ロボット聴覚のための2階層視聴覚統合を用いた音声認識システムの検討

○吉田 尚水[1], 中臺 一博[1 2], 奥乃 博[3]
[1] 東京工業大学大学院 情報理工学研究科,
[2]（株）ホンダリサーチ・インスティチュート・ジャパン, [3] 京都大学大学院 情報学研究科

Automatic Speech Recognition Improved by Two-Layered Audio-Visual Integration For Robot Audition

Takami YOSHIDA[1], Kazuhiro NAKADAI[1 2], Hiroshi G. OKUNO[3]
[1] Tokyo Institute of Technology,
[2] Honda Research Institute Japan Co., Ltd, [3] Kyoto University.

Abstract— The automatic speech recognition (ASR) for robot audition should be robust, because people usually communicate with each other using their voices. This paper presents a two-layered audio-visual integration to make ASR more robust against speaker's distance and environmental noises. The first layer is Audio-Visual Voice Activity Detection (AV-VAD) that integrates several AV features based on a Bayesian network. The second layer is Audio-Visual Speech Recognition (AVSR) that integrates the reliability estimation of acoustic and visual features by using a missing-feature theory method. Empirical results show that our system improves 9.9 and 16.7 points of ASR results and robustness against several auditory/visual noise conditions, respectively.

Key Words: audio-visual speech recognition, audio-visual voice activity detection, robot audition

1. はじめに

サービスロボットやホームロボットが人とコミュニケーションを行う日常環境では，環境ノイズやロボット自身が発するノイズなどの影響により，音声認識の性能が劣化する．加えて，それらのノイズの性質は事前に分かるとは限らない．そのため，ロボットは極力事前知識を持たずに，Signal-to-Noise Ratio(SNR) が低い信号を扱う必要がある．これを実現するため，2つのアプローチがある．一方は音源分離によりSNRの向上を図るアプローチであり，もう一方は視覚情報を利用するアプローチである．

音源分離の従来研究は，特に [1] で提唱されたロボット聴覚の分野で報告されており，ノイズにロバストであることが報告されている [2]．しかし，日常環境では雑音の種類や性質，音源などの音響条件が大きく変化し，報告されたようなロバストな音声認識が常に可能であるとは限らない．視聴覚情報を利用する従来研究は，視聴覚音声認識 (Audio-Visual Speech Recognition: AVSR) として数多く報告されている [3, 4, 5]．しかし，これらの研究では，唇の画像が常に高解像度で取得可能であると仮定している．ロボットへ適用する際必ずしもこの仮定は満たされない．

これらの問題に対処するため，我々は，これまでに人の認知機構を参考に，リップリーディングを用いた視聴覚統合，画像情報もしくは音声情報の信頼度が低い場合や一方が利用不可能な場合でも同一の枠組みで統合可能なミッシングフィーチャー理論，音声認識の認識単位の粒度を動的に変更する”Course-to-Fine”認識の3つのアプローチにより，音声認識のロバスト性が向上することを報告した [6]．この視聴覚音声認識システムは音声，画像の片方，もしくはその両方に雑音が混入した場合でも高いノイズロバスト性を示した．しかし，このシステムには，以下の問題点がある．

1. 発話区間が所与であると仮定している，
2. シングルチャンネル（マイク数 1）の入力を前提としている，
3. 評価実験において，クローズドテスト（評価用データと学習用データが同じ）しか行っていない．

実環境での音声認識では，発話区間検出 (Voice Activity Detection: VAD) の性能が音声認識の性能に大きな影響を与える．我々は視聴覚統合音声発話区間検出 (Audio-Visual Voice Activity Detection: AV-VAD) を用いることで，1. の問題解決を図る．2. を解決するため，我々はHARK[1]を導入する．HARK はロボット聴覚のためのオープンソースソフトであり，多チャンネル音響信号入力をサポートしており，音源定位，音源追跡，音源分離，音声認識といったモジュールが利用可能である．我々は視聴覚音声認識システムにこの HARK によるマイクアレイに基づいた音源分離システムを統合した．3. に対しては，システムの性能評価を単語オープンの評価実験により行った．

2. 発話区間検出における課題とアプローチ

発話区間検出の従来研究は，視覚情報のみを用いたVAD (Audio VAD: A-VAD)，聴覚情報のみを用いた

[1] http://winnie.kuis.kyoto-u.ac.jp/HARK/.

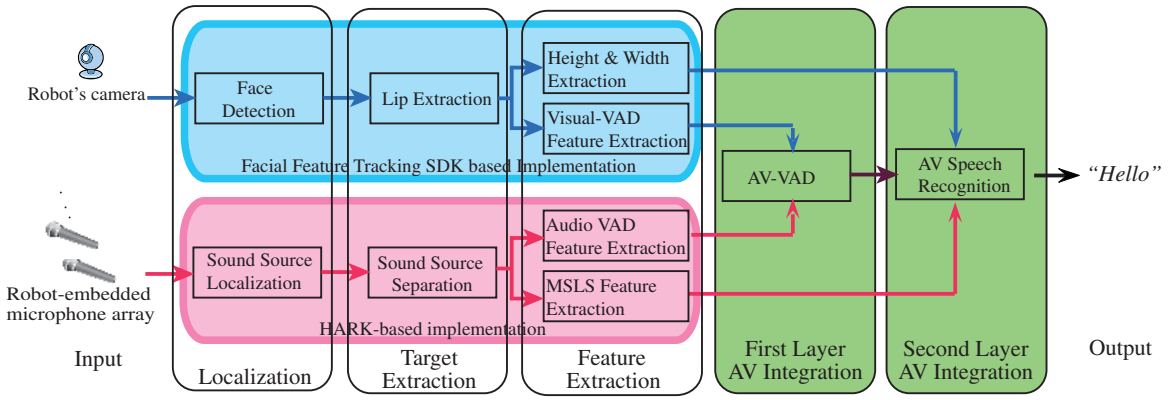

Fig.1 An Automatic Speech Recognition System with Two-Layered AV Integration for Robots

VAD (Visual VAD: V-VAD)，に大別できる．A-VADでは大きく分けて，A-1:音響信号の特徴を利用した手法，A-2: 音声特徴量を利用した手法，A-3:音声認識エンジンを利用する手法が報告されている．A-1 はパワー値や零交差数を用いた手法であり，クリーンな環境では高い検出性能を示すが，声の大きさの変化や雑音によって性能が大きく劣化する．A-2 は混合正規分布モデル（Gaussian Mixture Model: GMM）やカトーシスを用いた手法であり，想定した環境では高い検出性能を示すが，環境が大きく変化した場合には性能が劣化する．ロバスト性を向上させるには多くの学習データが必要となる．A-3 は音声認識の途中結果を用いた手法であり，高い検出性能とロバスト性を持つ．

V-VAD では，主成分分析を利用した手法 [7] や唇の縦横長を利用した手法 [8] が報告されている．これらの手法では，SNR が低い場合や画像から顔が検出できない場合では性能が劣化してしまう問題がある．

AV-VAD の研究では視覚情報を用いた V-VAD と聴覚情報を用いた A-VAD を 2 段階で行う手法が提案されている [9]．この手法も，視覚または聴覚情報の片方がノイズの影響を受けると検出の性能が劣化してしまう．

そこで本研究では，視覚/聴覚発話区間検出それぞれに用いられる特徴量をベイジアンネットワークによって統合して発話区間検出を行う．使用する特徴量は 1) Julius により計算された非発話である対数尤度 (x_{dvad})，2) 唇の縦横長から求める特徴量 (x_{lip})，3) Facial Feature Tracking SDK により求められた顔検出の信頼度 (x_{face}) の 3 種類である．1) は A-3 の手法である．2) は視聴覚音声認識での視覚特徴量としても応用が可能であるように唇の縦横長を元にした特徴量を使用する．まず，唇の縦長 h と横長 w を求め，近傍 5 点を用いて h, w のそれぞれを 3 次関数にフィッティングし，そのそれぞれ 3 次関数の 8 個の係数を特徴量として使用する (Fig. 2)．

これらの特徴量は誤差を含んでいるため，ベイジアンネットワークはこれらの特徴量の統合手法として適している．ベイジアンネットワークは，次のベイズの公式に基づいている．

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},\ j = 0, 1 \qquad (1)$$

ここで，x は $x_{dvad}, x_{lip}, x_{face}$ それぞれを表し，ω_j は ω_0, ω_1 がそれぞれ非発話，発話に対応する仮説を表す．

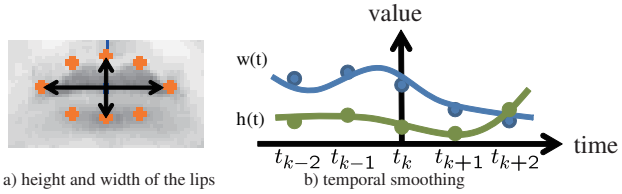

Fig.2 Visual feature extraction

各特徴量に対応する条件付き確率分布 $p(x|\omega_j)$ は 4 混合 GMM で近似し，事前に学習により求める．同時確率 $P(\omega_j|x_{dvad}, x_{lip}, x_{face})$ は，それぞれの特徴量が独立であると仮定し，

$$P(\omega_j|x_{dvad}, x_{lip}, x_{face}) = P(\omega_j|x_{dvad})P(\omega_j|x_{lip})P(\omega_j|x_{face}) \qquad (2)$$

により求めた．この同時確率を閾値処理し，発話，非発話の判別を行った．

3. システム構成

提案する AVSR システムの構成を Fig. 1 に示す．システムは 4 つのブロックから構成される．

- 画像特徴量抽出ブロック,
- 音声特徴量抽出ブロック,
- 視聴覚統合発話区間検出ブロック,
- 視聴覚統合音声認識ブロック.

以下で各ブロックについて説明する．

3·1 画像特徴量抽出ブロック

画像特徴量抽出ブロックは MindReader[2] に含まれる Facial Feature Tracking SDK を使用して実装されており，顔検出モジュール，唇検出モジュール，画像特徴量抽出モジュールの 3 つから構成される．顔検出と唇検出を Facial Feature Tracking SDK により行い，唇の周囲 8 点の座標を求める (Fig. 2)．画像特徴量抽出モジュールでは，もとめた唇の周囲 8 点のうち上下左右の 4 点を用いて上述の 8 個の係数を求める．

3·2 音声特徴量抽出ブロック

音声特徴量抽出ブロックは HARK を使用して実装されており，音源定位モジュール，音源分離モジュール，MSLS 抽出モジュールの 3 つのモジュールから構成される．音源定位モジュールでは，MUltiple SIgnal

[2]http://mindreader.devjavu.com/wiki.

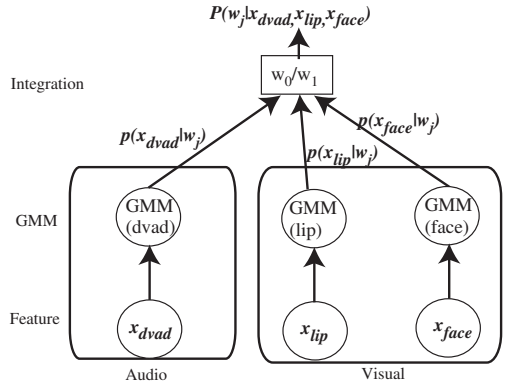

Fig.3 AV-VAD based on a Bayesian network

Classification (MUSIC) [10] を使用し，マイクアレイによって収録された多チャンネル音響信号から音源の方向を推定する．音源分離モジュールでは，Geometric Sound Separation(GSS)[11] を使用した．GSS は Blind Source Separation (BSS) とビームフォーミングを組み合わせた手法であり，BSS と BF の両方の特徴を受けている．MSLS 抽出モジュールでは，Mel Scale Logarithmic Spectrum (MSLS) [12] を抽出する．音声認識システムでは，一般に Mel Frequency Cepstrum Coefficient(MFCC) が用いられる．しかし，音源分離では分離音に分離歪が生じ，MFCC の場合にはこの分離歪が全ての特徴量に影響を与える．一方，MSLS は周波数領域の特徴量なので，分離歪は特定の周波数バンドにしか影響を与えないという利点がある．13 次元 MSLS と 13 次元 ΔMSLS，1 次元 Δlog power の 27 次元特徴量を使用する．

3·2.1　視聴覚統合発話区間検出ブロック

視聴覚発話区間検出ブロックは，Fig. 3 に示されるベイジアンネットワークを使用して画像特徴量と音声特徴量を統合し発話区間検出を行う．

3·2.2　視聴覚統合音声認識ブロック

視聴覚音声認識ブロックは，ストリーム重みを指定可能なマルチバンド Julian[13] を用いる．

4.　評価実験

構築したシステムを評価するため，以下の 2 種類の実験を行った．

Ex.1: 視聴覚発話区間検出実験
Ex.2: 音声認識実験

各実験では，男性 10 人，1 人当たり 266 単語（ATR 音素バランス単語 216 単語と ATR 重要単語 50 単語）の発話を収録した視聴覚データセットを使用した．音声データはクリーンな環境で 16bit，16kHz サンプリングで収録し，画像データはクリーンな環境で 8bit モノクロ, 640x480 ピクセル，100Hz で収録した．

AV-VAD モデルは，視聴覚データセットのうち ATR 音素バランス単語 216 単語，話者 5 人分のクリーンデータを使用し学習を行った．AVSR の音響モデルは，視聴覚データセットのうち ATR 音素バランス単語 216 単語，話者 10 人分のクリーンデータを使用し学習を行った．

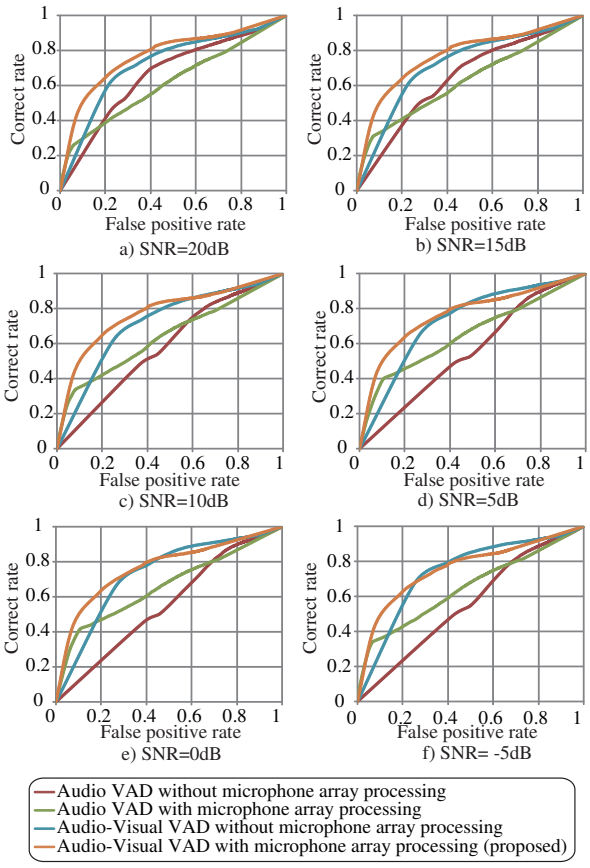

Fig.4 Results of Voice Activity Detection

音声データは，ロボットに搭載されたマイクアレイで測定した伝達関数を音声データに畳みこみ，正面（0 度方向）からの発話を 8ch マイクアレイで収録したデータを作成した．その後，雑音として音楽データを話者と 60 度をなす方向から来るように作成し，SNR が 20dB から-5dB まで 5dB 刻みとなるように調整して音声データに加えた．画像データは，3 枚毎に 1 枚づつ使用し，一般のカメラのフレームレートに近い 33Hz として使用した．評価実験は，学習用データには含まれないデータセットから作成した 8ch 視聴覚データセットを用いた．評価用データは，学習に用いたデータセットに含まれる話者 5 人が発話した ATR 重要単語 50 単語を使用した（話者クローズ，単語オープンテスト）．

Ex.1 では，A-VAD/AV-VAD とマイクアレイ処理あり/なしの組み合わせ 4 通りの条件でＶＡＤを行った．この視聴覚統合には，十分な解像度の画像データを使用した．**Ex.2** では，ASR，VSR，AVSR の孤立単語認識の性能比較を行った．

4·1　結果

Ex.1, 2 の結果を Fig. 4，5，6 に示す．Fig 4 に各条件での受信者動作特性（receiver operating characteristics: ROC）カーブを示す．音声発話区間検出は，SNR が低くなるにつれ性能が悪化するが，視聴覚統合により大きく性能が向上している．マイクアレイ処理は SNR を改善するため，マイクアレイ処理を行わない場合に比べ性能が向上している．この結果は，VAD における視聴覚統合の有効性，および本稿で提案する視聴覚統合とマイクアレイ処理を組み合わせた手法が性

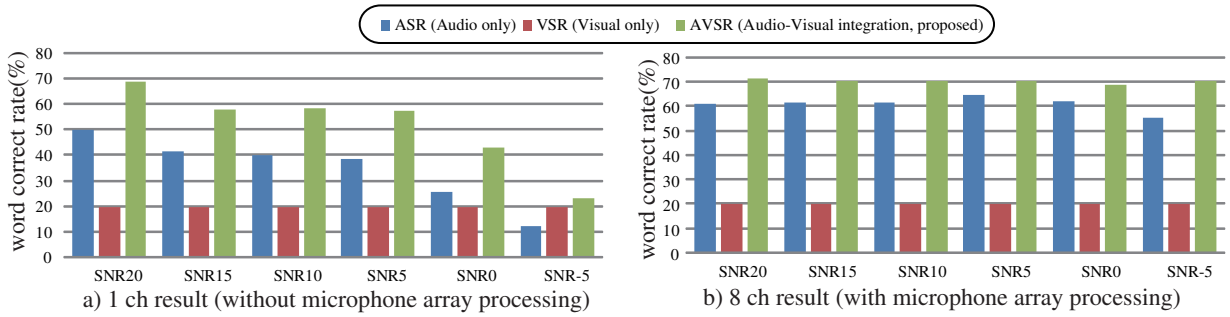

Fig.5 The effect of AV integration in ASR

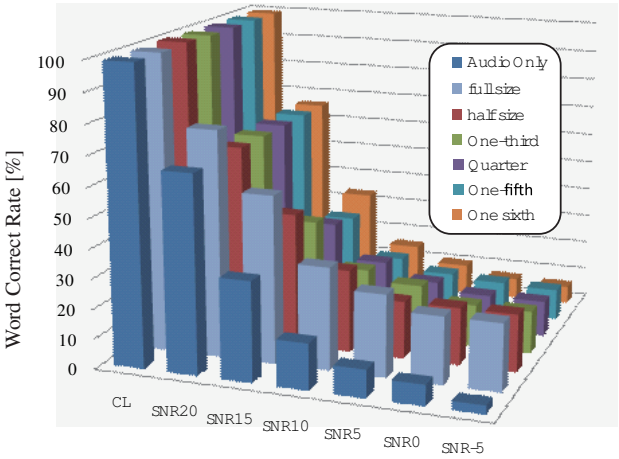

Fig.6 The robustness for face size changes

能をより向上させることを示している．

Fig 5 は音声認識実験の結果を表している．AVSR の性能が ASR，VSR に比べ向上している．単語に関してオープンな条件で評価であるが，提案手法では 70%の単語正解精度を達成した．音声入力にマイクアレイ処理を行わない場合，視聴覚統合により 16.7 ポイント性能が向上した．音声入力にマイクアレイ処理を行った場合，SNR 改善により ASR の性能が向上したにも関わらず，さらに 9.8 ポイント性能が向上した．

5. 終わりに

本稿では，音声認識の性能向上とロバスト性向上のため，発話区間検出でのベイジアンネットワークに基づいた視聴覚統合と音声認識でのミッシングフィーチャー理論に基づいた視聴覚統合の 2 段階で視聴覚統合を行う枠組みを提案した．提案した 2 層視聴覚統合音声認識システムをオープンソースソフトの HARK を用いて実装した．これにより，視聴覚統合と音源定位，音源分離から構成されるマイクアレイ処理と統合された．全体のシステムを単語オープンテストにより評価し，1) AV-VAD と AVSR を統合した手法が有効であること，2) マイクアレイ処理が SNR を改善し，音声認識の性能を向上させること，3) 2 階層視聴覚統合とマイクアレイ処理の統合によりさらに音声認識のノイズロバスト性が向上することを示した．

本稿では，音響ノイズと顔の大きさの変化についての評価を行ったが，実環境では残響や照明条件の変化，話者の顔の向きなども動的に変化する．これらの変化にロバストな手法の開発が今後の課題である．また，ロボットは動作することが可能であるという点を有効に利用し，音声認識の性能向上にロボットの動作を利用した手法の開発も今後の課題である．

謝辞 本研究に関して議論や助言を頂いた東京工業大学の井村順一教授，早川朋久准教授各氏に感謝いたします．また，MindReader の利用を許可して頂いた Massachusetts Institute of Technology の Dr. Rana el Kaliouby, Prof. Rosalind W. Picard 各氏に感謝いたします．

[1] K. Nakadai, *et al.*, “Active audition for humanoid,” AAAI-2000, pp. 832–839.
[2] S. Yamamoto, *et al.*, “Real-time robot audition system that recognizes simultaneous speech in the real world, ” IROS 2006, pp. 5333–5338.
[3] G. Potamianos, *et al.*, “A cascade visual front end for speaker independent automatic speechreading,” Speech Technology, vol. 4, pp. 193–208, 2001.
[4] S. Tamura, *et al.*, “A stream-weight optimization method for multi-stream hmms based on likelihood value normalization,” ICASSP 2005, SP-P5.2, 2005.
[5] J. Fiscus, “A post-processing systems to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” ASRU-97, pp. 347–354.
[6] T. Koiwa, *et al.*, “Coarse speech recognition by audio-visual integration based on missing feature theory,” IROS-2007, pp. 1751–1756.
[7] P. Liu, *et al.*, “Voice activity detection using visual information,” ICASSP, 2004, pp. 609–612.
[8] B. Rivet, *et al.*, “Visual voice activity detection as a help for speech source separation from convolutive mixtures,” Speech Communication, vol. 49, no. 7-8, pp. 667–677, 2007.
[9] K. Murai *et al.*, “Face-to-talk: audio-visual speech detection for robust speech recognition in noisy environment,” IEICE 2003, pp. 505–513.
[10] F. Asano, *et al.*, “Real-time sound source localization and separation system and its application to automatic speech recognition,” Eurospeech 2001, pp. 1013–1016.
[11] J.-M. Valin, *et al.*, “Enhanced robot audition based on microphone array source separation with post-filter,” IROS 2004, pp. 2123–2128.
[12] Y. Nishimura, *et al.*, “Noise-robust speech recognition using multi-band spectral features,” ASA Meetings, no. 1aSC7, 2004.
[13] Y. Nishimura, *et al.*, “Speech recognition for a humanoid with motor noise utilizing missing feature theory, ” Humanoids 2006, pp. 26-33.