対話型ロボットのための 口領域動画像と音情報に基づく発話推定

元吉 大介 †, 嶋田 和孝 ‡, 榎田 修一 ‡, 江島 俊朗 ‡, 遠藤 勉 ‡

九州工業大学大学院 情報工学研究科 †
九州工業大学 大学院情報工学研究院 知能情報工学研究系 ‡

1 はじめに

近年，人間と音声発話でコミュニケーションをとる対話型ロボットに関する研究が盛んに行われている．これらの対話型ロボットは，複数の人間に囲まれている場合などに，人間同士の会話に誤って応答してしまう可能性がある．そこで，対話型ロボットが要求者の発話のみに応じるため，発話推定技術の利用が考えられている．

先行研究 [1] では，口領域動画像からの特徴量として，オプティカルフローと絶対値差分和に基づく特徴量を利用し，判別対象フレームについて 10 フレーム前までの情報を素性として，分類器により発話か非発話かの判別を行う手法を提案した．この手法を考察した結果，(1) 発話ではない口の動きも発話と誤判別する，(2) 前フレームの情報のみを利用しているため，発話開始・終了時付近に多く誤判別する傾向があるという 2 つの問題点が挙げられた．(1) については，音情報を利用することで解決できると考えられる．また，(2) については，前フレームの情報だけでなく後のフレームの情報も用いることで解決できると考えられる．

そこで，本研究ではこれらの問題点を改善するため，口領域動画像からの特徴量と音情報からの特徴量を用いた発話推定手法を提案する．音情報からの特徴量として，人間の声の周波数範囲のパワーを利用し，その有効性を検証する．また，判別対象フレームの後のフレームの情報も素性として利用し，その有効性も検証する．

2 特徴量

本節では，先行研究 [1] で利用した口領域動画像からの特徴量と新たに追加する音情報からの特徴量について説明する．

2.1 口領域動画像からの特徴量

口領域動画像からの特徴量として，オプティカルフローと絶対値差分和に基づく特徴量を利用する．これらの特徴量を求める際，動画像中から口領域を検出する必要がある．口領域は，Viola ら [2] が提案し，Rainer ら [3] によって改良された物体検出器を用いた手法に基づいて検出する．オプティカルフローとは，画像中の物体の動きをベクトルで表現したものである．オプティカルフローに基づく特徴量として，ブロックマッチング法により求めたフローの大きさの総和をフローの個数で正規化した値を用いる．絶対値差分和とは，式 (1) で表される，1 つ前のフレームと現フレームの対応する全画素値の差の絶対値和である．

$$SAD = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} |I_t(i,j) - I_{t-1}(i,j)| \qquad (1)$$

ここで，w と h は画像の幅と高さ，$I_t(i,j)$ と $I_{t-1}(i,j)$ は現フレームと 1 つ前のフレームの画素値を表す．絶対値差分和に基づく特徴量として，絶対値差分和を口領域のサイズで正規化した値を用いる．

2.2 音情報からの特徴量

音情報からの特徴量として，マイク入力の情報を FFT によりスペクトル解析して求めた，人間の声の周波数範囲のパワーを利用する．本手法は，動画像のキャプチャ画像ごとに発話か非発話かの判別を行うため，動画像のサンプリング間隔に合わせて音情報からの特徴量を取得する必要がある．音情報として利用する WAVE データのサンプリング周波数は 44.1kHz であるのに対し，動画像のサンプリング間隔は約 15fps であるため，動画像のキャプチャ間に 2940 個 (44100/15) の音データが存在する．一般的に，FFT するデータ数は 2 のべき乗が適切である．画像キャプチャ間

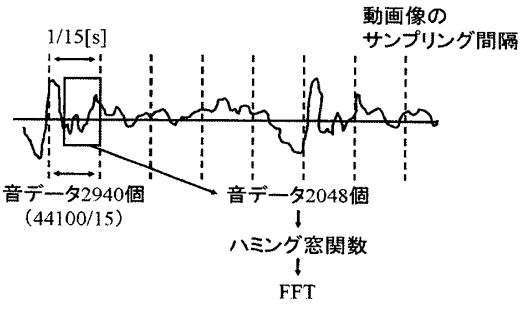

図 1 FFT までの流れ

Speech Activity Detection Using Mouth Image Sequences and Audio Information for an Interactive Robot
† Daisuke Motoyoshi, Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology
‡ Kazutaka Shimada, Shuichi Enokida, Toshiaki Ejima, and Tsutomu Endo, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology

の音データ数 2940 は 2 のべき乗ではないので，本手法では，判別対象フレームについて，キャプチャした時点から遡って 2048 個 (2^{11}) の音データを対象に FFT を行う．また，FFT を行う前には，ハミング窓関数を掛ける．この方法により，サンプリング定理よりサンプリング周波数の半分の 22.05kHz までのパワーを約 21.53Hz(44100/2048) 間隔で求めることができる．一般的に，人間の声の周波数は約 80Hz から 1.1kHz といわれている．本手法では，約 80Hz から 2kHz の範囲のパワーの総和を正面顔のサイズで正規化した値を音情報からの特徴量として用いる．

3 素性範囲

先行研究 [1] において，判別対象フレームについて，10 フレーム前までの情報を素性とした場合，発話開始時付近では，前フレームの情報の多くが非発話中の情報であるため，非発話中の情報を多く使って判別することになり，その結果，非発話と誤判別することが多かった．発話終了時付近についても，同様の理由で誤判別が多かった．この問題点について，判別対象フレームの前フレームの情報だけでなく，後のフレームの情報も素性として用いることで解決できると考えられる．そこで，本手法では判別対象フレームについて，前後 5 フレーム間の情報を素性として利用する．

4 実験

4.1 実験データ

実験データを得るため，USB カメラと指向性マイクを使って，被験者 3 名から発話区間 2322 フレーム，非発話区間 5451 フレームの計 7773 フレームを撮影した．発話内容は，施設内生活支援ロボットを想定したものとした．また，非発話区間は，口だけが動いている区間と周囲の人間の発話区間も含むものとした．これは，提案手法がカメラに写っている人間が口を動かして発話しているフレームだけを発話として判別できているかを確認するためである．

4.2 実験方法

判別対象フレームから前 10 フレームと前後 5 フレームの情報を素性とした場合の 2 パターンで実験を行った．また，音情報からの特徴量の利用の有効性を検証するため，前後 5 フレームの情報を素性とした場合において，口領域動画像からの特徴量のみと音情報からの特徴量のみ，そして全ての特徴量を用いた場合での実験も行った．分類器として AdaBoost(弱分類器は Random Forest) を利用して実験を行った．評価基準には，10 分割交差検定により求めた再現率，適合率，F 値を採用した．

4.3 実験結果

実験結果を表 1 と表 2 に示す．表 1 より，判別対象フレームから前後 5 フレームの情報を素性として用いることは有効であるといえる．また，表 2 より，音情報からの特徴量を用いることも有効であるといえる．

表 1 素性パターン別の実験結果

	再現率	適合率	F 値
前 10 フレーム	0.741	0.855	0.794
前後 5 フレーム	0.947	0.964	0.955

表 2 特徴量別の実験結果 (前後 5 フレーム)

	再現率	適合率	F 値
口領域のみ	0.322	0.572	0.412
音情報のみ	0.925	0.934	0.930
全て	0.947	0.964	0.955

5 考察

実験結果を調査したところ，前後 5 フレームの情報を用いることで，発話開始時と終了時付近の誤判別が大きく減少していることが分かった．また，周囲の発話の音量が大きい場合に発話と誤判別しやすい傾向にあることも分かった．これは，表 2 より，口領域動画像からの特徴量に比べて音情報からの特徴量の方が高精度であることから，音情報からの特徴量が大きな影響力を持っており，それが原因だと考えられる．

6 おわりに

本研究では，口領域動画像からの特徴量としてオプティカルフローと絶対値差分和に基づく特徴量を，音情報からの特徴量として人間の声の周波数範囲のパワーを利用し，判別対象フレームについて前後 5 フレームの情報を素性として，発話か非発話かに判別する発話推定手法を提案した．実験の結果，提案手法の有効性を確認することができた．

今後は，有効な特徴量の追加や判別手法の検討による精度向上を目指す．

謝辞

本研究は，次世代ロボット知能化技術開発プロジェクト（独立行政法人新エネルギー・産業技術総合開発機構）における「施設内生活支援ロボット知能の研究開発」の成果の一部である．

参考文献

[1] 元吉 大介, 嶋田 和孝, 榎田 修一, 江島 俊朗, 遠藤 勉, “ロボットとの対話のための発話推定に関する事例研究” , 画像の認識・理解シンポジウム (MIRU2008), (2008).

[2] P. Viola, M. Jones, “Robust Real-time Object Detection” , Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, pp.1–25, (2001).

[3] Rainer. L, Alexander. K, Vadim. P, “Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection” , MRL Technical Report, (2002).