認識することが多く,頑強な発話検出についても研究が行われている[2].しかし,発話者の画像情報がある場合でも認識に活用することは希であった.

そこで,我々は画像情報に基づいて音声認識に有用な情報を得るために,発話者の画像を活用する方法を検討している.画像から得られる情報は,読話と同様の発話内容ばかりではなく,口位置を検出することによる音源の同定,視線や顔の向きを検出することによる発話意思の検出や,本研究の発話区間検出などが考えられる.

筆者らは,画像により発話を検出する方法を提案した[3][4][5].これらの方法は騒音の影響を受けないことに加えて,カメラの手ぶれや話者の動きに対しても頑強である.しかし,後述するように発話動作は実際の発話区間と本質的に異なるため,検出誤差が生じ,認識率低下の要因となる.そこで,本研究では,画像と音声を融合することにより,頑強に発話を検出することを目的としている.

以下,発話検出と認識率の関係を第1章で,発話動作と発話の対応を第2章で検討した上で,発話動作を検出する方法と音声との融合を第3章で述べる.これを検証するための実験を第4章で説明し,その実験結果を第5章で示す.

1. 発話検出誤差と認識率

音声認識において,発話検出精度は認識率に大きな影響を与えると考えられている.そこで,実際にどの程度の影響があるかを実験に基づいて検証した.発話検出は,発話開始時刻を検出する始端検出と,発話終了時刻を検出する終端検出により発話区間を検出する.本研究では,始端と終端それぞれについて,検出精度が低下したことを想定した認識実験を行った.始端検出の誤差が 2,000ms～100ms であることを想定し,ハンドラベルにより発話区間をラベル付けした旅行対話の 123 発話(男性 2 名,読み上げ,SNR46dB)の始端を,実際の発話始端の前に

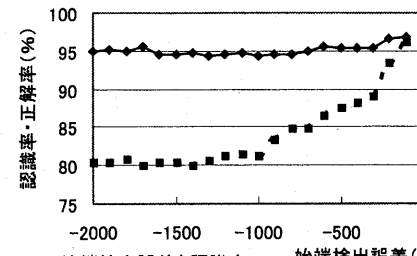

図1a.始端検出誤差と認識率

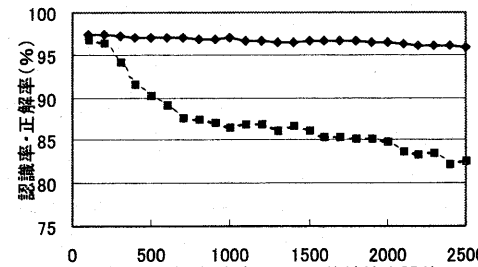

図1b.終端検出誤差と認識率

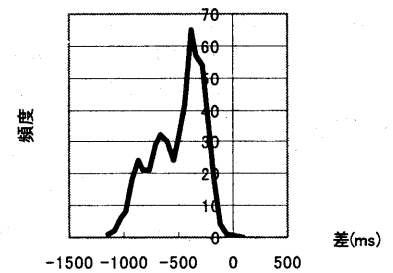

図2a.動作区間と音声区間の差(始端)

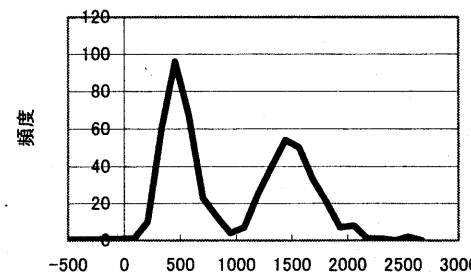

図2b. 動作区間と音声区間の差(終端)

100msおきに2,000ms～100msにわたって無音区間を延長し(発話区間に先立って,ノイズだけからなる発話の無い区間を付加)認識率を求めた.この結果を図 1a に示す.同様に,終端についても

100ms～2,500msの範囲で発話終端後に無音区間を延長したものについて同様に認識率を求めた(図 1b).図中,実線(◆)は単語正解率,鎖線(■)は単語正解精度である.始端,終端とも,誤差が大きくなるに従って単語正解精度が低下することが判る.単語正解率と単語正解精度の差は挿入によるもので,これが認識率の低下の主な要因である.ここでは,始端または終端のそれぞれについて実験したが,実際には始端・終端とも誤差の影響を受けると考えられる.

この実験の際に記録した画像から観察した発話動作は,音声の発話区間よりも,始端では最大1,200ms,終端で最大2,700ms長かった.図2aに発話始端時刻を0として、この長さのヒストグラムを示す.同様に,図 2b に終端(発話終端時刻=0)の結果を示す.以上から,画像による検出では認識率が低下すると考えられる.

2. 発話動作と発話区間

発話するためには調音し,開口する必要があり,発話に伴って口唇を含む調音器官が変形する.この変形は顔の外観の変化となるので,顔を含む動画像により観測することができる.本研究では,画像中の顔を検出し,その顔の口唇を含む調音器官の変形から発話を検出している.

通常,発話者は発話に先立って調音器官を発話の最初の音素の口形に変形(発話準備)し,発話が終了した後で口を元の形に戻す(発話始末).調音器官の変形は,発話準備の開始から発話始末の終了まで観察される.このため,本質的に調音器官の動きと発話区間とは一致しない.

口唇は,顔の動きなどにつれて,発話以外にも動く.また,呼吸などのために,発話をしない場合でも変形し,開口することがある.開口や動きを検出すると,これらの状況では誤検出となる場合がある.そこで,発話に伴う調音器官の動きを,他の動きと区別するために,実際に発話を撮影した画像により調査し,以下特徴が観察された.

- 発話準備から発話終了までの間,音素の変化とほぼ同じ速度で口唇が変形しつづける.
- 鼻音の一部と両唇音以外では,開口している.
- 発話していない場合でも開口が観察される.
- 発話の最後の鼻音を除き,閉口し続けているときには発話していない.

これらの特性に基づいて,画像による発話区間は**「速度が速い変形が続く状態」**として検出できる.以下では,画像により観察される発話準備から発話終了までの区間を「動作区間」,また,音声により観察される発話区間を「音声区間」と記す.図3に,孤立単語 520 語による音声区間と動作区間の長さの散布図を示す.この実験では,孤立単語の音声区間は345～1,248msの範囲に分布しているが,これに対応する動作区間は 1,034～3,220ms に分布している.このことから,変形が続く状態の長さは,1,000ms以上と想定できる.

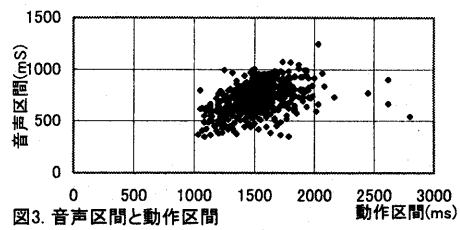

図3. 音声区間と動作区間

上述のように,動作区間と音声区間は異なるが,動作区間は音声区間に含まれる.そこで,画像により動作区間を検出し,検出した区間内で音声により発話検出を行えば,誤って無声音区間を検出する可能性を低減することができる.

3.発話検出

本研究の発話検出は以下の手順による.

1. 空間フィルターと配置による入力動画像中の顔の検出;
2. 口唇断面の検出;
3. 口唇断面の履歴による動作区間の検出;
4. 検出された動作区間内での音声による発話検出.

以下に,その詳細を述べる.

3.1 顔の検出

本研究では,Blob filter[6]と,レイアウトマスク[7]によって顔領域を検出している:

(1) 入力したRGB画像(図 4a)を,画素毎に 2R-G-Bによってグレースケール(図 4b)に変換する.この色変換は,照明条件の影響を除き,肌色を強調する方法である.

(2) Blob フィルターを適用する.顔の各器官は,画像上では水平方向の成分を多く含むため,本研究では横線を強調するように,Blob フィルターの形状を円形から楕円に変形した(図4c).

(3) フィルターを適用した画像に対し,両眼の距離(scale) と中点の座標(x, y)によって定まるレイアウトマスク中の平均値が最大となるように(scale, x, y)を探索する.レイアウト

マスクは,両眼,両眉毛,口をそれぞれ含む5つの矩形（図4d)からなる.

以上により,話者の顔を検出する.連続するフレーム間では、顔は大きく移動したり,大きさが、画像中の顔の位置と大きさはフレーム間で非常に強い相関があるので,最初のフレームでは(scale, x, y)を全探索するが,2フレーム以降では直前の値を初期値として,極大値が見つかるまで繰り返し近傍を探索している.また,色変換やBlob フィルターは比較的計算量が多いため,値が必要となった時点で計算することにより高速化した.本研究ではNTSC方式,DVフォーマットのデータ(29.97fps,幅720x縦480画素,インターレース)を基に,偶数・奇数フィールド毎に360画素,240画素59.94fpsにリサンプルした画像を用いたが,顔の全探索には13秒(PentiumIII 1GHz)を要するが,繰り返し近傍を探索する方法では1.26msで検出することができる.

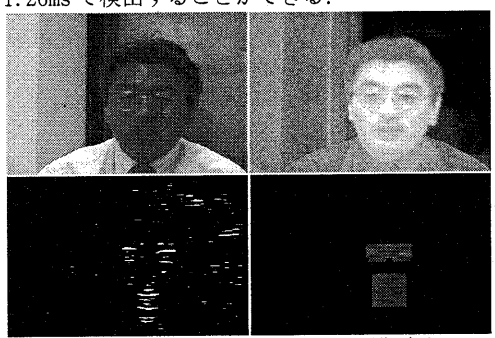

図4. a: RGB入力画像(左上), b: 2R-G-B画像(右上), c: Blob Filtered image(左下), d:検出した顔各器官(右下)

3.2 口唇断面の検出

筆者らの以前の研究[4][5]では,顔を色により検出し,重心から下方に線分を引いて口唇を跨ぐ線分を検出していた.しかし,顔が斜めに撮影された場合,下方に線分を引いても口唇を跨がないことが想定される.そこで,本研究では,両眼の角膜重心の垂直2等分線を用い,より安定して口唇断面を検出する方法を開発した.以下にその手順を示す.

(1) 両眼の矩形から色情報を用いて角膜(黒目)を検出し,その垂直2等分線を求める.眼は低彩度で,角膜は低明度のため,容易に検出できる.瞬きなどで角膜が検出できない場合には直前の情報を用いる.

(2) 求めた垂直2等分線のRGB画素の情報をフレーム毎に記録する.この際,両眼間の距離を用いて線分の長さを正規化する.

以上により検出した口唇断面を得る.顔の検出には2R-G-Bのグレースケールを用いたが,顔面内部では図4bのように低コントラストとなる.次節で述べる発話検出はCIE L*a*b*色情報を用いるため,元のRGB画像(図5)から画素毎にCIE L*a*b*色座標系に変換する.

3.3 口唇断面の履歴による動作区間の検出

前節で変換した口唇断面の履歴に基づいて,フレーム間エネルギー$E(f)$を算出する.

$$E(f) \equiv \min_{ofs} (e(f, ofs))$$

$$e(f, ofs) \equiv \sum_{j \in \text{断面}} \Delta ab(Lab(f, j) - Lab(f-2, j+ofs))$$

ここで,$Lab(f,j)$はフレームfの断面のL*a*b*色座標,ΔabはL*a*b*色空間上のEuclid距離を示す.切り出した画像は,位置の基準が両眼の角膜の重心であるため,瞬きなどにより垂直方向にシフトしやすい.そこで,$E(f)$を求める際,値が最小となるように調整する.本研究では,調整値(ofs)の範囲は,顔面上の距離でほぼ±15mm(±15画素)とた.また,入力画像はインターレース画像を基にしているため,リサンプルした画像の2フレーム前(入力画像の1フレーム前)の画像と比較することにより影響を受けないようにした.図6に,図5に対応する$E(f)$(太線)と,音声のパワー(細線)を示す.図から,動作区間とそれ以外で$E(f)$が大きく異なること,および,動作区間が音声区間よりも長いことがわかる.図6から,適宜の閾値により動作区間が検出できるか否かを判定するために,520単語の孤立発話について音声区間を含む動作区間(太

図5.口唇断面の履歴

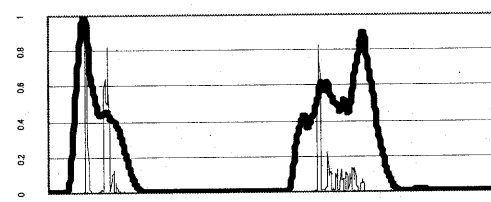

図6.フレーム間エネルギー(太線)と音声のパワー(細線)

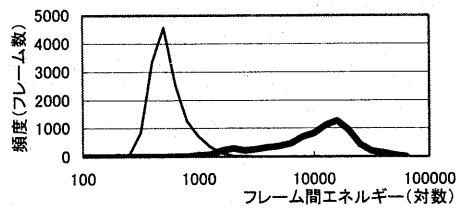

図7. 動作区間と音声区間のエネルギーと頻度

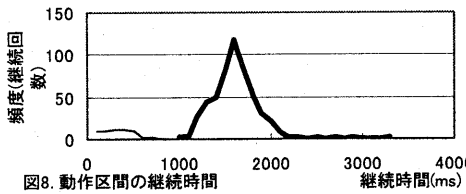

図8. 動作区間の継続時間

線)と,それ以外の区間(細線)での $E(f)$の頻度を求めた.その結果を図7に示す.図7から判るように,音声区間を含む動作区間の最小値付近(それ以外の区間の最大値付近)の領域では, $E(f)$が重複するために,エネルギーの値だけでは検出ができないことがわかる.そこで,エネルギーが大きい領域の継続時間に着目した.図8に520単語の孤立発話についての結果を示す.継続時間では重複する領域はなく,音声区間を含む動作区間を同定することがでる. このことから,以下の基準により音声区間を含む動作区間を判定する.

(1) 1秒以降の各フレームfについて,約1秒前のフレーム間エネルギーとの比較を行い,
$$E(f)>2E(f-60)$$
となるフレームfを動作始端候補とする.この不等式が成立するフレームが 60 フレーム連続した場合,動作始端とする.

(2) 発話区間中の各フレームfについて,約1秒前のフレーム間エネルギーとの比較を行い,
$$E(f)<\frac{1}{2}E(f-60)$$
となるフレームを動作終端候補とする.この不等式が成立するフレームが 30 フレーム連続した場合,この不等式が成立する最初のフレームを動作終端とする.

3.3 動作区間内の音声による発話の検出

前節までの方法により検出した動作区間に対し,パワーに基づいた発話区間検出を適用する.本研究では, ATR-Sprec[8]の発話検出部 ATR-EPD(ATR End Point Detection)を採用した. ATR-EPD は音声のパワーとその変化率により発話検出を行う.

4. 検出実験

上述の手法により, 提案法と ATR-EPD による音声のみの発話検出を行い,その結果に基づいて不音声認識実験を行った.以下に実験条件を述べる.

4.1 評価データ

日本語を母国語とする男性2名に日本語の旅行対話 56 対話を読み上げる内容のタスクを収録した。読み誤りを含め、合計 123 対話を収録した.

画像は、NTSC 規格のディジタルビデオカメラ(720x480 画素,29.97fps)により 2m 離れた正面から,顔全体が映るように撮影した.照明には商用電源(60Hz)で点灯した白熱灯を用い,シャッタースピードを 1/60 秒に設定した.

音声は接話型マイクロホンと DAT(48k サンプリング/S,16bit,SNR46dB)により収録し、ビデオ音声と同期して用いた.雑音に対する頑強性を確認するため,SNR が 10dB となるように展示会場のノイズ[9]を付加した音声も作成した.

4.2 認識システム

本研究では HTK3.0 を用いた。音響特徴パラメータは、サンプリングレート 16kHz,窓長 20ms,フレームシフト 10ms で抽出した 25 次元の特徴ベクトル(12 次元メルケプストラム,12 次元Δメルケプストラムと Δlog power)を用いた. 音響モデル,テストセットとも CMN を適用している.

4.3 音響モデルと言語モデル

本研究では,167 名の男性話者の旅行対話データを学習した性別依存の音響モデルを用いた.このモデルは,26音素,各音素モデル3状態,5混合ガウス分布,総状態数 1,400 の状態共有化 HMM(Hmnet)で表現されている.言語モデルは,旅行対話から作成したバイグラムモデルで,辞書は単語数 32,304 語である.

5. 実験結果

4 章の実験結果を表 1 に示す.提案法は,clean の単語正解率を除き,従来法よりも良い結果が得られている.従来法の認識結果を調査したところ,無音や雑音のみの区間を誤って発話区間として検出することがあり,誤検出した雑音のみの区間を「え…」「あの…」などのフィラー語などとして認識し,単語正解精度が低下していることが判った.提案法では,動作区間ではない区間を検出しないため,単語正解精度が向上した.

表1. 実験結果

	提案法	従来法 (ATR-EPD)
clean (単語正解率 %)	93.88	94.90
SNR10dB (単語正解率 %)	75.79	71.73
clean (単語正解精度 %)	87.55	78.28
SNR10dB (単語正解精度 %)	61.36	55.01

6. 結論

本研究は,雑音の影響を受けにくい発話検出により音声認識率を向上した.発話検出精度が認識率に与える影響を調べた結果,始点・終点とも検出精度が低下すると認識率が低下することが確かめられた.画像による発話検出は雑音の影響を受けないという利点はあるものの,発話に伴う動作は音声区間に比べて本質的に長いため,動作区間をそのまま発話検出としても,誤認識の要因となる.そこで,動作区間に対して音声モダリティーの発話検出を併用する方法を提案した.提案法は,初期値を設定することなしに画像中から顔と顔中の各器官を認識し,口唇断面を検出する.口唇断面の履歴に基づき,発話を含む動作区間を検出することができる.

画像と音声の2つのモダリティーを用いることにより,音声単独に比較して騒音下の音声認識率を改善することができた.

参考文献

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

[2] Jean-Claude Junqua, Ben Reaves and Brian Mak "A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM Recognizer", *Eurospeech* 1991 pp. 1371-1374, 1991

[3] 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

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

[6] 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

[7] Harashima et.al "Facial Image Processing System for Human-like "Kansei" Agent", *IPA*, 1998

[8] M. Naito, H. Singer, H. Yamamoto, H. Nakajima, T. Matsui, H. Tsukada, A. Nakamura, Y. Sagisaka, "Evaluation of ATRSPREC for travel arrangement task", proc. *ASJ Fall Meeting*, pp. 113-114, 1999

[9] Noise Database published by Japan Electronics and Information Technology Institutes Association, 1990