An Improvement in Audio-Visual Voice Activity Detection for Automatic Speech Recognition

Takami Yoshida[1], Kazuhiro Nakadai[1,2], and Hiroshi G. Okuno[3]

[1] Graduate School of Information Science and Engineering,
Tokyo Institute of Technology, Tokyo, Japan
yoshida@cyb.mei.titech.ac.jp
[2] Honda Research Institute Japan, Co., Ltd., Saitama, Japan
nakadai@jp.honda-ri.com
[3] Graduate School of Informatics, Kyoto University, Kyoto, Japan
okuno@kuis.kyoto-u.ac.jp

Abstract. Noise-robust Automatic Speech Recognition (ASR) is essential for robots which are expected to communicate with humans in a daily environment. In such an environment, Voice Activity Detection (VAD) strongly affects the performance of ASR because there are many acoustically and visually noises. In this paper, we improved Audio-Visual VAD for our two-layered audio visual integration framework for ASR by using hangover processing based on erosion and dilation. We implemented proposed method to our audio-visual speech recognition system for robot. Empirical results show the effectiveness of our proposed method in terms of VAD.

Index Terms: Audio-Visual integration, Voice Activity Detection, Speech Recognition.

1 Introduction

Service/home robots which are required to communicate with humans in daily environments should have a noise-robust Automatic Speech Recognition (ASR) function. In such an environment, there are many kinds of noises such as other speakers and robot's own noise. So, robots should cope with contaminated input signals.

To realize such a robot, there are two approaches. One is sound source separation to improve SNR of the input speech. The other is the use of another modality, that is, audio-visual (AV) integration.

For sound source separation, we can find several studies, especially, in the field of "Robot Audition" proposed in [1], which aims at building listening capability for a robot by using its own microphones. Some of them reported highly-noise-robust speech recognition such as three simultaneous speech recognition [2]. However, in a daily environment where acoustic conditions such as power, frequencies and locations of noise and speech sources dynamically change, the performance of sound source separation sometimes deteriorates, and thus ASR

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part I, LNAI 6096, pp. 51–61, 2010.

does not always show such a high performance. For AV integration for ASR, many studies have been reported as *Audio-Visual Speech Recognition (AVSR)* [3,4,5]. However, they assume that the high resolution images of the lips are always able to be available. And, most AVSR assumed that the voice activity is given in advance. Thus, their methods have difficulties in applying them to robots.

To solve the difficulties, we proposed a ***two-layered AV integration framework*** [6]. This framework applies audio-visual integration both to voice activity detection (the first layer AV integration) and to speech recognition (the second layer AV integration) for improving noise robustness. The AVSR system showed high noise-robustness and high speech recognition performance in acoustically and/or visually noisy conditions. However, the effectiveness of AV integration is shown when the image resolution is low. Thus, this integration method is of less use.

For this issue, we introduce two approaches. One is feature integration with time-lag. Visual features and audio features are not synchronized, so this method takes time-lag into account. The other is the use of pattern classification methods, that is, erosion and dilation.

The rest of this paper is organized as follows: Section 2 discusses issues in audio and visual voice activity detection, and Section 3 shows an approach for AV-VAD. Section 4 describes our automatic speech recognition system for robots using two-layered AV integration, that is, AV-VAD and AVSR. Section 5 shows evaluation in terms of VAD and ASR performance. The last section concludes this paper.

2 Issues in Audio and Visual Voice Activity Detection

This section discusses issues in voice activity detection (Audio VAD) and lip activity detection (Visual VAD) for robots and their integration (AV-VAD), because VAD is an essential function for ASR.

2.1 Audio VAD

VAD detects the start and the end points of an utterance. When the duration of the utterance is estimated shorter than the actual one, that is, the start point is detected with some delay and/or the end point is detected earlier, the beginning and the last part of the utterance is missing, and thus ASR fails. Also, an ASR system requires some silent signal parts (300-500 ms) before and after the utterance signal. When the silent parts are too long, it also affects the ASR system badly. Therefore, VAD is crucial for ASR, and thus, a lot of VAD methods have been reported so far. They are mainly classified into three approaches as follows:

A-1: The use of acoustic features,
A-2: The use of the characteristics of human voices,
A-3: The use of intermediate speech recognition results using ASR.

Common acoustic features for **A-1** are energy and Zero-Crossing Rate (ZCR), but energy has difficulty in coping with an individual difference and a dynamic change in voice volume. ZCR is robust for such a difference/change because it is a kind of frequency-based feature. On the other hand, it is easily affected by noise, especially, when the noise has power in speech frequency ranges. Therefore, a combination of energy and ZCR is commonly used in conventional ASR systems. However, it is still prone to noise because it does not have any prior knowledge on speech signals.

For **A-2**, Kurtosis or Gaussian Mixture Model (GMM) is used. This shows high performance in VAD when it is performed in an expected environment, that is, an acoustic environment for a VAD test is identical to that for GMM training. However, when the acoustic environment changes beyond the coverage of the model, VAD easily deteriorates. In addition, to achieve noise robust VAD based on these methods, a large number of training data is required.

A-3 uses the ASR system for VAD, and thus, this is called decoder-based VAD. An ASR system basically has two stages for recognition. At the first stage, the ASR system computes log-likelihood of silence for an input signal at every frame. By using the computed log-likelihood, VAD is performed by thresholding x_{dvad} defined by

$$x_{dvad} = \log(p(\omega_0|x)) \tag{1}$$

where x is audio input, and ω_0 shows the hypothesis that x is silence.

Actually, this mechanism is already implemented on open-sourced speech recognition software called "Julius"[1]. It is reported that this approach shows quite high performance in real environments Although this approach sounds like the chicken-or-egg dilemma, this result shows that integration of VAD and ASR is effective.

Thus, each method has unique characteristics, and none of them are suitable for all-purpose use. **A-1** is still commonly-used, **A-3** has the best performance.

2.2 Visual VAD

Visual VAD means lip activity detection in visual speech recognition which corresponds to audio VAD in ASR. The issues in visual VAD for integration with audio VAD and AVSR are as follows:

B-1: The limitation of frame rate,
B-2: The robust visual feature.

The first issue is derived from the hardware limitation of conventional cameras. The frame rate of a conventional camera is 30 Hz, while that of acoustic feature extraction in ASR is usually 100 Hz. Thus, when we integrate audio and visual features, a high speed camera having a 100 Hz capturing capability or a synchronization technique like interpolation is necessary.

For the second issue, a lot of work has been studied in the AVSR community so far. A PCA-based visual feature [7], and a visual feature based on width and

[1] `http://julius.sourceforge.jp/`

length of the lips[8] were reported. However, these features are not robust enough for VAD and AVSR because visual conditions change dynamically. Especially, the change in a facial size is hard to be coped with, since the facial size is directly related to facial image resolution. Thus, an appropriate visual feature should be explored further.

2.3 Audio-Visual VAD

AV integration is promising to improve the robustness of VAD, and thus, audio and visual VAD should be integrated to improve AVSR performance in the real world. In this case, we have two main issues. One is AV synchronization as described above. The other is the difference between audio and visual VAD. The ground truth of visual VAD is not always the same as that of audio VAD, because extra lip motions are observed before and after an utterance to open/close the lips. AV-VAD which integrates audio and visual VAD should take their differences into account. To avoid this problem, Murai *et al.* proposed two-stage AV-VAD [9]. First, they extract lip activity based on a visual feature of inter-frame energy. Then, they extract voice activity by using speech signal power from the extracted lip activity. However, in this case, when either the first or the second stage fails, the performance of the total system deteriorates.

In robotics, AV-VAD and AVSR have not been studied well although VAD is essential to cope with noisy speech. Asano *et al.* used AV integration for speech recognition, but their AV integration was limited to sound source localization [10]. Nakadai *et al.* also reported that AV integration in the level of speaker localization and identification indirectly improved ASR in our robot audition system [11]. However, in their cases, VAD was just based on signal power for a speaker direction which is estimated in AV sound source localization, that is, they indirectly used AV integration for VAD.

3 Improved AV-VAD

This section describes AV-VAD in our two-layered AV integration.

3.1 Audio VAD

For audio VAD, three approaches are described in the previous section, and the **A-3** approach has the best performance. Thus, we used decoder-based VAD as one of **A-3** approaches.

3.2 Visual VAD

We use a visual feature based on width and length of the lips, because this feature is applicable to extract viseme feature in the second layer of AV integration, i.e., AVSR.

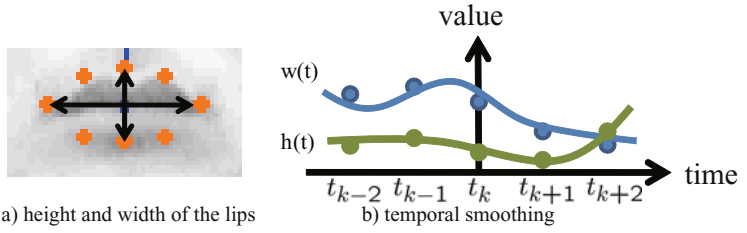

Fig. 1. Visual feature extraction

To extract the visual feature, we, first, use Facial Feature Tracking SDK which is included in MindReader[2]. Using this SDK, we detect face and facial components like the lips. Because the lips are detected with its left, right, top, and bottom points, we easily compute the height and the width of the lips, and normalize them by using a face size estimated in face detection shown in Fig. 1a).

After that, we apply temporal smoothing for the consecutive five-frame height and width information by using a 3rd-order polynomial fitting function as shown in Fig. 1b). The motion of the lips is relatively slow, and the visual feature does not contain high frequency components. Such high frequency components are regarded as noise. This is why temporal smoothing is performed to remove the noise effect. Thus, we can get four coefficients such as a_i – d_i for height and another four for width described as below.

$$y_i(t) = a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3, \tag{2}$$

where $y_i(t)$ corresponds to fitting result of height or width information ($h(t)$ or $w(t)$ in Fig. 1b). In total, eight coefficients are obtained as a visual feature vector.

For the frame rate problem mentioned in Section 2-B, we propose to perform up-sampling for the extracted eight coefficients so that they can easily synchronize with audio features. As a method of up-sampling, we used cubic spline interpolation

3.3 Audio-Visual VAD

AV-VAD integrates audio and visual features using a Bayesian network shown in Fig. 2, because the Bayesian network provides a framework that integrates multiple features with some ambiguities by maximizing the likelihood of the total integrated system. Actually, we used the following features as the inputs of the Bayesian network:

- The score of log-likelihood for silence calculated by Julius (x_{dvad}),
- Eight coefficients regarding the height and the width of the lips (x_{lip}),
- The belief of face detection which is estimated using Facial Feature Tracking SDK (x_{face}).

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

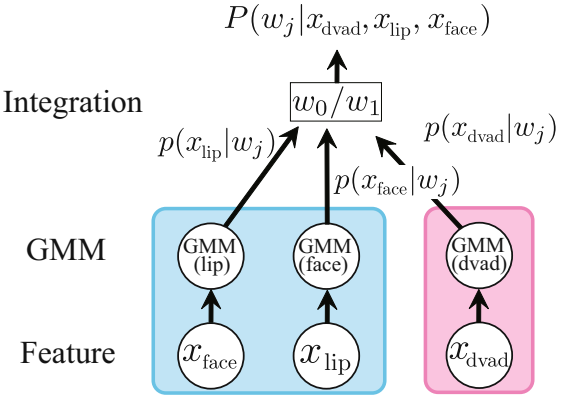

Fig. 2. AV-VAD based on a Bayesian network

Since the score of log-likelihood tend to extract voice activity with delay. So, we use the log-likelihood of previous frame, instead of current frame.

Audio and visual features have errors more or less, the Bayesian network is an appropriate framework for AV integration in VAD. The Bayesian network is based on the Bayes theory defined by

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}, \qquad j = 0, 1 \tag{3}$$

where x corresponds to each feature such as x_{dvad}, x_{lip}, or x_{face}. A hypothesis ω_j shows that ω_0 or ω_1 corresponds to a silence or a speech hypothesis, respectively. A conditional probability, $p(x|\omega_j)$, is obtained using a 4-mixture GMM which is trained with a training dataset in advance. The probability density function $p(x)$ and probability $P(\omega_j)$ are also pre-trained with the training dataset. A joint probability, $P(\omega_j|x_{dvad}, x_{lip}, x_{face})$, is thus calculated by

$$P(\omega_j|x_{dvad}, x_{lip}, x_{face}) = P(\omega_j|x_{dvad})P(\omega_j|x_{lip})P(\omega_j|x_{face}). \tag{4}$$

By comparing this probability and threshold, we estimate voice activity. Next, we perform dilation and erosion for the temporal sequence of estimated voice activity. Dilation and erosion are commonly used in pattern recognition. Fig. 3 shows the results of these processes. In dilation, a frame is added to the start-point and end-point of voice activity as below.

$$\hat{V}[k] = \begin{cases} 1 & \text{if } V[k-1] = 1 \text{ or } V[k+1] = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $V[k]$ is the estimated voice activity at k frame and $\hat{V}[k]$ is the result of dilation. In erosion, a frame is removed from the start- and end-point of voice activity as below.

$$\hat{V}[k] = \begin{cases} 0 & \text{if } V[k-1] = 0 \text{ or } V[k+1] = 0 \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

AV-VAD performs these processes several times and decides voice activity.

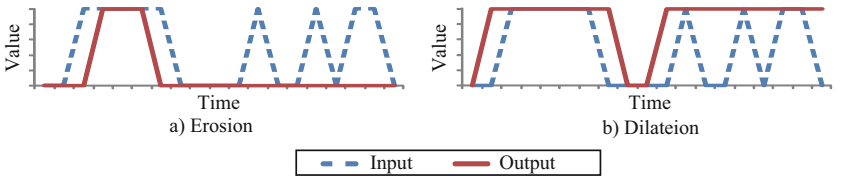

Fig. 3. Erosion and dilation

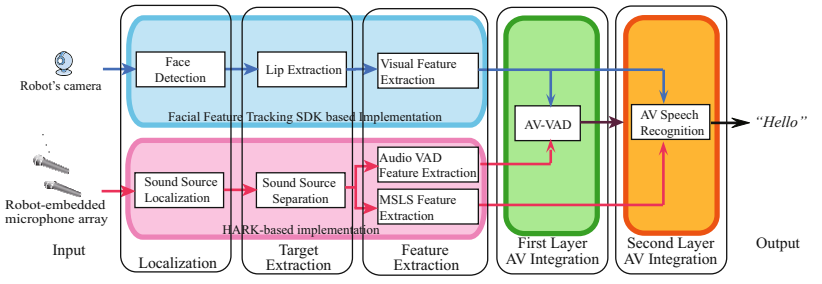

Fig. 4. Two-Layered Audio-Visual Integration Framework

4 System Implementation

Fig. 4 shows our automatic speech recognition system for robots with two-layered AV integration, that is, AV-VAD and AVSR. It consists of four implementation blocks as follows;

- Facial Feature Tracking SDK based implementation for visual feature extraction,
- HARK-based implementation for microphone array processing to improve SNR and acoustic feature extraction,
- The first layer AV integration for AV-VAD,
- The second layer AV integration for AVSR.

Four modules in *Facial Feature Tracking SDK based implementation block* were already described in Section 3.2, and *the first layer AV integration for AV-VAD* was also explained in Section 3.3. Thus, the remaining two blocks are mainly described in this section.

4.1 HARK-Based Implementation Block

This block consists of four modules, that is, sound source localization, sound source separation, audio VAD feature extraction, and MSLS feature extraction. Their implementation is based on HARK mentioned in Section 1. The audio VAD feature extraction module was already explained in Section 3.1, and thus, the other three modules are described. We used an 8 ch circular microphone array which is embedded around the top of our robots head.

For sound source localization, we used MUltiple SIgnal Classification (MUSIC) [12]. This module estimates sound source directions from a multi-channel audio signal input captured with the microphone array.

For sound source separation, we used Geometric Sound Separation (GSS) [13]. GSS is a kind of hybrid algorithm of Blind Source Separation (BSS) and beamforming. GSS has high separation performance originating from BSS, and also relaxes BSS's limitations such as permutation and scaling problems by introducing "geometric constraints" obtained from the locations of microphones and sound sources obtained from sound source localization.

For an acoustic feature for ASR systems, Mel Frequency Cepstrum Coefficient (MFCC) is commonly used. However, sound source separation produces spectral distortion in the separated sound, and such distortion spreads over all coefficients in the case of MFCC. Since Mel Scale Logarithmic Spectrum (MSLS) [14] is an acoustic feature in a frequency domain, and thus, the distortion concentrates only on specific frequency bands. Therefore MSLS is suitable for ASR with microphone array processing. We used a 27-dimensional MSLS feature vector consisting of 13-dim MSLS, 13-dim ΔMSLS, and Δlog power.

4.2 The Second Layer AV Integration Block

This block performs AVSR. We simply introduced our reported AVSR for robots [15] as mentioned in Section 1, because this AVSR system showed high noise-robustness to improve speech recognition even when either audio or visual information is missing and/or contaminated by noises. This kind of high performance is derived from missing feature theory (MFT) which drastically improves noise-robustness by using only reliable acoustic and visual features by masking unreliable ones out. In this paper, this masking function is used to control audio and visual stream weights which are decided to be optimal manually in advance. For ASR implementation, MFT-based Julius [16] was used.

5 Evaluation

In this experiment, we used a Japanese word AV dataset. This dataset contains 10 male speech data and 266 words for each male. Audio data was sampled at 16 kHz and 16 bits, and visual data was 8 bit monochrome and 640x480 pixels in size recorded at 33 Hz. For training an AV-VAD model, we used 216 acoustically and visually clean AV data by 5 males in this AV dataset. For AVSR acoustic model training, we used 216 clean AV data by 10 males in this AV dataset.

The audio data was 8 ch data converted from 1 ch data so that each utterance comes from 0 degrees by convoluting a transfer function of the 8 ch robot-embedded microphone array. After that, we added a music signal from 60° as a noise source. The SNR changed from 20 dB to 0 dB at 10 dB increments. For the test dataset, another 50 AV data which were not included in the training dataset were selected from the synthesized 8 ch AV data.

In experiment, **C-1** AV-VAD reported by [6] and **C-2** proposed method were examined. Fig. 5 and Fig. 6 shows VAD results in various conditions using ROC

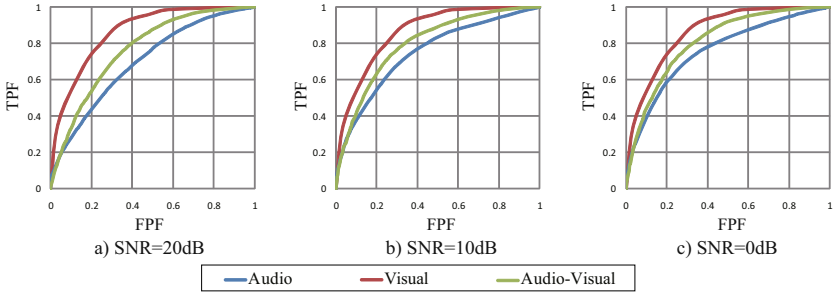

Fig. 5. The ROC curve of VAD with **C-1**

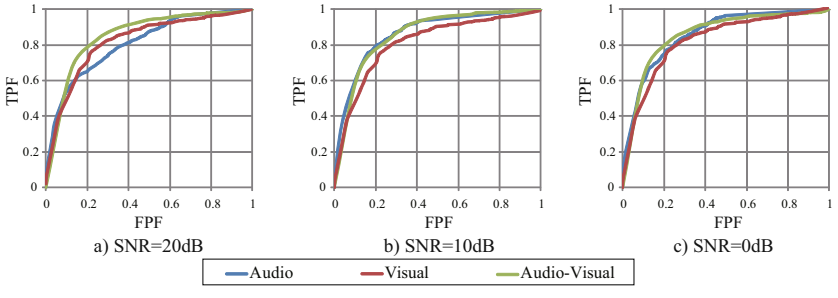

Fig. 6. The ROC curve of VAD with **C-2**(proposed method)

curves. By comparing Fig.5 and Fig. 6, we can see that proposed method improves Audio-VAD performance. The AV-VAD performances were the best in every condition with **C-2**, while Visual-VAD performances were the best with **C-1**. This result shows that the combination of proposed method and AV integration is effective in VAD.

6 Conclusion

We introduced two method for AV-VAD. And we implemented these two methods to two-layered audio-visual integration AVSR system. VAD performance was evaluated using high resolution images, and we show that our proposed method improves AV-VAD performance when the input images are high resolution.

The future work is to cope with visual noises such as reverberation, illumination, and facial orientation. In this paper, we evaluate robustness for acoustical noises and face size changes, but other dynamic changes exist in a daily environment. To apply AVSR to a robot, AVSR should be robust for such changes.

Acknowledgments

We thank Prof. R. W. Picard and Dr. R. E. Kaliouby, MIT for allowing us to use their system.

References

1. Nakadai, K., Lourens, T., Okuno, H.G., Kitano, H.: Active audition for humanoid. In: Proceedings of 17th National Conference on Artificial Intelligence, pp. 832–839 (2000)
2. Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.M., Komatani, K., Ogata, T., Okuno, H.G.: Real-time robot audition system that recognizes simultaneous speech in the real world. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5333–5338 (2006)
3. Potamianos, G., Neti, C., Iyengar, G., Senior, A., Verma, A.: A cascade visual front end for speaker independent automatic speechreading. Speech Technology, Special Issue on Multimedia 4, 193–208 (2001)
4. Tamura, S., Iwano, K., Furui, S.: A stream-weight optimization method for multi-stream hmms based on likelihood value normalization. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 469–472 (2005)
5. Fiscus, J.: A post-processing systems to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In: Proceedings of Workshop on Automatic Speech Recognition and Understanding, pp. 347–354 (1997)
6. Yoshida, T., Nakadai, K., Okuno, G.H.: Automatic speech recognition improved by two-layered audio-visual speech recognition for robot audition. In: Proceedings of 9th IEEE-RAS International Conference on Humanoid Robots, pp. 604–609 (2009)
7. Liu, P., Wang, Z.: Voice activity detection using visual information. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 609–612 (2004)
8. Rivet, B., Girin, L., Jutten, C.: Visual voice activity detection as a help for speech source separation from convolutive mixtures. Speech Communication 49, 667–677 (2007)
9. Murai, K., Nakamura, S.: Face-to-talk: audio-visual speech detection for robust speech recognition in noisy environment. IEICE TRANSACTIONS on Information and Systems E86-D, 505–513 (2003)
10. Asano, F., Motomura, Y., Aso, H., Yoshimura, T., Ichimura, N., Nakamura, S.: Fusion of audio and video information for detecting speech events. In: Proceedings of International Conference on Information Fusion, pp. 386–393 (2003)
11. Nakadai, K., Matsuura, D., Okuno, H.G., Tsujino, H.: Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. Speech Communication 44, 97–112 (2004)
12. Asano, F., Goto, M., Itou, K., Asoh, H.: Real-time sound source localization and separation system and its application to automatic speech recognition. In: Proceedings of International Conference on Speech Processing, pp. 1013–1016 (2001)
13. Valin, J.M., Rouat, J., Michaud, F.: Enhanced robot audition based on microphone array source separation with post-filter. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2123–2128 (2004)

14. Nishimura, Y., Shinozaki, T., Iwano, K., Furui, S.: Noise-robust speech recognition using multi-band spectral features. Acoustical Society of America Journal 116, 2480–2480 (2004)
15. Koiwa, T., Nakadai, K., Imura, J.: Coarse speech recognition by audio-visual integration based on missing feature theory. In: Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems, pp. 1751–1756 (2007)
16. Nishimura, Y., Ishizuka, M., Nakadai, K., Nakano, M., Tsujino, H.: Speech recognition for a humanoid with motor noise utilizing missing feature theory. In: Proceedings of 6th IEEE-RAS International Conference on Humanoid Robots, pp. 26–33 (2006)