

TR-S-0012

口周囲画像による頑強な発話検出
A Robust Speech Detection by Mouth Image

野間啓介
村井和昌

2000.9.29

現在使用されている発話区間検出手法として音声のパワーに対して発話区間を検出するものがあるが、騒音環境に対しては、音声のパワーが正確にとれないことにより正確に発話区間が検出できない。そこで、話者が発話していない際には口を閉じるという仮定のもとで、話者の唇の輪郭情報を active contour model SNAKES を用いて検出し、発話区間検出において騒音環境下の頑強さを高めることを目的とする。

1 実習の目的

現在使用されている発話区間検出手法として音声のパワーに対して発話区間を検出するものがあるが、騒音環境に対しては、音声のパワーが正確に取れないことにより正確に発話区間が検出できない。そこで話者が発話していない際には口を閉じるという仮定のもとで、話者の唇の輪郭情報を active contour model, SNAKES を用いて検出し、発話区間検出において騒音環境下の頑健さを高めることを目的とする。

2 唇情報による発話区間検出

2.1 唇の曲率情報による発話区間検出

発話区間検出に際して、発話している場合は、口が開いた状態になっており、発話していない場合には口は閉じられているものと仮定した。この仮定に基づき、SNAKES により二次曲線に近似した唇の二次曲線の二次の係数（以後曲率と呼ぶ）は、発話時に上昇が見られると仮定し検出を行った。

2.2 データの作成

発話検出の実験を行う為に ATR216 単語を用いてデータを収録した。その際に顔の検出や、唇など

の顔の部位の検出は多方面で研究が進められており現実的にはこれらの研究の成果を活用することができる。本研究では検出された唇の情報から発話区間を検出することが主眼であるから唇が容易に検出できるように、上唇を赤、下唇を青に着色し、発話したものをデータとして収録した。収録したデータの 1 フレーム目を図 1 に示す。今回作成したデータの話者数は 2 人である。

2.3 口からの発話区間パラメータの抽出

収録したデータの各フレーム毎に着色した色からの式差を算出し、それぞれ上唇、下唇を抽出した唇領域だけを抽出したデータの 1 フレーム目を図 2 に示す。唇領域のみを抽出したデータに対して、動的輪郭モデルの一種である SNAKES[1]-[3] を用いて唇の形を近似した。

2.3.1 動的輪郭モデル (SNAKES)

動的輪郭モデルとは、エネルギーが最小になる場所を探索しながら、曲線の形や位置を変化させていくものである。2次元の動的輪郭モデルに用いられるスプライン曲線 v については、式 (1) の様に表さ



図 1: 収録したデータの様子



図 2: 着色した唇の抽出

れる。

$$v(s) = (x(s), y(s)) \quad (1)$$

x, y は二次元空間を表し, s は画像の開空間 $[0, 1]$ を表す。この係数が線分のエネルギーを表す。

SNAKES のエネルギー計算式 (2) は 3 つの項からなる。

$$E_{total} = \frac{1}{2} \int_{\omega} E_{internal} + E_{image} + E_{constraint} \quad (2)$$

$E_{internal}$ は曲線の内部的なエネルギーを表し, その曲線の形に依存する。 $E_{internal}$ の式を式 (3) に表す。

$$E_{internal} = \alpha \left| \frac{\partial v}{\partial s} \right|^2 + \beta \left| \frac{\partial^2 v}{\partial s^2} \right| \quad (3)$$

α は曲線の張力に対する重みであり, これが大きくなるに従い, スプライン曲線の任意の 2 つの制御点が離れた場合にその曲線に対するエネルギーが大きくなる。 β は, 曲線の屈曲に対する重みであり, β を大きくするに従い, また曲線の屈曲が大きくなるに従い, その曲線に対するエネルギーが大きくなる。

E_{image} は, 画像からのエネルギーを表し, 画像の画素の値とエッジのエネルギーを与え, 式 (4) の様に表される。

$$E_{image} = \omega_1 I(x, y) + \omega_2 \left| \nabla I(x, y) \right|^2 \quad (4)$$

$E_{external}$ は, SNAKES の線分に対して, 外部から意図的に与えるエネルギーであり, スプライン曲線の場合は隣接する制御点の距離から算出するエネルギーと, 曲線上のの対象となる点と点によるエネルギーを用いたりなどして, 用途に応じて作成する。

2.3.2 唇を近似するために作成した SNAKE

今回は, 発話していない場合には, 口は閉じられていると考える。その場合には上唇と下唇は離れていないと考えた。通常 SNAKES で用いられているスプライン曲線では唇の全体の開きが表されていない, 2 次を超える高次の曲線は口の開きに対しての変

化を表現するのが困難である。これより SNAKES により作成する曲線はスプライン曲線を用いず, 2 次曲線で近似した。この場合のエネルギー関数は, E_{image} は式 (4) で表されるものを用い, $\nabla I(x, y)$ はソーベルフィルタを用いた。外部からのエネルギーについては, 上唇, 下唇の上部を近似する場合, 上方から, 下方へ向かうほうへエネルギーを置き, 下部を近似する場合には, 下方から上方へ向かうエネルギーを与える。SNAKES の初期点配置の公式は, 近似した画像の縦軸, 横軸でそれぞれ画素の値の分布が閾値を超えた場所に配置した。SNAKES の各制御点の動きは,

- 制御点の位置から 8 近傍に対し現在位置とのエネルギーを比較して, 最小のエネルギーを持つ位置へ移動する
- 8 近傍に対して比較した際に現在位置が最小値であれば, 移動は行わない
- 全ての点が移動を終了した時点で終了

作成した SNAKES により, 唇の形状を表す情報にパラメータ化する。SNAKES による唇の曲線近似の様子を図 3 に示す。

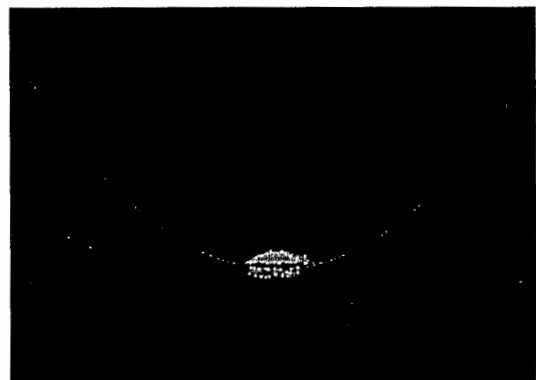


図 3: SNAKES による二次曲線近似の様子

3 発話と曲率の関係

3.1 発話と唇の曲率との同期

発話区間の発話時のパワーの上昇に対する発話のパワーの上昇のグラフを図4に示す。このグラフより、発話の際には、口の曲率が上昇し、同期しているのがわかる。これを用い、発話区間検出を行う。

3.2 発話区間検出のアルゴリズム

3.2.1 音声による発話区間検出のアルゴリズム

今回用いた音声側の発話区間検出のアルゴリズムは、ATRepdを用いた。ATRepdのアルゴリズムの概要は、

- 音声のフレーム毎にパワーの平均値 P を計算する
- 平均値 P が閾値 $P_{Threshold}$ を超えた場合、発話の開始ポイントとみなす
- あるフレーム数置いて、フレーム内のパワーの分散 d を計算し、分散値が閾値 $d_{highThreshold}$ を超えた場合、発話の開始ポイントから、発話とする。
- その後分散値 d が、発話終了の閾値 $d_{lowThreshold}$ を下回ると、発話終了とする。

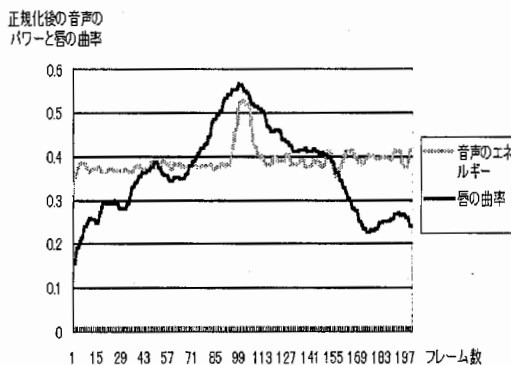


図4: 口の曲率と音声のパワーの同期の様子

3.2.2 音声と画像の統合による発話検出

今回表1により確認された発話と、開口の同期を用いて発話区間検出を行う。それぞれのフレームに対し、

$$\mu = k_a P + k_v co \quad (k_a + k_v = 1) \quad (5)$$

また発話開始ポイントとみなした後に、

$$\mu = k_a d + k_v co \quad (k_a + k_v = 1) \quad (6)$$

を計算する。ただし開口を表す曲率 co 、音声の重み k_a 、画像の重み k_v である。ただし co, P, d は正規化している。この場合の発話区間検出のアルゴリズムは、

- 式 (5) の値が閾値 $k_a P_{highThreshold} + k_v co_{lowThreshold}$ を超えた場合発話の開始ポイントとみなす
- あるフレーム数置いて、式 (6) が閾値 $k_a d_{highThreshold} + k_v co_{highThreshold}$ を超えた場合、発話の開始ポイントから、発話とする。
- その後式 (6) が、発話終了の閾値 $k_a d_{highThreshold} + k_v co_{highThreshold}$ を下回ると、発話終了とする。

音声を用いた発話区間検出が困難な場合であっても画像との混合により発話区間検出が可能であり、混合することにより音声のみもしくは、画像のみの情報により発話区間検出を行う場合と比較し、より正確に発話区間検出が行える。

4 実験評価

4.1 発話区間の含有率と区間検出誤りの評価

実験評価として、発話の発生に対し、人手により正確に発話を切り出した時間情報を正解として

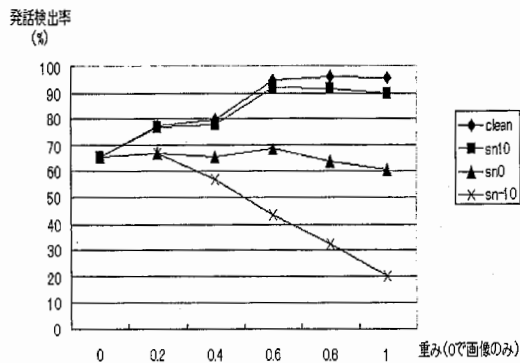


図 5: 話者 1 の発話区間検出率

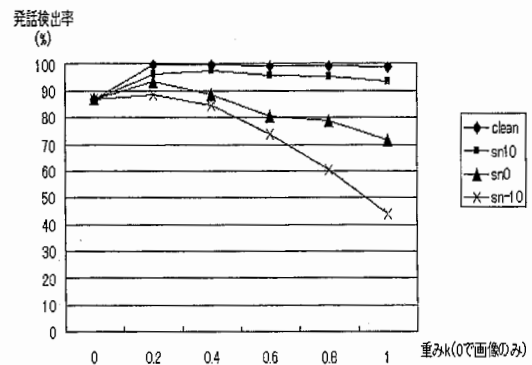


図 7: 話者 2 の発話区間検出率

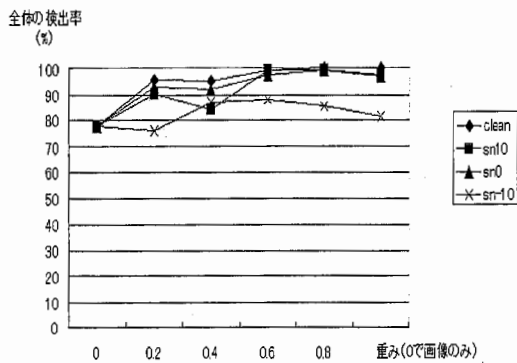


図 6: 話者 1 の全体の正解率

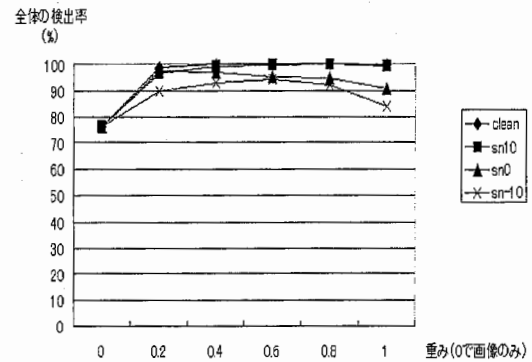


図 8: 話者 2 の発話外誤り率

考える。その正解に対し、発話開始、発話終了が含まれていたならば発話区間検出の正解、含まれていない場合は誤りとして正解率を $\text{発話検出率} = (\text{正解数}) / (\text{発話した単語数})$ で表した。また発話区間検出で正解を含まず、発話としてみなした場合を発話外誤りとして全体の検出率 $= (\text{正解数}) / (\text{発話区間としてみなした数})$ として評価した。ただし、作成した音声データの閾値は、話者 1、話者 2 それぞれに音声側は、ATRepd の閾値設定モジュールを用いて最適と思われる閾値を設定した。画像側の閾値は 88 発話の発話検出を数パターンの閾値を変化させて行い最も検出率が高かったものに閾値を設定した。また音声での発話区間検出が困難な状態を想定し、音声側に白色ノイズを付加し、同様の実験を行った。音声、画像の混合比 k を変化させた際

の話者 1、話者 2 それぞれの発話区間検出率と、発話外誤り率を図 5, 6, 7, 8 に示す。ただし、グラフ中では収録したデータを clean として表し、ノイズ付加データはその音声データへのノイズを付加した SN 比で表している。

4.2 考察

それぞれの話者に対しての傾向として、音声のみの情報もしくは画像のみの情報を用いて発話区間検出を行うより検出率が向上しているのが、確認できる。話者 1 の画像による発話区間検出率が低い要因として、216 単語を 4 つに分けて撮影したが、そのうちの一つに対しては、閾値が変化してしまっ

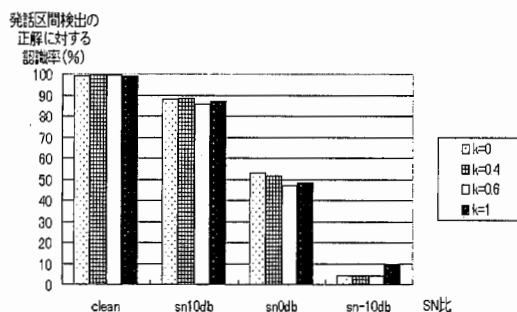


図 9: 話者 1 の検出正解数に対する認識正解数

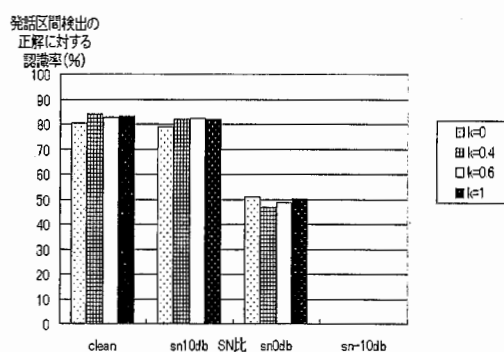


図 10: 話者 2 の検出正解数に対する認識正解数

たので発話区間検出率が、低下してしまった。ここで4分割されたうちの残り3つのみを用いた場合の発話検出率を付録に付加しておく。今回の実験では閾値を話者ごとに設定したが、撮影環境の変化についても画像の閾値を変化させる必要があることがわかった。

また図5, 図7より、音声と画像を混合することで、ノイズを付加しない場合においても発話検出率が上昇している。またノイズを付加した場合には、発話検出率が最も高くなっている重みが画像側に変化しているが、画像のみを用いる場合と比較して、音声と画像を混合した場合のほうが検出率が高くなっている。

4.3 本手法による発話区間検出の妥当性の評価

画像による発話区間検出が、従来の音声での発話区間検出に対し、現在の音声のみで発話区間検出を行った場合と認識に影響を及ぼさないことを評価する。そのため、検出に成功した場合に限り認識を行い、(認識正解数) / (検出成功数) で評価を行った。実験に用いた音響モデルはIPAの男性モデルのモノフォンを用い216単語認識を行った。

4.4 考察

話者1, 話者2の(認識正解数) / (検出成功数)のグラフを9, 10に表す。これより、画像からの発話区間検出が音声による検出と同様に認識に耐えるものと考えられる。また話者1, 話者2の人手により正確に発話を切り出した音声による認識率はそれぞれ93.517(%), 77.209(%)である。話者2の認識率が低い理由としては口を動かすことを意識しすぎ、モデルとのミスマッチを起こしたと考えられる。

5 まとめと今後の課題

現在の発話区間検出のアルゴリズムに対して画像を唇情報を用いた発話区間検出を行った。これにより、音声による発話区間検出が困難な場合においても、唇情報を用いることにより発話区間検出が可能であると考えられる。

今後の課題として、今回は閾値を人手により設定したが、顔画像の撮影条件が同一でない場合や、不特定話者に対しては、対応できていない。これより撮影条件や不特定話者に対しての閾値の動的設定な設定が必要である。