音声言語情報処理 34－13
(2000. 12. 21)

口周囲画像による頑強な発話検出

村井和昌 † 野間 啓介 ‡ 熊谷 建一 †‡ 松井 知子 † 中村 哲 †

† ATR 音声言語通信研究所

〒 619-0288 京都府相楽郡精華町 2-2-2

‡ 奈良先端科学技術大学院大学 情報科学研究科

〒 630-0101 奈良県生駒市高山町 8916-5

E-Mail:{kmurai, xkumata, tmatsui, nakamura}@slt.atr.co.jp,
keisu-no@is.aist-nara.ac.jp

あらまし 音信号や画像信号による発話認識において, 発話区間検出精度は認識率に大きな影響を与える. 特に, 騒音下では音声が騒音に埋もれてしまい, 音信号のみを用いて発話区間を正確に検出することは難しい.
本稿では, 発話検出の一手法として, 騒音に影響されることなく発話区間を検出することが可能な発話者の顔画像から発話区間を検出する方法を提案する. 本方法では、まず、色情報から肌色領域を検出し、発話器官を含む領域を推定する. 次に、この領域の画像の変形測度から発話を検出している. 評価実験により、本方法は画像ノイズに対しても頑強であることが確認された. 音信号 (SNR25dB) のみによる発話区間検出では 97.5% の検出率であったのに対し、本方式では画像ノイズの有無に関わらず 99.8% であった.

キーワード 音声認識, 発話区間, 顔画像, 肌色, 発話検出

A Robust End Point Detection by Speaker's Facial Image

Kazumasa MURAI †‡ Keisuke NOMA ‡ Kenichi KUMATANI †‡ Tomoko MATSUI † Satoshi NAKAMURA†

† ATR Spoken Language Translation Research Laboratories

2-2-2 Seika, Soraku, Kyoto, 619-0288 JAPAN

‡ Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara, 630-0101 JAPAN

E-Mailkmurai, xkumata, tmatsui, nakamura}@slt.atr.co.jp,
keisu-no@is.aist-nara.ac.jp

Abstract In this paper, we propose a method to detect the end points of speaking sections (EPD: End Point Detection) by visual information. It is well known that the accuracy of EPD affects speech recognition accuracy. Detecting the speech end points from a noisy audio signal is difficult because the speech is masked by the audio noise. We propose a method for EPD that uses image of the speaker's facial motion that are not affected by audio noise. Our method locates the skin area by color information and estimates the area that includes the speech organs. Then the end points are detected by the speed at which the image alternates. An evaluation experiment also confirms that the proposed method is robust with respect to visual noise. Its accuracy with/without visual noise is 99.8% while audio(SNR 25dB) EPD is 97.5%.

Key words Speech Recognition, Speaking Section, Facial Image, Skin Color, End Point Detection.

1. はじめに

一般に,音声認識においては発話区間を検出してから認識を行う.騒音の少ない環境においては,音声信号のパワーによって比較的容易に発話区間を検出することができるが,騒音下では問題がある.そこで,簡易な音声認識処理を行いながら,その結果に基づいて発話区間を検出するなどの方法が提案され,検討されている[1][2].

一般に,発話検出には以下の課題が考えられる.

(1) 発話区間を含む最小区間の検出
　検出した開始端点が発話開始よりも遅れている場合や,終了端点が発話終了よりも進んでいる場合には、発話の一部を欠くこととなる。これは、音声認識を行う上で誤認識の主な原因の1つである。逆に発話区間以外を含む場合には,発話区間以外の騒音により誤認識することがある.

(2) 騒音・外乱への適応
　音信号による発話検出では,騒音により精度が低下する.

(3) 不特定話者への対応
　話者により発話検出精度がばらつく.

(4) 判別用閾値の設定
　発話検出は一般に、予め設定した閾値と,入力信号から導出される値との比較によって発話を検出する.そのため,検出結果はそれらの定数に依存する.入力信号の特性に頑強な発話区間検出を行うためには,閾値が入力信号に依存しないような検出方法や,入力信号から適応的に導出できる方法が望ましい.

(5) 実時間での検出
　実時間で音声認識を行うためには,発話区間も実時間で検出する必要がある.

本研究では,発話者の顔画像を用いることにより,(2)騒音・外乱と、(4)判別用閾値の設定に関して頑強な発話検出方法を提案する。以下では、検出対象となる発話と動画像の特徴について2章で検討し、この特徴に基づいた発話検出方法を3章で提案する。提案の方法により実験を行った結果を4章で示す。

2. 発話動作の特徴と動画像の特徴

動画像による発話検出は、発話に伴う動作を撮影した画像に基づいて発話を検出する。そこで、発話と画像の関係について調査した。また、外乱に頑強な検出を行うため、想定される動画像における外乱について検討した。

2.1 発話と画像の関係

話者が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,この動画像を認識することにより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する口形に変形(発話準備)し,発話が終了した後で口を閉じる(発話始末).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.しかし,通常の発話では調音器官の動きは発話区間を含む.

また,口唇は,顔の動きなどにつれて,発話以外にも動くこともあるし,呼吸などのために,発話をしない場合でも開口することがある.単純に開口や動きを検出すると,これらの状況では誤検出となってしまう.

そこで,発話に伴う調音器官の動きの特徴を,実際に発話を撮影した画像により調査した.その結果,以下特徴が見られた.

発話をしていない状態

閉口している場合には発話は観察されなかったが,発話していない状態でも開口が観察された.開口している場合の動きは,数秒単位の比較的ゆっくりした動きや,停滞が見られた.また,発話していない状態でも,顔全体の動きが見られた.

発話準備から発話始末中の状態

鼻音(/n/,/ng/)の一部や,両唇音(/p/,/b/, /m/)の場合の短時間の閉口を除き,開口していることが観測された.発話中は,比較的高速な口唇の変形が観測された.また,発話開始が両唇音であっても,発話準備のための変形が観測された.

以上から,比較的高速な口唇の動きを観測することにより,発話(発話準備～発話始末)区間を検出できることがわかった.

2.2 動画像における外乱(ノイズ画像)

画像は音の影響を受けないため,画像が入力できれば著しい騒音下でも発話を検出することができるが,画像にも騒音と同様に外乱の影響を受ける.一例として,照明条件の変動,カメラのぶれ(手ぶれ),被験者の動き,フォーカスのずれ,画像信号自体のノイズや歪み,照明とシャッタースピードの干渉によるちらつき(フリッカー)が挙げられる.これらのノイズは,発話検出にも影響するために,適切な対策が必要である.これらのうち,本研究では,フリッカーや,画像信号事体のノイズや歪みに関しては既存の画像処理技術等により対処可能であると考えられるので,照明条件の変動,カメラの手ブレ,被験者の動

きについて検討を行った.

(1) 照明条件の変動

照明,被験者の皮膚の向き,撮影位置により,カメラに入射する光量が著しく変化する.これに伴い,口唇や皮膚の同じ部位であっても画像上の明度が変動する.一般に,照明の色度(色味)や,画像上の色度は殆ど影響を受けない.

(2) 手ブレ

カメラを手に持ち撮影した場合,手が動くことによりカメラの位置や向きが変動する.この変動に伴い,画像も変動する.スチル(静止画)の写真でもブレが生じるように,動画の各々の単一フレームでもブレが生じる事があるほか,動画ではフレーム間の位置のずれが生じる.単一フレームのブレはシャッタースピードを高速にすれば解決できる.一方,フレーム間のブレによる画像は全体的に平行移動した画像となる.画像上は,カメラの向きの変動による影響が大きく,ブレの量は撮影対象とカメラの間の距離にほぼ比例する.画像上のブレは撮影距離に依存するため,被験者3名にビデオカメラを手で持ち,撮影対象1人の顔画像を撮影するようにインストラクションを与えたところ,何れも 1m～2m 離れた位置で撮影したことから手ブレのノイズ量は,距離2mの手ブレが想定される.本研究では、手ブレをシミュレーションするために、距離 7m にある目標を手で持ったカメラから撮影した手ブレのデータベースを作成した。このデータベースに基づいて、距離2mの手ブレを算出し、被験者の顔面上の変位(単位 mm)としたものを図1に示す.

(3) 被験者の動き

通常,発話の有無に関わらず呼吸や所作などに伴って体は動き,動きの量は状況(立位,座位,歩行中など)によって非常に大きく変化する.発話を検出する際にはこれらの動きを含んだ発話者の画像中から調音器官の変形を検出する必要がある.画像上では,変位については手ブレと同様の平行移動となるが,顔の回転(うなずき等)は画像の動きの他,顔の向きに応じて上述した照明条件の変動も見られる.

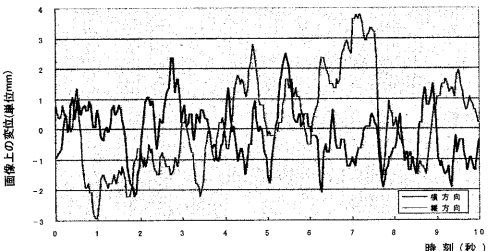

図1 手ブレの実測値

上記のノイズのほか,現実的にはカメラと被写体の間を手や書類等の障害物が通過する場合や,被験者が振り向くなどにより,画像中に被験者の調音器官が映らない場合も想定できるが,本研究ではこれらのノイズは考慮していない。

本研究では、座位の被験者が概ね正面を向いた顔の全体を含む画像を距離 2m から固定したカメラまたは手で持ったカメラで撮影した顔画像を対象としている.

3. 発話検出方法

ここでは、前章の検討に基づいた発話検出方法を提案する。調音器官の変形を検出するためには,顔や唇,目などの部位をパターンマッチングなどにより検出し,認識した唇の形状の変化を検出する方法と,唇などの位置を特定せず,少なくとも唇を含む領域の変形を検出する方法が考えられる.前者では,部位の検出ができれば,顔の向きや色,照明などに依存せずに頑強な発話検出が可能であると考えられるが,認識時間が長いという課題がある.後者では,瞼の瞬きなど,発話と近い動きをする器官の動きによって誤認識することが想定されるが,前者に比較すると短時間で算出が可能である.

本研究では,特に口唇を認識することなしに口唇を含む領域を検出し,その変形に基づいて発話を検出することとした.口唇の検出は,動画の各フレームから,色情報によって顔領域(肌色)を検出し,その重心位置から口唇を跨ぐ線分を検出している.その断面の変化から発話を検出している(図2).

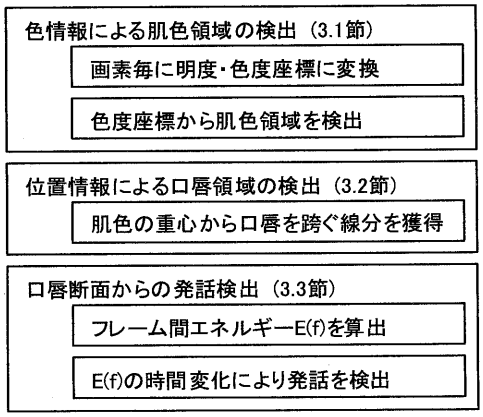

図2. . 発話検出の流れ

3.1 色情報による肌色領域の検出

一般に,肌色と唇色の色度は,人種ごとに非常に狭い範囲に分布していると考えられている.これが正しいかを検証するため,公開されている日本人の顔を含む静止画 RGB 画像 10 件を,sRGB(A Standard Default Color Space for the Internet： standard RGB)[3]色画像と仮定して調査した.その結果を,代表的な明度・色度分離の色空間である CIE L*a*b*(D50) 色空間上で評価した.ここで,L*値は明度を,a*と b*は色度（色味）を D50 光源において示す直交色座標系である.sRGB と,CIE L*a*b*(D50)色空間の色座標は定義式に基づく一対一対応である.

皮膚および唇の CIE L*a*b* (D50) 色空間における a*b*平面上の分布を図 3 に示す.図中では,1 画像ごとに肌色の平均値を○で,唇色の平均値を×で示している.図 3 より明らかなように,肌色・唇色ともに比較的広範囲に分布し,かつ,分布が重複していることが明らかになった.この要因としては, 光源色などの照明条件や撮影条件が一定しないこともあるが,本質的に日本人の肌色と唇色の分布が広いことが考えられる.

それを検証する目的で, 光源や撮影条件が等しい同一画像中の単独被験者の肌色及び口唇の色分布を画素毎に a*b*平面で調査した(図 4).この分布は被験者毎に異なるが,調査した範囲では,同一顔面中でも,a*b*色差で 15 程度の範囲に分布し,かつ,唇色と肌色が重複している.また,実測値の明度は照明条件等の影響により,非常に広範囲に分布している.ただ,同じ部位については,顔の向きにより明度は影響を受けるものの,色度はほぼ一定である.

肌色と唇色の分布は重なるため、色情報だけではそれらを区別することは困難である。そこで,本研究では以下の手順により肌色と唇色の両方の領域を検出し、唇は位置情報により検出することとした.

(1) 撮影した動画像から,フレームごとの RGB 信号の入力
(2) 入力した RGB 画像を,sRGB を仮定して CIE L*a*b* (D50) へ変換
(3) 色度 (a*, b*) により肌色領域を検出

この際,前述したように肌色の色域は話者に依存するため,話者ごとに a*b*平面上での肌色の基準点を 1～3 点指定することにより色域を定め,入力画像中の各画素の a*b*値が,何れかの基準点からの色差が 5 以下であれば肌色領域であると認識している.図 4 のように比較的広範囲に分布している肌色と唇色の領域でも、1～3 個の半径 5 の円で殆どを被覆することができる。このように、肌色を指定することと,明度情報を無視することにより,照明条件や,その変動による影響を防ぐことができる.また,フレームごとに肌色領域を検出しているため、検出される領域の形状はフレーム間でのブレや、被験者の動きには影響を受けない。

3.2 位置情報による口唇領域の検出

前述したように,色情報だけにより顔画像中の唇領域を検出することは困難である.そこで,本研究では,位置情報により口唇領域を検出している.画像認識により口唇を認識するためには煩雑な計算が必要であるが,発話検出では,画像領域から唇が変形しているか否かだけが認識できれば十分であるので, 以下の簡単な方法を考案した.

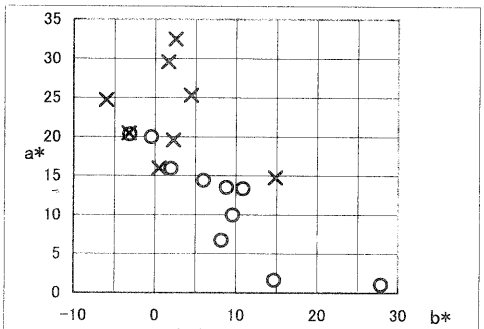

図3. 画像ごとの肌色、唇色の分布

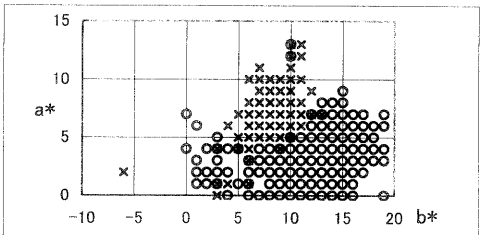

図4. 同一顔面中の肌唇色の分布

本研究では,話者の顔を概ね正面から,正立して撮影していると仮定した.この状態では,前節で説明した肌色領域(図 5)の重心が,概ね口の上部となる.話者が正立している場合,この重心から,画像の下方に向かって直線を引くと,ほぼ口唇の中央部を跨ぐ線となる.基準となる点が肌色領域の重心であるため,顔の変位や,撮影時の手ブレ等の影響を受けず,顔の中心に対する口唇の相対的な断面を得ることができる.この画素列は,特に唇等の位置を特定しなくても,画素値の変化として動きを検出することができる.以下では,この一列の画素を口唇断面と記す.尚,インターレース画像を用いる場合には,フレーム間で

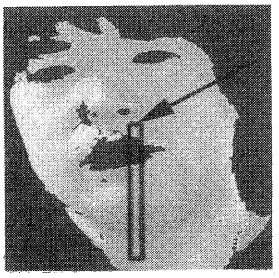

図5. 肌色領域と口唇を跨ぐ線分の検出

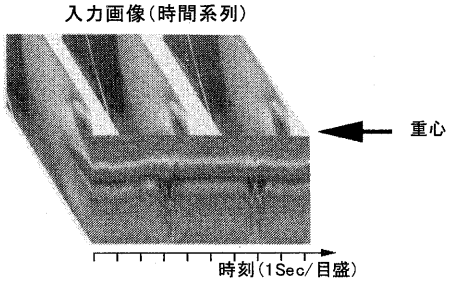

図6. 画像の時系列から検出した口唇断面

のブレが生じることがある.本研究の実験では,このブレによる影響を抑制するために,各フレームで検出した 1 列の画素を,他のフレームとの誤差が最小となるように顔面上の距離で 2.5mm 以内の上下方向の平行移動を行っている.画像の時系列から検出した口唇断面を図 6 に示す.

3.3 口唇断面からの発話検出

前節で示した検出方法により得られた画像を図 7 に示す.この図は,各フレームごとに検出した肌色領域の重心 (上端) から下方向に 150 画素(縦軸)の口唇断面を,時間軸 (横軸) 上に約 9 秒間分並べたものである. この図では,「あい」,「あいだ」の 2 つの発話を含んでおり,目視では容易に発話のための開口を観察できる.

本研究では,肌色の検出については基準色を初期設定しているが,発話区間については,画像の最初の1秒間を発話していない区間であると想定し,この画像から閾値を算出することにより明示的な指示を不要とした.また,口唇の変形は発話区間よりも長いため,実際の発話が極めて短時間であったとしても,動きのある時間はある程度の区間を想定する事ができるので,ここでは発話検出区間の最短の長さを 1 秒 (30 フレーム分) ,発話間の最小のポーズを 0.5 秒 (15フレーム)と仮定した.後述する実験において、

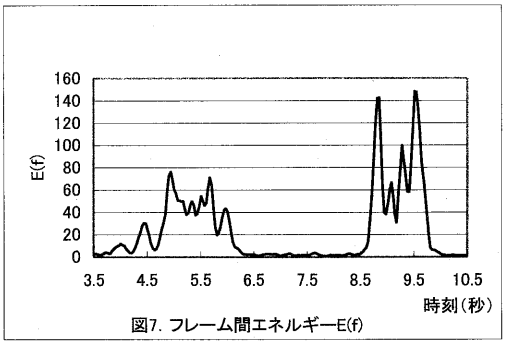

図7. フレーム間エネルギーE(f)

最短の発話は 0.349 秒であるが、口唇の動きは 1 秒を超えているために、この発話も検出することができる。

(1) 各フレーム f と,その直前のフレーム $f-1$ の口唇断面との差を算出する.

$$E(f) \xleftarrow{def} \sum \{L(f,j) - L(f-1,j)\}^2$$

ここで, $L(f,j)$は,フレーム f 中の口唇断面の上から,j 番目の画素の明度である.これをフレーム間エネルギー$E(f)$とする(図 7).入力した画像は,毎秒 29.97 フレームのディジタルビデオ (DV) 規格の画像である.

(2) 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,

$$E(f) > 2E(f-30)$$

となるフレーム f を発話区間の開始候補とする.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,発話区間であると認識する.

(3) 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,

$$E(f) < \frac{1}{2}E(f-30)$$

となるフレームを発話区間の週両候補とする.この不等式が成立するフレームが15 フレーム連続した場合,発話区間の終端であると認識する.

図 7 から明らかなように,発話区間は大きく変形するため, $E(f)$は発話していいない区間に比較して 1 桁以上の大きな区間が継続する.発話していない区間でも 1 桁程度の変動があるが,継続時間が短いため,上述の方法によって容易に発話を検出できる.

4. 検出実験

本研究の方法により発話検出実験を行った。固定したカメラで撮影した発話者の画像のほか、外乱に対する頑強さを確認する目的で、この画像から手ブレを想定した画像を合成し、検出実験を行った。ま

た、音信号による発話検出実験を行い、結果を比較した。

4.1 実験条件

日本語を母国語とする成人男性 1 名について,日本語の 520 単語を発話する内容のタスクを,NTSC 規格のディジタルビデオカメラ(720x480 画素,29.97 フレーム/Sec)により正面から,顔全体が映るように撮影した.画像上の 1 画素は,発話者の顔の上ではほぼ 0.5mmであった.撮影時間は,発話間の無音区間を含めて 35 分となった.

発話者は着席しているが,特に拘束せず,自然な状況で発話を行った.また,顔画像の検出精度を評価する目的で,顔上部にマーカーを取り付けたが,本研究の実験では,このマーカーを用いる事なく十分な精度が得られたため,位置検出の精度に付いては検証を行っていない.また,撮影した色の精度を確認する目的で,画像中に Macbeth チャートを撮影し,目視で色の変動がないことを確認した.照明は交流電源 60Hz の蛍光燈を用い,ビデオカメラのシャッタースピードを 1/60Sec とした.また,手ブレの影響を検討するために,ビデオカメラを手持ちにして手ブレのデータ(図1)を作成し,そのデータに基づいて各フレームの画像を平行移動することによって手ブレした動画像データを合成した.この際,被験者との距離は 2 mを想定している.音声は接話マイクで画像と同時に収録した。この音声の SNR は 25 dB であった。

4.2 実験結果

本研究の方法により,発話区間検出を行い,音声による発話区間を音声発話区間のハンドラベリングの結果と比較した.前述したように,画像による検出は音声の発話区間よりも長くなるため,検出結果がハンドラベリング区間を含む場合には検出が成功したものとし,無音区間を発話区間と検出した場合と、発話区間を欠いた場合を誤検出とした.その結果,本研究の方法では、520 単語の発話中,1 回の誤検出が観測され,検出率は99.8%となった.従来の音信号によるEPDでは,520発話中,13回の誤りがあり,検出率は97.5%であった.検出時間は,Pentium3, 866MHz において画像の再生時間の 4.2 倍である.

手ブレを付加した動画像の発話検出を行ったが,口唇断面が重心位置からの相対位置により検出されるため,検出結果に変化は見られなかった.更に,撮影距離が 7 mを想定した場合の検出実験も行ったが,顔画像の一部がフレーム外にはみ出し,重心の位置がずれる場合もあり,発話区間が最大 2 フレーム変動する場合も見られたが,検出率は同一であった.

表 1. 各 EPD の検出率

	音声 EPD	画像 EPD	
	SNR25dB	clean	手ブレ有
検出率	97.5%	99.8%	99.8%

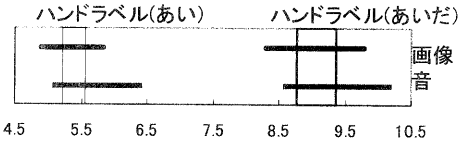

図8. 発話区間検出結果

検出区間の一例として、図 8 に発話「あい」,「あいだ」の発話検出結果を示す.図中,上から順に、本研究の方法、音信号による発話検出(ATR-EPD)、ハンドラベルによる発話区間検出を示す.

5. まとめ

音声による発話区間検出は,騒音が比較的小さい環境であっても,2.5%の検出誤りがあったが,画像による発話検出は画像にノイズを加えても 0.2%の検出誤りであり,音声よりも精度よく発話区間を検出できることがわかった.

現状では顔領域の検出のため,カラー画像を必要とし,肌色の基準点を手作業で与える必要があること,正面画像でなければ発話が検出できないことなどの課題がある.今後,音情報との融合や、高度な顔検出アルゴリズムと組み合わせることにより,設定を不要とし,正面以外の画像や,白黒画像でも検出ができるものと期待される.また,現状では画像の入力と顔の検出が処理時間の大半を占めているが,画像の解像度などを最適化することにより改善できるものと考えられる.

参考文献

[1]Jean-claude JUNQUA, Ben REAVES, Brian MAK, "A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM Recognizer", Eurospeach 1991 pp1371-1374

[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] http://www.w3.org/Graphics/Color/sRGB